Journal Information
Journal ID (publisher-id): jgi
ISSN: 1910-7595
Publisher: Centre for Addiction and Mental Health

Article Information

Received Day: 2 Month: July Year: 2002
Accepted Day: 22 Month: July Year: 2003
Publication date: October 2003
Publisher Id: jgi.2003.9.1
DOI: 10.4309/jgi.2003.9.1

Meta-analysis: A 12-step program

L. Streiner David, PhD, CPsych

Affiliation: Baycrest Centre for Geriatric Care, Toronto, Ontario, Canada, E-mail: dstreiner@klaru-baycrest.on.ca

For correspondence: David L. Streiner, PhD, CPsych. Director, Kunin-Lunenfeld Applied Research Unit Baycrest Centre for Geriatric Care 3560 Bathurst Street Toronto, Ontario, Canada M6A 2E1, Telephone: (416) 785-2500, x2534, Fax: (416) 785-4230, E-mail: dstreiner@klaru-baycrest.on.ca

After graduating from the clinical psychology department at Syracuse University (New York), I joined the faculty of health sciences at McMaster University (Hamilton, Ontario), in the departments of psychiatry and clinical epidemiology and biostatistics. My aim was to stay for about two years. Thirty years later, I retired from McMaster, and the next day, moved to the Baycrest Centre for Geriatric Care in Toronto as director of the Kunin-Lunenfeld Applied Research Unit and assistant V.P., research; and as a professor in the department of psychiatry (University of Toronto). My main research interests are (a) determining which types of woods work best for furniture I make for my grandchildren, and (b) whether songs sound better played on the banjo or guitar. In between these activities, I have published four books, 10 book chapters, and about 200 articles spanning a range of research areas, from statistics to schizophrenia, to scale development and the sequelae of extremely low birth weight.

This article prints out to about 23 pages

Abstract

Meta-analysis is a technique for combining the results of many studies in a rigorous and systematic manner, to allow us to better assess prevalence rates for different types of gambling and determine which interventions have the best evidence regarding their effectiveness and efficacy. Meta-analysis consists of (a) a comprehensive search for all available evidence; (b) the use of applying explicit criteria for determining which articles to include; (c) determination of an effect size for each study; and (d) the pooling of effect sizes across studies to end up with a global estimate of the prevalence or the effectiveness of a treatment. This paper begins with a discussion of why meta-analyses are useful, followed by a 12-step program for conducting a meta-analysis. This program can be used both by people planning to do such an analysis, as well as by readers of a meta-analysis, to evaluate how well it was carried out.

The purpose of this article is to describe a technique called meta-analysis to people engaged in counselling those with gambling problems, to enable them to either read meta-analyses with greater understanding or perhaps even conduct one on their own. The value in understanding the bases of meta-analyses comes with being able to read one and assess if it has sound methodology. We can expect more treatment outcomes to be assessed through meta-analyses, and it serves clinicians well to understand how such an analysis was completed, not simply to accept it on faith or the author's reputation. My aim is to make this paper relevant for the broadest range of readers: those with research-oriented PhDs as well as community college graduates. For those who are comfortable with statistics, the relevant formulae are provided. However, readers who wish to gain mainly a conceptual understanding of meta-analysis without going into the details can easily skip the technical parts, which are set off in boxes to make them easier to avoid.

Let's start off with a tongue-in-cheek multiple-choice question.

Which of the following options reflects current thinking about meta-analysis?

1. a. Meta-analysis is a rigorous method for objectively combining the results of many different studies to arrive at a better estimate of truth.
2. b. Meta-analysis is the greatest boon to humanity since the invention of the double bed.
3. c. Meta-analysis is a way of combining the results of many inadequate studies to arrive at an inadequate answer.
4. d. Meta-analysis is the new growth industry of social science and biomedical research, allowing people to build up their C.V.
5. e. All of the above.

If you chose option (e), you'd be a winner. Meta-analyses have indeed swept the worlds of psychology and medicine, and this has even led to the creation of a large international group (the Cochrane Collaboration) devoted to their production and dissemination. In 1991, Chalmers (1991) found 150 meta-analyses of randomized controlled trials (RCTs). Using MEDLINE with the search term “meta-analysis,” I identified 609 articles published in 1996 (the first year for which that search term could be used); and by 2001, there were more than twice this number (1,251), with no sign that this trend is slowing down.

On the other hand, there are some people who feel that, with meta-analysis, “bad science drives out good by weight of numbers” (Wachter, 1988, p. 1407); that is, summing the findings of many poorly done studies with the results of a few good ones with opposite conclusions will overwhelm the latter. In fact, there are sometimes discrepancies between the findings of meta-analyses and those of large clinical trials (Furukawa, Streiner & Hori, 2000; Ioannidis, Cappelleri & Lau, 1998), and some researchers have advocated a more qualitative synthesis of “best evidence” rather than a quantitative summation of all evidence (Slavin, 1986).

So, what is all this debate about? In this article, I will first outline the rationale for using meta-analysis. Then, as the readers of this journal are no doubt familiar with treatment programs designed for people with addictions, I will give my own 12-step program for dealing with meta-analyses. This program can be used in two ways: for people contemplating doing a meta-analysis, it can serve as a how-to guide, to what they should do, in what order, and with references to resources for more advanced information; for readers who do not have training in statistics, as a quality control checklist, to see if an author took adequate care to ensure results that are relatively unbiased, fair and accurate. Readers in the latter category can safely skip over the statistics and equations, which are set off in boxes (unless they are masochistically inclined). Many of the examples come from health literature (and outside gambling studies) because that is where most of the current literature resides and where some crucial findings originate. However, the applicability of meta-analyses from other areas to studies within the field of gambling should be readily apparent.

