

Meta-analyses of Clinical Drug Trials—Gold Standard Reviews or Statistical Alchemy?

E S Y Chan,*^{BSc, BVMS, PhD}

Abstract

Renewed emphasis on the critical appraisal of published evidence has in turn led to critical appraisal of the way clinical studies have been reviewed. Reviews perform the vital function of summarising a wealth of research information in an efficient and trustworthy manner, but traditional narrative reviews have fallen short in this area, not because they are narrative, but because they have allowed reviewer and publication bias to creep in under the guise of expert authority. Meta-analysis or systematic review emphasises the use of a review protocol based on logical principles to minimise bias, together with qualitative and appropriate quantitative summaries of the data. Like any review, the perceived impact of a meta-analysis is dependent on the quality of the reviewed studies, therefore randomised clinical trials are the favoured substrate for the meta-analysis of drug efficacy because of their potentially high internal validity. The importance of meta-analysis as a tool for clinical drug evaluation has prompted this selective account of its nature, theory, practice and interpretation, with particular emphasis on meta-analysis as a review methodology rather than a statistical methodology, the critical distinction between design and result homogeneity and the need to explore issues of heterogeneity rather than gloss over them by using some statistical shortcut.

Ann Acad Med Singapore 2000; 29:610-5

Key words: *Exchangeability, Fixed-effects model, Heterogeneity, Homogeneity, Individual patient data meta-analysis, Nested meta-analysis, Random-effects model, Study precision, Systematic reviews, Weighted average*

Introduction

Informal review and synthesis of scientific evidence has been practised for as long as experiments have been replicated. The idea of quantitatively combining the evidence from different samples can be traced back to the early 1900s, when Karl Pearson analysed the preventive effect of serum inoculations against enteric fever.¹ In 1976, Glass² who was working in the social and psychological sciences contrasted a “rigorous alternative” he called meta-analysis to the “casual, narrative discussions of research studies which typify our attempts to make sense of a large volume of research literature”. He defined it as: “... the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings”.

With the rise of evidence-based medicine³ and its emphasis on using evidence from the very best study designs, it was logical that the requirement of rigour in conducting individual studies should be extended to the way these studies were combined in a review. The meta-analytic approach to reviewing and synthesising social study data was seized upon to render similar service to Medicine, and

over the past 10 years, meta-analysis of evidence from clinical research, especially that of randomised clinical trials (RCTs), has taken on an increasingly important role. This paper discusses some of the issues, misconceptions and errors within meta-analysis, in the hope that the reader will leave with a clearer and wider perspective of this important topic.

Systematic, not Statistical

While Glass’ definition emphasised the narrower interpretation of meta-analysis as a collection of statistical techniques, the real strength and value of meta-analysis lies in its advocacy of a systematic process to the conduct of a review. Unfortunately, the clinician sees and focuses far more on the statistical side, which admittedly has a certain seductive mystique. Two extreme attitudes stem from the misconception that meta-analysis is more about statistics than about a system of reviewing. The sceptic dismisses meta-analysts as statistical alchemists. In his eyes, they are trying to turn “base” studies into “golden” guidelines. The gull on the other hand, worships at the shrine of statistics, believing that it can even resurrect dead studies. Both

* Head and Co-ordinator

Division of Evidence-based Medicine

NMRC Clinical Trials and Epidemiology Research Unit, Ministry of Health, Singapore

Address for Reprints: Dr Edwin Chan Shih-Yen, Clinical Trials and Epidemiology Research Unit, 10 College Road, Singapore 169851.

E-mail: edwin@cteru.gov.sg

attitudes are wrong and divert attention from the real issue, which is that a review, just like an RCT, needs to be conducted according to a protocol which minimises bias. The quality of a meta-analysis should be judged first on the evidence that a valid protocol based on logical principles has been followed, there being no necessity that it should conclude with a quantitative combination of the results. With this perspective, we can perhaps see why the name “systematic review” is a more accurate appellation.⁴

What are these logical principles that a meta-analysis protocol should embody? Most authorities⁵ would agree that a valid systematic review protocol should at least declare:

1. A clear aim.
2. Criteria for studies to be included which specify the kind of study design, patient type, disease, treatment and relevant clinical outcomes to be measured.
3. A strategy for searching out relevant studies.
4. The type of data to be extracted from each study.
5. The method for assessing individual study quality.
6. The comparisons to be made including criteria for defining comparable studies.
7. The method for combining studies if this is judged appropriate.

