BMC Medical Research Methodology

Research article

Meta-analysis: Neither quick nor easy

Nancy G Berman*[1] and Robert A Parker[2]

Address: [1]Department of Pediatrics, Harbor-UCLA Medical Center, 1000 West Carson Street, Torrance, CA, USA and [2]Biometrics Center/E-GZ814, Beth Israel Deaconess Medical Center and Department of Medicine Harvard Medical School, 330 Brookline Avenue, Boston, MA 02215

E-mail: Nancy G Berman* - berman@gcrc.rei.edu; Robert A Parker - robert_parker@caregroup.harvard.edu

*Corresponding author

Published: 9 August 2002

BMC Medical Research Methodology 2002, **2**:10

Received: 22 May 2002
Accepted: 9 August 2002

This article is available from: http://www.biomedcentral.com/1471-2288/2/10

Abstract

Background: Meta-analysis is often considered to be a simple way to summarize the existing literature. In this paper we describe how a meta-analysis resembles a conventional study, requiring a written protocol with design elements that parallel those of a record review.

Methods: The paper provides a structure for creating a meta-analysis protocol. Some guidelines for measurement of the quality of papers are given. A brief overview of statistical considerations is included. Four papers are reviewed as examples. The examples generally followed the guidelines we specify in reporting the studies and results, but in some of the papers there was insufficient information on the meta-analysis process.

Conclusions: Meta-analysis can be a very useful method to summarize data across many studies, but it requires careful thought, planning and implementation.

Background

Meta-analysis has been defined as 'the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings.' [1] Although there has always been some controversy about its validity [2–8], meta-analysis has become increasingly popular as the number of studies with similar protocols has grown. By systematically combining studies, one attempts to overcome limits of size or scope in individual studies to obtain more reliable information about treatment effects.

A meta-analysis goes beyond a literature review, in which the results of the various studies are discussed, compared and perhaps tabulated, since it synthesizes the results of the individual studies into a new result. A meta-analysis also differs from a 'pooled data' analysis because the summary results of the previous studies, not the results on individual subjects, are combined for analysis.

Meta-analyses are fairly common in some fields of research and are relatively rare in others. In fact, a March 2002 Medline search revealed 1,610 articles with the keyword 'cancer', and only 19 with the keyword COPD and 41 with the keyword 'epilepsy' among the 9,055 articles indexed under meta-analysis. This may reflect a common belief that meta-analyses should be based on multiple clinical trials, which are very common in cancer studies and less common in other fields. However, a meta-analysis of small trials may provide sufficient information on treatment effects to avoid the delay and expense of a large-scale randomized clinical trial. If proper methods for selecting and combining studies are used, observational studies can also be included in a meta-analysis [9–11]. Guidelines for meta-analysis of observational studies have

been published [12]. In particular, meta-analysis may be used for combining studies in research where clinical trials would not be practical or would be unethical.

Because a meta-analysis does not involve human subjects or experimental animals directly, it is often considered an easy study that can be done with a minimum of effort and little attention is often paid to details of design and implementation. A valid meta-analysis, however, requires the same careful planning as any other research study. In this paper, we will focus on the important design issues underlying a meta-analysis: formulating the study question, identification of research studies, collecting and evaluating information about these studies, and extracting results. Simple methods for analyzing the data once it is collected are described briefly.

Methods

A meta-analysis is similar to a record review. A record review, sometime referred to as a chart review, is a clinical study in which all the information is derived from existing records, usually either hospital charts or a data base. Major problems in performing a record review involving identifying all the relevant records and then obtaining adequate data from these records. Exactly the same problems occur when performing a meta-analysis. In a meta-analysis, the individual studies are the subjects, so that locating study results is equivalent to identifying and locating patient records. Evaluating whether to include studies in the meta-analysis is equivalent to reviewing the patient's record and deciding whether to include the patient in the record review. Although data collection may appear simple, it requires as much, if not more, attention to detail than in an actual clinical study. Needed data may be missing, or, even worse, be ambiguous so that it is not clear what the actual data are. In a record review, this might require review of additional records, data on other hospital visits or access to the hospital or government data bases. In a meta-analysis, it is often necessary to review a number of separate publications to determine the "correct" data, or to contact investigators to resolve questions. Table 1 (see additional file: Table 1) lists the 7 major parts of a meta-analysis protocol and summarizes the similarities and differences between a meta-analysis and record review study.

The bulk of the time and effort in a meta-analysis is in the first four steps of the study. Therefore, in this paper, we will focus on these activities and discuss data abstraction and analysis only briefly.

Defining the Objectives of the Study

The first step is to identify the problem. This includes specifying the disease or condition of interest, the population of interest, the specific treatments or exposures being studied and the outcome measurements (efficacy, adverse reactions or both) being studied. Additional clinical or biological measurements of interest that might be potential confounders of the results should also be identified at this time, although other factors may be recognized during the evaluation or data collection phase of the meta-analysis.

The goals of the study should be defined at this stage. Meta-analyses attempt to meet one or both of two goals: summarizing the available data or explaining the variability between different studies. When the objective is to summarize the effects of an intervention, ideally all studies would have similar patient characteristics and the outcome measures would be consistent across studies. Thus, the summary measure resulting from the meta-analysis would reflect the effect of the treatment being studied. In practice, however, there is always variability between studies both in patient characteristics and in outcome measures, which is, of course, the primary motivation for performing a formal meta-analysis.