Although there have been meta-analyses of diagnostic instruments (Hasselblad & Hedges, 1995), and even one on the genetics of gambling (Walters, 2001), the vast majority of meta-analyses address issues of the effectiveness and efficacy of treatment interventions. Consequently, this article will focus mainly on this type of study, although the principles can be applied to meta-analyses of any kind.

The rationale for meta-analysis

No one who has tried to keep abreast of advances in his or her own field needs to be convinced of the growth of published articles. Busy clinicians, and even researchers, have always needed some way of keeping up-to-date without having to find the original articles, get them from a library and read them. In large measure, this role was (and continues to be) filled by textbooks and chapters in annual reviews. This presents two problems. First, the publication lag between starting a chapter and finally seeing it in print can be two or three years, during which time the field has moved on. A more serious problem, though, is the potential for bias to creep (or storm) in. There is no guarantee that the authors of the review chapter have tried to locate all of the relevant articles, much less those that do not agree with their pre-existing beliefs or prejudices. Furthermore, they may deploy their methodological rigour differentially, reserving their harshest criticism for articles they disagree with, and passing over faults in those supportive of their position. For example, Munsinger (1975) and Kamin (1978) reviewed the same articles about the relative effects of genetics and environment on intelligence, but arrived at diametrically opposite conclusions; conclusions that not surprisingly supported their own views. In the area of gambling, two non-systematic reviews of naltrexone similarly came to opposite conclusions regarding its effectiveness (Hollander, Buchalter & DeCaria, 2000; Modesto-Lowe & Van Kirk, 2002).

The first step in addressing some of the faults of review chapters or papers is to do a systematic search of the literature, to maximize the chances that all of the relevant articles will be found (although problems with this are outlined below in Step 3), and spell it out in sufficient detail that the reader could replicate the search and end up with the same articles. The issue then becomes what to do with all of the findings. The simplest solution is simply “vote counting"; how many articles come to one conclusion and how many the opposite. Yet again, however, two problems rear their heads. The first is what to do when the vote is close. For example, of 27 articles that reported the relationship between obesity and socioeconomic status in men, 12 found a positive correlation, 12 found a negative one and three reported no relationship (Sobal & Stunkard, 1989). The second problem is that vote counting does not take the quality of the study into account. A study with a correlation of +0.2 is balanced by one with a correlation of –0.9; and one with a sample size of 50 given the same weight as one with 10 times the number of subjects.

The solution is to add a further step to a systematic review, and that is to combine the results in a way that takes the magnitude of the findings, the sample size and the quality of the research into account. This is what Smith and Glass (1977) have called “meta-analysis.” So, a meta-analysis consists of

- a comprehensive search for all applicable articles;
- explicit and reliable criteria for selecting which articles to include;
- derivation of a measure of the magnitude of the effect of the intervention; and
- a method for combining the results of all of the studies.

Step 1. Defining the question

It may seem that defining the question to be addressed by the meta-analysis is a simple and straightforward task. As with most things in life, if it looks easy and problem-free, it's a sure bet that there are major problems ahead, and this is no exception. A question like, “What treatment programs work for problem gambling?” is too broad to yield meaningful conclusions. It will result in trying to combine studies looking at the many different types of gambling (e.g. betting on horse races, using slot machines, playing cards, lotteries, etc.) with many different populations (people who enter therapy on their own initiative as opposed to those who are ordered into therapy by spouses or the courts; men versus women; people who have been gambling for years versus those who have just started; people in a stable relationship with a non-gambler versus those whose marriage ended in divorce because of the gambling versus those who have never married; “action” versus “escape” gamblers; and so forth). It is quite possible that treatments that work for one type of gambling or with one group of people may not work for others. If the results of these different studies are combined, one of two misleading results may be drawn. First, unsuccessful studies may wash out the results of successful ones, so that we erroneously conclude that no intervention is successful. Second, the opposite effect may occur, where we reach the global conclusion that all the treatments work for all people, whereas in reality only certain ones may be effective and only for some types of gambling or some groups of people.

The more focussed the question, the more useful and more accurate the results of the meta-analysis. So, the question should be phrased more like, “How effective is treatment A for gambling problem B with this specific group of people?” There is a downside to being too specific (this is the “So what else is new?” effect). Once we've done the search and applied the inclusion and exclusion criteria (steps that will be explained later), we may find that there are no articles that address the question. At this point, we will have to broaden one or more of the parts of the question (e.g. by including different types of gambling) and repeat the steps. It may be necessary to do this a few times, depending on how many articles exist in the literature and their quality. If there are many, high quality interventions with different groups of gamblers, then we may end up with a highly focussed meta-analysis. Conversely, if most of the reports do not have control groups, or consist of self-selected, highly motivated people, then we may have to keep the question broad to get any meaningful results.

In Step 11 of our program, we will examine some ways of teasing apart what is successful from what is not when we do have a heterogeneous mix of studies.