The role of the meta-analysis protocol is thus analogous to that of the study protocol in a RCT of drug efficacy. The same requirement for internal validity that applies to the conduct of RCTs should also apply to the conduct of reviews of RCTs. Committing to do a review accordingly will go a long way in countering the criticisms of reviewer bias and lack of methodological information that are the Achilles’ heel of narrative reviews. It is for this reason of being systematic rather than statistical, that meta-analysis could be viewed as the gold-standard approach to conducting reviews.

Combining Qualitative and Quantitative Appraisal

The traditional narrative review has tended not only to be unsystematic but biased towards qualitative appraisal. In general, a meta-analysis will include both qualitative and quantitative aspects. For example, the description of the available studies in terms of their design and methodological strengths and weaknesses is important for conveying an impression of the quality of studies to be found “out there”. More importantly, it forms the logical basis for claiming that studies are similar enough to combine. If the studies are deemed not combinable, reasons should be given together with a simple tabulation of the individual study characteristics and results. The results of combinable studies (typically relative risks, odds ratios or mean differences) will be pooled to form a weighted average which represents a statistically more precise and (hopefully) more clinically

reliable estimate of the true effect of treatment. The drug development process has need for both of these aspects. In the planning phase, there will be a need to summarise existing research on the performance of current therapies and the need for a new drug. Regulatory submissions call for integrated summaries of efficacy, safety and risk-benefit. I would suggest that a meta-analysis is an ideal format to meet both these requirements.

Information Quality

Most meta-analyses are based on the published literature, and therefore hostage to both the quality of reporting and conduct of those studies. Ambiguities in the report may sometimes be ameliorated by contacting the authors for further information, that is if they have not moved, forgotten or lost the results or refused to reply. Alternatively, when patient follow-up is ongoing over an extended period such as is common in survival trials, data can be collected on individual patients through a collaborative network of trialists. This so-called “individual patient data” (IPD) meta-analysis⁶ differs from a multi-centre trial in that the trials in an IPD meta-analysis were designed independently of each other which in effect makes them more heterogeneous. Nevertheless, it is superior to meta-analysis of the published literature because it permits checking of the raw data (as opposed to having only the published summary data), and provides a means of updating the results. Perhaps of more importance, getting trialists directly involved in the review gives them a powerful incentive to provide quality and timely information. In reality, an IPD meta-analysis is much more costly and time-consuming, thus the majority of meta-analyses will continue to rely on published data. A potential step-up from IPD meta-analysis would be planning trials with a future meta-analysis in mind (prospective meta-analysis). Of course the most satisfactory situation with regards to data pooling will still be when all studies use a standard protocol and are part of a single large multi-centre trial.

Study Combinability—Design Homogeneity

The preoccupation with the statistical facet of meta-analysis has led to an unfortunate misunderstanding of the correct way to decide on whether study results should be statistically combined. Combination should be based on an *a priori* specification of common study design features, before the results have been looked at, and not on the *post hoc* result of a statistical test of homogeneity. To be at all interpretable, an average outcome must represent a sample or population that can logically be thought of as homogeneous. That is, it must make sense to think of the group of studies as a single collection with clear common features whatever their other differences may be. It makes no sense to say that the average superiority of drug A over

B from a group of studies is Y, unless we can also specify the common kind of diseased patients treated, a common pair of interventions, a common measure of clinical improvement and a common study design. I will refer to the situation where a uniform set of design features which justify study combinability as design homogeneity, others have called the opposite condition clinical heterogeneity.⁷

Study Combinability—Result Homogeneity

Result (statistical) homogeneity refers to the situation where the results (and not the design) of a group of studies are judged to be “similar” as determined by applying a statistical test, often misleadingly called a test of homogeneity.⁸ A *P* value greater than 0.05 (a statistically non-significant result) is taken to mean that the studies are homogeneous enough to combine. This would be a wonderfully convenient way to decide on combinability if it were not so utterly wrong.