Step 2. Detailing the selection criteria

Once the question has been defined, prepare a checklist so that the criteria can be applied in a uniform fashion. The list need not be long, but should include all of the reasons for accepting or rejecting articles. For example, if the question reads, “Is cognitive behavioural therapy (CBT) effective for adults with a problem with electronic gambling machines (EGMs)?" (EGMs include video lottery terminals, slot machines, poker and video poker machines), and we want to look only at RCTs, then the checklist can be as simple as the one shown in Table 1.

The reason for the phrases "At least one CBT group" and "Identifiable EGM group" in the checklist is that we want to include articles that may also involve other types of therapy or other forms of gambling, as long as the results allow us to look at the specific groups of interest. That is, if a study used people with various forms of problem gambling, but the results were reported separately for each type, or the author was willing to provide us with the necessary data, then we would include that article. On the other hand, if the results were reported with all types of gambling combined and we were unable to gain access to the raw data, then we would reject the study.

Step 3. Doing the search

The next step is to actually find the articles. Computerized databases, such as MEDLINE, PsycINFO, EMBASE, CINAHL and the like have made our lives infinitely easier in this regard. However, we should not be lulled into thinking that, because we do a computerized search, all *or even most of* the relevant articles will be identified. A number of studies have found that even small changes in the search strategy result in very different sets of articles being retrieved (Haynes et al., 1985); and that even competently done searches may find no more than 30% of the existing papers (Dickersin, Hewitt, Mutch, Chalmers & Chalmers, 1985). Although MEDLINE has definitely improved since these articles were done, especially by adding the term "RCT" to the MeSH vocabulary and reclassifying nearly 100,000 trials it hadn't originally labelled as RCTs (Lefebvre & Clarke, 2001), the problem still remains that many articles will be missed. This means that other strategies must be used in addition to searching the computerized databases.

Perhaps the easiest, most fruitful method is to look through the reference lists of the articles that have been found, to see if they have identified studies you may have missed. This is then supplemented by hand-searching the five to 10 most relevant journals, such as the *Journal of Gambling Studies, Addictive Behaviours, Addictions* and *International Journal of the Addictions* from the gambling side; and *Behaviour Research and Therapy, Journal of Clinical and Consulting Psychology, Behavior Modification, Archives of General Psychiatry, American Journal of Psychiatry* and *British Journal of Psychiatry* from the treatment end.

Within the past few years, three other resources have been developed that are extremely useful. One is the Cochrane Database of Systematic Reviews (CDSR), which, as its name implies, is an on-line set of meta-analyses. There are a number of advantages to this database. The main advantage is that someone has already done the work for you (although this may deprive you of a publication if you had your heart set on doing one yourself). Second, strict methodological criteria have been used in the selection of the primary articles, so you can be sure that all of the articles in the review have met fairly exacting standards.

The downsides are, first, that the reviews are limited, for the most part, to RCTs of interventions; few reviews of diagnosis or clinical course are present. Second, the CDSR is a strictly volunteer effort, so a review of a particular topic exists only if someone had an interest in that area. No one is overseeing the entire effort and identifying reviews that should be done, so it's quite possible that the topic you want may not be present. Third, the vast majority of reviews come from the areas of medicine and psychology; relatively few are from the field of gambling. Indeed, a search for meta-analyses of treatments for pathological gambling turned up only one citation, that of Oakley-Browne, Adams and Mobberley (2001). Finally, "strict methodological criteria have been used in the selection of the primary articles, so you can be sure that all of the articles in the review have met fairly exacting standards." If this sounds exactly like one of the advantages, that's because it is. Many reviews start off by identifying 50 to 100 potential articles, but after the methodology filters have been applied, only one article remains. While this will undoubtedly be a very well designed and executed study, it is likely that there are many other studies that have been excluded but may still contain useful information. That is, the criteria may be *too* strict in some cases, especially for those of us who are not true believers that RCTs are the only road to truth and beauty.

The second Cochrane database that may be extremely useful is DARE, the Database of Abstracts of Reviews of Effectiveness. These are structured abstracts of excellent reviews of treatment effectiveness, which have the same advantages and disadvantages as the CSDR. Finally, there is the Cochrane Controlled Trials Register (CCTR), which is a listing of RCTs that has been compiled by the Cochrane collaborators. As of June 2002, the CCTR contains over 300,000 trials, including many RCTs that have not yet been combined into systematic reviews.

Finally, an excellent source is Dissertation Abstracts. Graduate students are, for the most part, far more obsessive than we are, and it's quite possible they have located some published articles that we missed. So, it's often worthwhile to get a copy of the dissertation and scan the reference list.

Unfortunately, these search strategies cover only published articles. The problem is that there is a "publication bias" (Begg & Berlin, 1988; Cooper, DeNeve & Charlton, 1997; Gilbody, Song, Eastwood & Sutton, 2000; Marshall et al., 2000), in that it is much easier to get articles with significant results accepted by editors than those that fail to find significance (we will discuss this further in Step 8). The difficulty is how to find this "grey literature" of unpublished results. One strategy is to write to authors and ask if they have studies sitting in file drawers that haven't seen the light of day. This is most useful if there are not too many researchers in the area, and most are known to you. It will miss people who may have done a few studies, failed to get them published, and moved on to more rewarding fields. Proceedings of meetings are another source of unpublished material. Abstracts from some meetings are sometimes published by a journal, especially if an organization sponsors both the meeting and the journal; and databases such as PsycINFO are starting to include some proceedings. Finally, for studies of medications, the reviewers can write to the drug manufacturers that may have sponsored some trials.

Step 4. Selecting the articles

This step consists of applying the selection criteria devised in Step 2 to the articles found in Step 3. The important point of this step is to avoid any suspicion that articles were rejected because they failed to show what the reviewer wanted, rather than not meeting the criteria. The best way to ensure this is to have two or more independent reviewers evaluate each article; ideally, one of the reviewers doesn't even know the question that's being asked, just the criteria. It's a good idea for the reviewers to start off by rating about 10 or so articles that definitely would not be included in the meta-analysis, such as those looking at a different type of gambling or a different population than those targeted by the review. Any disagreements should be discussed to determine why they occurred, and to clear up any ambiguities in interpreting the criteria. This should be repeated until their reliability is over 90%. At this point, the reviewers can turn their attention to the articles that may be included in the meta-analysis.

If more than 50 articles were located, though, it may be too onerous a task for two people to review each study. In this case, 10 to 20 articles can be randomly selected for both reviewers to look at. If their agreement is high, then it's fairly safe to divide the remaining articles between them, thus reducing the workload. The authors should report the level of agreement for the articles evaluated in common (likely using Cohen's kappa; Norman & Streiner, 2000, pp. 96-97) and how discrepancies were resolved.

There is one other point to bear in mind in selecting articles. Some authors feel that if they've gone to all the trouble to do a study, the world should know of their findings, over and over again. Unfortunately, it's not unusual to find the same study in different journals (with minor modifications to slip under the copyright rules). Another ploy is to publish with, say, 50 subjects, and then publish again after the sample size has grown to 75. A third tactic, used in multi-centre trials, is for each study centre to publish its own results, in addition to one paper giving the global results. If you suspect that this is the case, use only the last publication, or the one that has the findings for all of the centres; otherwise, the study will have a disproportionate weight (and the authors will have been rewarded for their dubious tactics).

Step 5. Appraising the articles

Step 4 addressed the minimal criteria for an article to be included in the meta-analysis. However, there are studies and then there are studies. In other words, not all research is created equal. A study can be flawed in many ways, and allow biases to creep in. A useful framework was presented by Cook and Campbell (1979), who differentiate between the internal and external validity of a study. Internal validity refers to how well the study itself was conducted, and the degree to which we can believe the findings; external validity relates to the ability to generalize the results from the study sample to the population at large. Issues that pertain to the internal consistency of a study cover areas such as the number of people who drop out before the end, the adequacy of the outcome measures, how well the treatment and control groups were matched at the beginning, the fidelity with which the intervention was carried out, blinding of the raters and the proper analysis of the data. When we look at external validity, we are concerned primarily with issues of subject selection and reproducibility of the treatment. For example, were the participants self-defined gamblers or were diagnostic criteria applied; were people with co-morbid disorders screened out or entered into the trial; were they primarily community dwellers or a sample of convenience of university psychology students? As regards the intervention, was a manual used so that all therapists followed the same protocol; were sessions videotaped to ensure adherence to treatment guidelines; and most importantly, was it an intervention that could easily be applied by practitioners in the field? Unfortunately, in many instances, there is a trade-off between internal and external validity, so that the better the design, the less the study resembles what is actually done in the real world (Streiner, 2002). The reviewers have to decide at what point violations of internal and external validity jeopardize the study.

Over the years, a number of checklists have been developed that allow people to evaluate the design and execution of a study, although they are almost all restricted to RCTs (see Moher et al. (1995) for a good review; and Jÿni, Altman & Egger (2001) for a critique of the scales). Perhaps the most widely used are those of Jadad et al. (1996) and Chalmers et al. (1981). Scales such as these can be used in two ways: to set a minimum criterion for a study to be included in the meta-analysis, and to assign a score to each study to reflect its methodological adequacy. In Step 11, we will see how we can use this score to determine if the results of studies are influenced by research design issues. If the Jadad or similar scales are used, the reviewers should independently rate the same 10 to 20 articles and the reliability should be reported using an intra-class correlation

(Streiner & Norman, 2003).

Step 6. Abstracting the results

Key elements of each study now have to be abstracted from the articles and entered into a spreadsheet, or a program specifically designed to do meta-analyses; a review of available programs is in Stern, Egger and Sutton (2001). What should be abstracted? At an absolute minimum, it would be the data necessary to calculate effect sizes (described in Step 7). First, this would include the final sample size in each group (that is, the initial sample size minus those who dropped out, were lost to follow-up, or died). Second, if the outcome is measured on a continuum (e.g. the South Oaks Gambling Screen (SOGS); Lesieur & Blume, 1987), then the mean score and standard deviation (SD) for each group at the end of treatment is required; if the outcome is dichotomous (e.g. have or have not betted within the last 12 months), then we need the numbers in each category. These criteria are so minimal that you would expect every published article to meet them. However, as an example of the fatuousness of this belief, in preparing a meta-analysis of anti-depressants (Joffe, Sokolov & Streiner, 1996), we found that only 9 of 69 (13.0%) of articles reported even these elements (Streiner & Joffe, 1998). In many cases, we had to photo-enlarge graphs and estimate mean values.