To base combinability on the *P* value of a hypothesis test would be to commit two mistakes. The first being that demonstrating result homogeneity logically implies that the studies must be homogeneous in design and therefore combinable. This is the *post hoc* fallacy, for while we do expect design-homogeneous studies to also evince result homogeneity, the converse is certainly not true. After all it is the study design features that “cause” the result and not vice versa. A related but even more mistaken belief is that demonstrating result homogeneity is the sole requirement for justifying combinability. The second fallacy is thinking that the statistical test of homogeneity actually tests for homogeneity, alas it does not. The reason is that the test procedure starts with the assumption that the results are homogeneous (the null hypothesis). It should be obvious that no test can validly prove any hypothesis for which it also requires as an assumption. If this were not true, deductive reasoning would not exist and mathematicians throughout the centuries would have been wasting their time and ours. The homogeneity test is in fact a misnomer — it is actually a test of heterogeneity⁷ because when $P \leq 0.05$ we can reject the hypothesis of result homogeneity and accept the alternative hypothesis of result heterogeneity. Even so, result heterogeneity has not been “proven” in the sense of a mathematical proof but is only strongly suggested [by demonstrating that the result would be highly unlikely (with probability equal to the *P* value) under the assumption of result homogeneity]. One final word, and that is to warn the reader that the test has low statistical power to demonstrate heterogeneity.⁷ This means that the ability to reject the null hypothesis when there really is heterogeneity in results is weak. It seems that we are not permitted to take much comfort in a non-significant heterogeneity test result! For it cannot be interpreted as demonstrating design or

result homogeneity, neither is the test adequately powered to detect result heterogeneity.

The Decision to Combine

What then is the use of the heterogeneity test? In the case when we cannot argue there is design homogeneity, studies should not be combined and performing the test is pointless. When we can justify design homogeneity, getting a non-significant *P* value is what we would expect and deciding to combine is easy. But what if the test indicates result heterogeneity even though we believe there is design homogeneity? Assuming that the calculations are correctly done, the first thing to do is thoroughly review the study characteristics to see if something that might explain the result heterogeneity has been missed. This is generally only practicable for the meta-analyst and not the reader who does not have all the studies close at hand. The investigation may show up previously unnoted study features that could have caused the heterogeneity, this exploratory process^{5,7} is a vital part of the meta-analysis procedure and should be described. The studies may be regrouped into smaller more homogeneous sets based on these newly identified common design features and re-analysed. Now one hopes to see a non-significant test result.

If the source of result heterogeneity cannot be found or the hoped for non-significant test result does not materialise, then one has the option of deciding not to combine the studies but to report them separately.⁹ This is an option, which I believe is not used as often as it should be. However, if the belief in design homogeneity is strong, meta-analysts may use a statistical method (random-effects model) to incorporate the extra variability observed into the final pooled result. It should also be borne in mind that the observed result heterogeneity may not in fact be true (a type I error).

Combining Results—The Weighted Average

Assuming that study pooling is justified, the basic approach to combining results is to form a weighted average of all the individual results.⁸ The results pooled should not be the *P* values of the significance tests but rather the direct numerical estimates of the comparative outcome benefit. In drug RCTs reporting a binary outcome, this means averaging the relative risk, odds ratio or risk difference of a cure/response and for those reporting continuous outcomes, the mean difference in response between the two trial groups. The rationale for weighting each study result differently comes from our knowledge that despite design homogeneity, studies differ in sample size and this will affect the precision of each study's result. All things being equal, a result from a larger study has a greater precision or smaller variance (narrower confidence

interval). It seems to make a lot of sense that the average result should somehow reflect the precision with which each component study result was estimated. When the study precision is defined as the inverse of the study variance, the “fixed-effects” precision-weighted average of all study results is calculated by multiplying each result by its precision, taking their sum and dividing this by the total of all the study precisions.

The term “fixed-effects” needs further explanation.¹⁰ It is a translation of our assumption of design homogeneity into a mathematical assumption. First, we are to assume that all the studies are randomly sampling subjects from the same population of patients, all of whom have potentially the same average drug response (patient exchangeability). Each study design is also assumed to be equivalent (design exchangeability) with all others, in the sense that, using study design A in place of design B (but with the same sample size as in study B), would give exactly the same results as study B save for random error. Another way of saying this is that design A generates the same sampling distribution of study results as design B when based on the same sample size. Mutual exchangeability further implies that all study designs generate the same sampling distribution of results when based on the same sample size. In reality, because different studies also have different sample sizes, the sampling distributions will have differing variances, but still have the same mean. Our aim is to estimate the mean drug efficacy of our exchangeable patients by estimating the common mean of these sampling distributions, which is the fixed-effects weighted average.

The Random-effects Model

Earlier on, reference was made to a random-effects model that may be used to average studies homogeneous in design but heterogeneous in results. As stated before, the principle of forming the weighted average remains the same. The only difference is that each study’s variance is increased (and hence the precision is decreased) by some constant amount from the original value under the fixed-effects model. This is seen as a way of taking into account the observed extra result variability. The main effect is to produce a wider confidence interval for the weighted average than one gets using the fixed-effects model.¹¹ Many reviewers choose to use the “random-effects” weighted average over the “fixed-effects” average because they want to express greater uncertainty about the true treatment effect.

The random-effects model differs from the fixed-effects model by allowing the sampling distributions of study results to have different means. We could postulate this as coming about in several ways. I think the most plausible would be to imagine that differing populations of patients

(non-exchangeable) are being sampled and studied in each trial using non-exchangeable study designs. Each study outcome will belong to its own sampling distribution with differing means. These means are further assumed to be normally distributed. This “distribution of means” has itself a mean which is what we are estimating when we calculate the random-effects weighted average.

While this model may appear more flexible (for it includes the fixed-effects model as a special case), the supposed advantage over the latter is questionable because of the difficulty of interpreting the random-effects weighted average. The “distribution of means” is an assumed form about an artificial population made up of non-exchangeable patients, who have been lumped together as a convenient way of modelling inter-study variability. The meaning of the average of this population is anybody’s guess because it does not refer to a logically homogeneous group. People have a convenient habit of ignoring logical difficulties when it suits them and I have identified three pragmatic attitudes when reviewers report results using the random-effects model:

1. The results are homogeneous, but reviewers believe that the random variability is wider than assumed by the normal distribution and therefore seek a way of getting wider confidence intervals than the fixed-effects ones.
2. The results are heterogeneous, but reviewers, still believing in overall design homogeneity, seek to factor in the observed variability.
3. The results are heterogeneous and the designs are heterogeneous, yet the reviewers still want to combine the results.

Choosing Models

There is no clear choice between fixed-effects and random-effects models. Rather I would like to submit some personal views. If the studies are design and result homogeneous, the use of the fixed-effects model seems a reasonable choice based on the logical consistency between the model’s assumptions and the principles for combining studies. Using the random-effects model to incorporate result heterogeneity (regardless of whether it was statistically significant) appears to be also defensible, if the argument for design homogeneity is convincing. The numerical estimate of the random-effects average would then be interpreted as an estimate of a “variability adjusted” fixed-effects average. In this way, the problem of interpreting the random-effects average is arguably circumvented. This still begs the question whether the random-effects approach is the most appropriate way to factor in variability.¹⁰ If patients are demonstrably non-exchangeable, then the random-effects model can be used to estimate the means of the separate patient populations¹⁰ because these are the

only interpretable summaries, but the random-effects average itself has no useful interpretation. Finally, the difficulty with both models is the untenable assumption that the component studies are a random sample from their respective populations. The general attitude to this difficulty is reminiscent of our approach to generalising results from RCTs — we know the study subjects are not a random sample but we interpret the result as if they were.

Hierarchical or “Nested” Meta-analysis

Given the difficulty of justifying the combination of one set of studies, it boggles the mind when meta-analyses are recklessly combined to produce a magnificent nested analysis much like my daughter’s Russian doll. An infamous example of this was the recent systematic review on mortality risks from colloid versus crystalloid solutions.¹² The authors first performed three separate meta-analyses on studies treating three different types of injuries (note non-exchangeability of patients). One of the pooled averages was not statistically significant, the other two were barely significant. They then combined these three meta-analyses and did an overall meta-analysis. Not surprisingly, the final result was even more significant. I see big problems with such a practice aside from the operational decisions on the extent of such hierarchical pooling, for the reviewers are in real danger of producing an average result which applies to an ill-defined clinically irrelevant population. If there are problems interpreting the average result of just one meta-analysis, it is not clear how one explains the even more nebulous results from a hierarchy of combinations. There is also the little problem of justifying the statistical model for combining meta-analyses.

Old Wine in a New Skin?

In his presidential address to the Royal Statistical Society, Nelder¹³ urged statisticians to abandon the word meta-analysis, for he did not see anything special about the statistical methodology that distinguished it from the rest of statistics. This should be comforting to those who harbour a suspicion that meta-analysts are Dr Franksteins who have mutated statistics into a monstrous teller of lies. It is not necessarily so, for the fixed-effects weighted average model is the same method we have long used to summarise results from a stratified RCT or case-control study.¹⁴ If we take the narrow view of meta-analysis, Nelder’s argument is persuasive but we should take the broader view of meta-analysis as a distinct methodology for doing good reviews—what a pity it has such an ugly name.