One decision that should be made before the articles are abstracted is which outcome measure to use when two or more are reported. It isn't kosher to use more than one outcome result (although there are exceptions that we'll discuss in a moment), because that would result in studies contributing more to the overall findings simply because they used more measures. There are two options. The first is to pool all of the outcomes into one measure: how to do this is discussed by Rosenthal and Rubin (1986). The second, more common method is to select one outcome. For example, in our meta-analysis of anti-depressants (Joffe et al., 1996), we decided *a priori* that, if both were given, we would select objective measures over subjective; and for the possible objective indices, we devised a hierarchy of which scales would be preferred over others. The exception to the one study-one outcome rule is when the meta-analysis itself is addressing a number of outcomes. For example, a meta-analysis of CBT versus drug therapy for escape gamblers may look at effectiveness, measured by how many times the person has gambled within a six-month period, and acceptability of the treatments, evidenced by the drop-out rate. Within each outcome area, though, only one measure per study should be used.

What else to abstract depends on what else you think may influence the magnitude of the findings from one study to the next. For example, if the meta-analysis is focussing on drug treatments for people with gambling problems, it may be worthwhile to code the type of medication and the average dose. A meta-analysis of CBT may code the average number of sessions, whether the therapists were professionals or students, whether there was a treatment manual they had to follow, and so forth. If you believe that the treatment is changing over time (hopefully, improving), then the date of publication would be a variable of interest. Finally, if a methodology checklist was used, its score should be recorded for each study.

Step 7. Calculating effect sizes

One major problem in combining various studies is that they often use different outcome measures. Some may look at the number of times a person has gambled in a six-month period, others may use one year; some report frequency of gambling, others focus on the amount of money wagered. Yet other studies may rely on scores on a questionnaire, such as the SOGS. The issue is to find a common yardstick, so that the results are all reported using the same metric. For therapy trials, the most commonly used measure is the effect size (ES).

Effect size

ES comes in two main flavours: effect size for continuous measures (e.g. SOGS scores ranging from 0 through 22) and for dichotomous ones (e.g. treatment success or treatment failure). The general form for continuous measures is

$$ES=\frac{\overline{X}_T - \overline{X}_C}{SD}$$

[1]

where $\overline{X}_T$ is the mean for the treatment group; $\overline{X}_C$ that of the control group; and SD is the standard deviation. When calculated in this way, the ES expresses the results in standard deviation units. For example, if the outcome in one study was time since the person last gambled, and its SD was four months, then a two-month difference between the group means would yield an ES of 0.5. (i.e. half the standard deviation) A different study could have used an outcome of the amount gambled, with an SD of $2,000. If the group means differed by $500, then that would be equivalent to an ES of 0.25. In this way, these two studies, using very different outcomes, can be directly compared with one another, and their results pooled with those from other studies.

Another advantage of this ES is that it allows us to use the table of the normal curve to figure out what proportion of people in the treatment group did better than the average person in the control group.

Where the formulae differ is what to use for the SD. One option, called Cohen's *d* (Rosenthal, 1994), is to use the pooled SD of both groups. Its advantage is that it uses all of the data and so is a more stable estimate. Its disadvantage is that it uses all of the data, so that if the intervention affects not only the mean but also the SD of the treatment group, the resulting ES will be biased. Glass's Δ (Glass, 1976) gets around this problem by using only the SD from the control group. The downside is that it uses only half of the data, and so is less efficient than Cohen's d.

For dichotomous outcomes (e.g. treatment success or failure), the usual indices of ES are the odds ratio (OR) for case-control studies; and the relative risk (RR) for RCTs and cohort studies. Those who want to understand the important differences between the concepts of odds ratios and relative risk can find a useful discussion on-line at http://bmj.com/cgi/content/full/316/7136/989

Because the OR and RR have some undesirable properties (e.g. there's a lower bound of 0 but no upper bound; and no intuitive relationship between an OR or RR and its reciprocal, although both express the same result; see Streiner, 1998), we most often use the logarithm of the OR or RR, which removes these problems.

Step 8. Checking for publication bias

In Step 3, we mentioned that there is a strong bias against submitting articles that failed to show significant results (Cooper et al., 1997) and an equally strong bias against publishing those that have been submitted (Begg & Berlin, 1988). The exclusion of negative studies leads to biased results and will overestimate the overall effect size. Perhaps the most widely used method for determining if publication bias may be operating is to draw a funnel plot (Light & Pillemer, 1984), a fictitious example of which is shown in Figure 1. Some index of the ES (e.g. the ES itself, or the log of the odds ratio) is on the *X*-axis and an index of the study's size on the *Y*-axis. This could be the sample size itself, or the reciprocal of the standard error (if we used the standard error itself, the funnel would be upside down). The rationale for the plot is that smaller studies have less precise estimates of the true ES, and so their results would vary from one study to the next. With larger sample sizes (or smaller standard errors), the estimates of the ES should cluster closer to the true ES, resulting in the pyramidal shape.

If publication bias is present, then the funnel is asymmetrical, as in Figure 2, because the non-significant studies have been excluded. Needless to say, this only works if there are a large number of studies (“large” is one of those statistical terms that means, “I can't give you an exact number”).

Rosenthal (1979), who coined the term “the file drawer problem,” derived a formula for estimating how many studies with negative results (i.e. with ESs of 0) have to be stuck away in a filing cabinet in order to negate the conclusions of a meta-analysis. If the number is large (same definition as before) in comparison to the number of trials that were found, then we can relatively safely say that it's unlikely there would be this many, and the results would hold. On the other hand, if the number is small (again, the same definition), we should be far more cautious because even a few unpublished, negative findings would wipe out the overall effect.

Step 9. Testing for homogeneity

It's important to determine how similar their results are before combining the results of the individual studies. In statistical jargon, the issue is the homogeneity of the findings. If all of the studies report ESs in the same ballpark, then we are more confident that they're all reporting the same phenomenon and that the pooled ES is a good estimate of what's really going on. On the other hand, if there is a lot of variability from one study to the next, then it's possible that we're trying to compare apples with oranges. That is, the studies may differ so much from each other in terms of the sample, the intervention, or other aspects of the design, that it may not make sense to combine them. Also, the results of testing for heterogeneity (the opposite of homogeneity) may dictate how we analyze the data (which we will look at in Step 12).

Testing for homogeneity

The most general test for homogeneity, which can be used for any index of ES (Hardy & Thompson, 1998) is

$$Q = \Sigma_{w_i} (\hat{\Theta}_i - \overline{\Theta})^2$$

[2]

where w_i is a weight for each study, which we will discuss in the next step; $\hat{\Theta}_i$
is the ES for Study i, and $\overline{\Theta}$
is the mean ES. Q is distributed as χ^2 with $k - 1$ degrees of freedom, where k is the number of studies.

If it appears as if one or two studies are outliers, in that their ESs are much larger or much smaller than all of the others, it may be worthwhile removing them and seeing if *Q* becomes non-significant. If so, the final analyses should be done with and without such studies, to test the degree to which they may be influencing the conclusions.

Step 10. Combining the studies

Once the ES has been derived for each study, we have to summarize (or “pool”) them in some way to get an estimate of the mean; that is, an overall estimate of the effectiveness or ineffectiveness of the intervention. The simplest way is to add them up and divide by the number of ESs; after all, that *is* what we mean by the “mean.” But (and there's always a “but”), this method gives equal weight to studies that looked at 10 patients and those that looked at 1,000. Intuitively, it seems obvious that we should give more credit to larger studies, because their results are more stable. We do this by weighting each effect size (which we denote by *θ*) by some index of the sample size.

Weighting the studies

The weight that is applied to each study is the reciprocal of its squared standard error(SE):

$$w_i = \frac{1}{SE(\Theta_i)^2}$$

[3]

Since the standard error is strongly influenced by the sample size, larger studies will have a smaller SE, and therefore a larger weight. The weighted ESs are then averaged using the formula:

$$\bar{\Theta} = \frac{\Sigma_{wi}\Theta_i}{\Sigma_{wi}}$$

[4]

For more about calculating standard errors for different types of ES, see Deeks, Altman and Bradburn (2001).

Step 11. Looking for influential factors

Even if the test for homogeneity is not statistically significant, there will be some degree of variability among the ESs. We can now look to see what accounts for the differences. Basically, we run a multiple regression, where the ESs are the dependent variable, and the design features we coded in Step 6 are the predictors.

For example, we (Joffe et al., 1996) found that how the diagnosis of depression was made had a major influence on the results. Studies that used strict, research-based criteria tended to have larger ESs than studies that relied on the judgement of a single psychiatrist. In studies of treatments for gambling, possible predictors could be the number of therapy sessions, whether a person is self- or other-referred, the quality of the research (based on one of the scales mentioned in Step 5), the presence or absence of other co-morbid conditions, and so on. Bear in mind, though, that the number of predictor variables you can have is limited by the number of articles. The rough rule of thumb is that there should be around 10 articles for each predictor (Norman & Streiner, 2000); so, if you found 20 articles, you should have no more than two predictors.

Step 12. Selecting the type of analysis

There are two general approaches to analyzing the results of meta-analyses: a fixed-effects model and a random-effects model. We will not go into the mathematics of the differences between the two (for which we can all give a heartfelt thanks), but rather discuss the issue on a conceptual level. A fixed-effects model assumes that there is a “true” effect size that underlies all of the studies, and that they differ among each other only because of sampling error. A random-effects model makes the assumption that there is a population of effect sizes, from which the studies in the meta-analysis are a random sample (Hedges & Vevea, 1998). The reason that this distinction is important is that, in many situations, the two types of analyses yield different results. A fixed-effects model is less conservative and may give statistically significant results in some situations when a random-effects model will not.

So, which model is it appropriate to use and when? A fixed-effects model is appropriate if we want to draw conclusions about the particular set of articles in the meta-analysis. That is, it does not allow us to say anything about studies that may have been missed or those that will be done in the future. On the other hand, a random-effects model is perhaps more realistic in two regards. First, by saying that there is a population of effect sizes, the model acknowledges the fact that studies differ with respect to the sample, the procedures used and other aspects of the design, all of which may result in different findings. Second, it allows us to generalize from this particular set of articles to studies of this phenomenon in general; studies we did not include and studies yet to be done. Note that this distinction is not based on the tests of homogeneity we discussed in Step 9, but only on the type of inferences we wish to make (Erez, Bloom & Wells, 1996; Hedges & Vevea, 1998). In most situations, and especially if the test of homogeneity is significant, we would be wise to go with a random-effects model.

Summary

Meta-analysis is neither the answer to all of the world's ills, nor the greatest scourge visited upon humanity since the Black Plague. Carefully done and used intelligently it can be a very powerful tool for synthesizing the literature in a field, sometimes bringing clarity where there had been confusion. This is particularly true when the effect we are looking for is small, and even very large trials may not have sufficient power to tease out a definitive conclusion. For example, there were six relatively large trials looking at the effects of ASA following a myocardial infarct. Because the outcomes were dichotomous and the event rate rare (fortunately for us; unfortunately for the researchers), none showed statistically significant results. However, a meta-analysis showed that by combining these studies, there was a clear advantage to taking ASA (Canner, 1983); and a similar conclusion was made regarding beta-blockade, again on the basis of individually non-significant studies (Peto, 1987).