Clinical versus Statistical Significance—Unlearned Lessons from RCTs

It hardly seems worth mentioning that the reader should

be looking for a result that is clinically significant and not simply statistically significant, for exactly the same concern applies to the interpretation of individual trials. However, the danger of just looking for statistical significance is much more acute in meta-analysis. Why? For the simple reason that statistical significance is so much more easily achieved. Remember that many meta-analyses involve the pooling of small studies that may in themselves not be statistically significant. The process of combining studies has the guaranteed effect of making the confidence interval of the weighted average narrower than any single study’s interval because of the increased number of subjects. If enough studies are combined, the confidence interval of the weighted average will eventually become narrow enough to be statistically significant. Therefore demonstrating statistical significance in itself, while often yearned for and pursued diligently, is when found, really quite trivial.

It is sobering to imagine how many mistakes we might have avoided by noting important parallels in the principles for the conduct and interpretation of RCTs that apply equally well to meta-analysis, I have discussed three of these at some length and now summarise them below:

1. The need to use a valid protocol which minimises bias.
2. The need to compare similar patients using similar protocols.
3. The need to look for clinical significance rather than statistical significance.

Great Expectations

In the early days there was the great expectation fueled by the illusory sophistication of statistical methods that meta-analysis could produce once-for-all answers to questions in which the evidence was conflicting. The resulting uncritical pooling of data by misguided reviewers has resulted in much justified criticism but in some cases has led to the baby being thrown out with the bathwater.¹⁵ I have reviewed the mistakes, misunderstanding and unresolved issues in some key areas. Yes, the bathwater is dirty, but I hope the reader is convinced that the baby is worth saving! There are many other areas such as publication bias, sensitivity analysis, choice of outcome measures, meta-regression for which it would be impossible to do justice in the allotted space and therefore I have chosen to remain silent. Given the uncoordinated way we have been doing RCTs, it is my belief that the number of combinable studies for most meta-analyses of drug trials will be small, say from five to ten at most. The vast majority of meta-analyses will also not surprise us with dramatic and heretofore unexpected evidence of efficacy. What the well-conducted “garden-variety” meta-analysis can be expected to do is to provide more insight into the problem, so that in the spirit of “the best is yet to be” better RCTs can

be designed. For all of us, better understanding of the aim, principles and methods of meta-analysis will enable us to return safely from the other side of the looking-glass, equipped now with a more refined tool which will allow us to do most of the things we want to do, whilst restraining us from doing some things which perhaps we should not.

REFERENCES

1. Pearson K. Report on certain enteric fever inoculation statistics. *BMJ* 1904; 3:1243-6.
2. Glass G V. Primary, secondary and meta-analysis of research. *Education Rec* 1976; 5:3-8.
3. Sackett D L, Rosenberg W M C, Gray J A M, Haynes R B, Richardson W S. Evidence-based medicine: What it is and what it isn't [editorial]. *BMJ* 1996; 312: 71-2.
4. Dickersin K, Higgins K, Meinert C L. Identification of meta-analyses. The need for standard terminology. *Contrib Clin Trials* 1990; 11:52-66.
5. Anello C, Fleiss J L. Exploratory or analytic meta-analysis: should we distinguish between them? *J Clin Epidemiol* 1995; 48:109-16.
6. Clarke M J, Stewart L A. Obtaining data from randomised controlled trials: how much do we need for reliable and informative meta-analyses? In: Chalmers I, Altman D G, editors. *Systematic Reviews*. London: BMJ Publishing Group, 1995:37-47.
7. Thompson S G. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994; 309:1351-5.
8. Petitti D B. *Meta-analysis, decision analysis and cost-effectiveness analysis*. New York: Oxford Univ Press, 1994:90-114.
9. Marsoni S, Torri V, Taiana A, Gambino A, Grilli R, Liati P, et al. Critical review of the quality and development of randomized clinical trials and their influence on the treatment of advanced epithelial ovarian cancer. *Ann Oncol* 1990; 1:343-50.
10. Normand S L T. Meta-analysis: Formulating, evaluating, combining and reporting. *Stat Med* 1999; 18:321-59.
11. Poole C, Greenland S. Random-effects meta-analyses are not always conservative. *Am J Epidemiol* 1999; 150:469-75.
12. Cochrane Injuries Group Albumin Reviewers. Human albumin administration in critically ill patients: systematic review of randomised controlled trials. *BMJ* 1998; 317:235-40.
13. Nelder J A. Statistics, science and technology. *J Roy Stat Soc* 1986; A149:109-21.
14. Selvin S. *Statistical analysis of epidemiologic data*. Oxford: Oxford Univ Press, 1996:234-9.
15. Feinstein A R. Meta-analysis: statistical alchemy for the 21st century. *J Clin Epidemiol* 1995; 48:71-9.