On the other hand, meta-analyses do not do away with the need for judgement and decision making. Two people reviewing the same literature may use different criteria in deciding which articles to include and which to discard; how the effect size should be calculated; which type of analysis to use; and so forth. Consequently, meta-analyses should not be regarded as “truth,” only as a better approximation of it than individual studies. Used in this way, and tempered by clinical experience, they can assist the clinician in deciding what may work and what won't for a particular patient.

References

Begg, C.B.. Berlin, J.A.. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society*, 151, 419-463.

Canner, P.L.. (1983). Aspirin in coronary heart disease: Comparison of six clinical trials. *Israel Journal of Medical Science*, 19, 413-423.

Chalmers, T.C.. (1991). Problems induced by meta-analysis. *Statistics in Medicine*, 10, 971-980.

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

Cook, T.D.. Campbell, D.T.. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

Cooper, H.. DeNeve, K.. Charlton, K.. (1997). Finding the missing science: The fate of studies submitted for review by human subjects committee. *Psychological Methods*, 2, 447-452.

Deeks, J.J.. Altman, D.G.. Bradburn, M.J.. (2001). Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In Egger, M.. , Smith, G.D.. & Altman, D.G.. (Eds.), *Systematic Reviews in Health Care: Meta-analysis in Context* (2nd ed.) (pp. 285–312). London: BMJ Books.

Dickersin, K.. Hewitt, P.. Mutch, L.. Chalmers, I.. Chalmers, T.C.. (1985). Perusing the literature: Comparison of MEDLINE searching with a perinatal trials database. *Controlled Clinical Trials*, 6, 271-279.

Erez, A.. Bloom, M.C.. Wells, M.T.. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. *Personnel Psychology*, 49, 275-306.

Furukawa, T.A.. Streiner, D.L.. Hori, S.. (2000). Discrepancies among megatrials. *Journal of Clinical Epidemiology*, 53, 1193-1199.

Gilbody, S.M.. Song, F.. Eastwood, A.J.. Sutton, A.. (2000). The causes, consequences and detection of publication bias in psychiatry. *Acta Psychiatrica Scandinavica*, 102, 241-249.

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

Hardy, R.J.. Thompson, S.G.. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, 17, 841-856.

Hasselblad, V.. Hedges, L.V.. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, 117, 167-178.

Haynes, R.B.. McKibbon, K.A.. Walker, C.J.. Mousseau, J.. Baker, L.M.. Fitzgerald, D.. , et al. (1985). Computer searching of the medical literature: An evaluation of MEDLINE searching systems. *Annals of Internal Medicine*, 103, 812-816.

Hedges, L.V.. Vevea, J.L.. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3, 486-504.

Hollander, E.. Buchalter, A.J.. DeCaria, C.M.. (2000). Pathological gambling. *Psychiatric Clinics of North America*, 23, 629-642.

Ioannidis, J.P.A.. Cappelleri, J.C.. Lau, J.. (1998). Issues in comparisons between meta-analyses and large trials. *Journal of the American Medical Association*, 279, 1089-1093.

Jadad, A.R.. Moore, R.A.. Carrol, D.. Jenkinson, C.. Reynolds, D.J.. Gavaghan, D.J.. , et al. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary?*Controlled Clinical Trials*, 17, 1-12.

Joffe, R.. Sokolov, S.. Streiner, D.L.. (1996). Antidepressant treatment of depression: A meta-analysis. *Canadian Journal of Psychiatry*, 41, 613-616.

Jÿni, P.. Altman, D.G.. Egger, M.. (2001). Assessing the quality of randomized controlled trials. In Egger, M.. , Smith, G.D.. & Altman, D.G.. (Eds.), *Systematic Reviews in Health Care: Meta-analysis in Context* (2nd ed.) (pp. 87–108). London: BMJ Books.

Kamin, L.J.. (1978). Comments on Munsinger's review of adoption studies. *Psychological Bulletin*, 85, 194-201.

Lefebvre, C.. Clarke, M.J.. (2001). Identifying randomised trials. In Egger, M.. , Smith, G.D.. & Altman, D.G.. (Eds.), *Systematic Reviews in Health Care: Meta-analysis in Context* (2nd ed.) (pp. 69–86). London: BMJ Books.

Lesieur, H.. Blume, S.. (1987). The South Oaks Gambling Screen (SOGS): A new instrument for the identification of pathological gamblers. *American Journal of Psychiatry*, 144, 1184-1188.

Light, R.J.. Pillemer, D.B.. (1984). *Summing Up: The Science of Reviewing Research*. Cambridge: Harvard University Press.

Marshall, M.. Lockwood, A.. Bradley, C.. Adams, C.. Joy, C.. Fenton, M.. (2000). Unpublished rating scales: A major source of bias in randomised controlled trials of treatments for schizophrenia. *British Journal of Psychiatry*, 176, 249–252.

Modesto-Lowe, V.. Van Kirk, J.. (2002). Clinical uses of naltrexone: A review of the evidence. *Experimental & Clinical Psychopharmacology*, 10, 213-227.

Moher, D.. Jadad, A.R.. Nichol, G.. Penman, M.. Tugwell, P.. Walsh, S.. (1995). Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials*, 16, 62-73.

Munsinger, H.. (1975). The adopted child's IQ: A critical review. *Psychological Bulletin*, 82, 623-659.

Norman, G.R.. Streiner, D.L.. (2000). *Biostatistics: The Bare Essentials* (2nd ed.). Toronto: B.C. Decker.

Oakley-Browne, M.A.. Adams, P.. Mobberley, P.M.. (2001). Interventions for pathological gambling. Cochrane Database of Systematic Reviews, Issue 4. In: The Cochrane Library, 4, 2001. Oxford: Update Software. Abstract available: http://www.cochranelibrary.com/Abs/ab001521.htm

Peto, R.. (1987). Why do we need systematic overviews of randomized trials. *Statistics in Medicine*, 6, 233-240.

Rosenthal, R.. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, 86, 638-641.

Rosenthal, R.. (1994). Parametric measures of effect size. In Cooper, H.. & Hedges, L.V.. (Eds.), *The Handbook of Research Synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rosenthal, R.. Rubin, D.B.. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, 99, 400-406.

Slavin, R.E.. (1986, November). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. Educational Researcher, 5–11.

Smith, M.L.. Glass, G.V.. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32, 752-760.

Sobal, J.. Stunkard, A.J.. (1989). Socioeconomic status and obesity: A review of the literature. *Psychological Bulletin*, 105, 260-275.

Stern, J.A.C.. Egger, M.. Sutton, A.J.. (2001). Meta-analysis software. In Egger, M.. , Smith, G.D.. & Altman, D.G.. (Eds.), *Systematic Reviews in Health Care: Meta-analysis in Context* (2nd ed.) (pp. 336–346). London: BMJ Books.

Streiner, D.L.. (1998). Risky business: Making sense of estimates of risk. *Canadian Journal of Psychiatry*, 43, 411-415.

Streiner, D.L.. (2002). The two Es of research: Efficacy and effectiveness trials. *Canadian Journal of Psychiatry*, 47, 347-351

Streiner, D.L.. Joffe, R.. (1998). The adequacy of reporting randomized controlled trials in the evaluation of antidepressants. *Canadian Journal of Psychiatry*, 43, 1026-1030.

Streiner, D.L.. Norman, G.R.. (2003). *Health Measurement Scales: A Practical Guide to Their Development and Use* (3rd ed.). Oxford: Oxford University Press.

Wachter, K.W.. (1988). Disturbed by meta-analysis?*Science*, 241, 1407-1408.

Walters, G.D.. (2001). Behavior genetic research on gambling and problem gambling: A preliminary meta-analysis of available data. *Journal of Gambling Studies*, 17, 255-271.

Figures

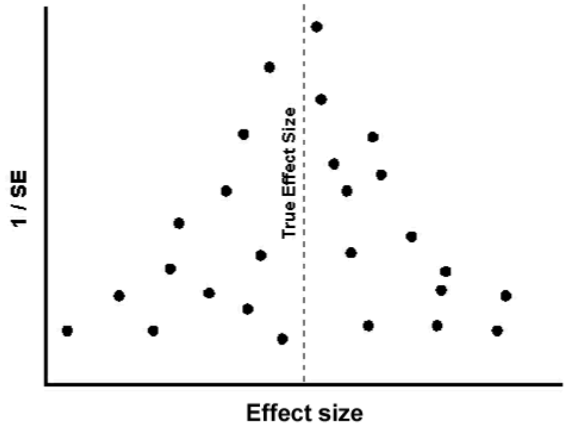

Figure 1.
A fictitious funnel plot with no publication bias

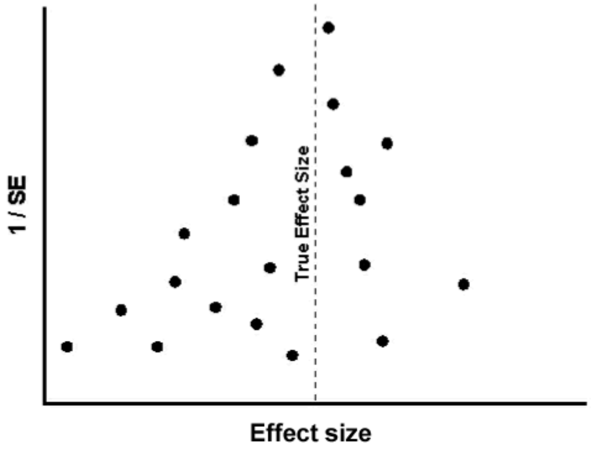

Figure 2
The same plot showing publication bias.

Tables

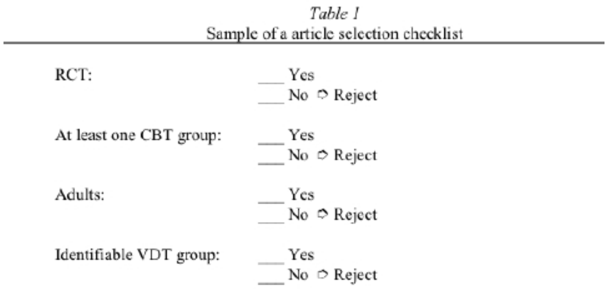

Table 1
Sample of a article selection checklist

RCT: ___ Yes
___ No ➲ Reject

At least one CBT group: ___ Yes
___ No ➲ Reject

Adults: ___ Yes
___ No ➲ Reject

Identifiable VDT group: ___ Yes
___ No ➲ Reject

Table 1

Article Categories:

- Feature