Alternatively, one might attempt to model the variability between studies to understand why different studies had different results [13,14]. This would suggest that as wide a range of studies should be included as possible. Frequently both objectives can be served in the same meta-analysis, by providing summary statistics of treatment or exposure effects in subgroups, often referred to as a sensitivity analysis, and modeling the heterogeneity across studies as a function of patient characteristics. Given the amount of work involved in performing any meta-analysis, we recommend that a meta-analysis attempt to meet both these goals.

Defining the Population of Studies Included in the Meta-Analysis

Inclusion and exclusion criteria for studies are as necessary in a meta-analysis as they are in clinical studies to safeguard against selection bias. These criteria need to be specified in the meta-analysis protocol, just as inclusion / exclusion criteria are specified in a clinical protocol. The criteria should follow immediately from the objectives of the study. An analysis aimed at providing a summary result in a specific subgroup of subjects will necessarily have more restrictive criteria than one designed to investigate heterogeneity. The inclusion criteria should address at least the following.

Type of study

Will the analysis be restricted to randomized clinical trials only, or will other designs be included? In our opinion, trials other than RCT are allowable. As we mentioned in the introduction, there is some disagreement on this subject [5,6,9,11,12,15,16]. The type of study may be used as a classification variable in a sensitivity analysis or in as-

sessing the quality of the study as described below. Olkin [14] has developed a hierarchy of strength of evidence based on the type of design, ranging from case reports (the weakest) to randomized trials (the strongest), with case-control studies somewhere in the middle. This hierarchy may serve as a guideline for inclusion of different designs. Example 3 [17], discussed below, shows how a valid meta-analysis may be based only on observational studies when clinical trials would not be ethical.

Patient characteristics

These include age, gender, ethnicity, presenting condition, duration of illness, and method of diagnosis. Again, this needs to reflect the goals of the study. We recommend that a meta-analysis be as inclusive as possible, e.g., you may exclude studies in children if they are known to be different from adults, but all studies in adults should be included without regard to age or gender. If you limit the meta-analysis to a very restricted population it will probably provide little new information. We recommend including population factors as covariates in estimates of the overall effect and examining their effects in a sensitivity analysis.

Treatment modalities

Allowable treatment type, dosage, and duration of treatment should be addressed. Since analysis of a given treatment is frequently the reason for the meta-analysis, most meta-analyses are limited to tests of a specific treatment or variations within that treatment, such as formulation, route of delivery, or dosage. Variations in treatment may be deliberately included for comparison or for sensitivity analysis. If controlled studies are included, then these criteria should specify the acceptable control groups, e.g., placebo control or a standard treatment. If more than one standard treatment can be used for the condition being studied, then the protocol must either define which ones will be acceptable or methods to address this possible source of variability between studies.

Outcome measures

Many studies have multiple outcome measures. The protocol for the meta-analysis should specify the outcome measure(s) of interest, including the allowable methods of measurement. For example, percent body fat may be measured by DEXA scan, by underwater weighing, by bioimpedance or by anthropometry. Different methods of measurement, if allowed, should be accounted for in the analysis. When allowed, the protocol must specify whether every study must report all of them, any one of them, or at least one or two specific ones. We recommend that the protocol allow only one or two primary outcomes to focus the analysis and avoid the impression of a fishing expedition. We realize that once the studies have been located and evaluated, investigators are reluctant to neglect any information and may want to perform additional analyses on other outcomes. However, it is likely that not all studies relating to these other outcomes will have been obtained since they were not the initial purpose of the analysis.

An important point, sometimes neglected, is that one should include only one set of results from a single study, even if multiple publications are available. Thus, it is necessary to have a method for deciding which paper(s) will be included. Most often it is reasonable to specify that this will be the latest paper published, or the paper with the most complete data on the outcome measures of interest. In any case, the decision rule needs to be specified in the meta-analysis protocol before reviewing results in the papers so that selection cannot be influenced by the results.

Locating Studies

Locating **all** studies is by far the most difficult and the most frustrating aspect of any meta-analysis but it is the most important step. A structured plan is necessary to manage the frequently large number of papers. Most meta-analyses begin with a search using the NLM Medline system. This should be supplemented by the use of other computerized indices, such as in-house research listings and reports from professional organizations. Properly done, this will give you most of the **published** articles relating to your topic.

There are several options for finding unpublished studies. Peer consultation, i.e. networking among your professional colleagues and contacting specific investigators who are known to be active in the area can help identify additional studies and investigators. Since abstracts are often not included in computer indexes, it is necessary to manually review special meeting issues of journals from the major professional organizations in the field. In addition, one might publish a request for information at meetings and in newsletters. References to 'unpublished data' in published studies must be followed up. The NIH and NLM maintain registries of clinical trials for some diseases; public non-profit organizations, such as the American Diabetes Association, can usually supply information about trials and other studies that they are sponsoring. The Cochrane Library contains a bibliography of controlled trials as well as abstracts of reviews of the effects of healthcare. The Internet is becoming increasingly important for identifying studies, using resources such as news groups or mailing lists.

In meta-analysis, one is particularly concerned with publication bias, i.e. the effect of failing to detect unpublished trials. The most common reason for not publishing is nonsignificant or uninteresting results [16,18]. Clearly, leaving out negative studies in any meta-analysis will sub-

stantially bias the result so that treatment will appear more effective than it actually is. Other factors associated with failure to publish include type of study, with clinical trials being most likely to be published, and funding source, with externally funded studies having a higher publication rate. Olkin [14] has noted that the results of very large studies are usually published whereas the publication of small studies may depend on timing, with early small studies having a higher chance of publications than later small studies. Other causes of publication bias include language restrictions [13,19] and imperfect search techniques. For the later, we recommend that investigators seek the support of the institutional librarians. Negative studies are more likely to be published in 'local' journals and not in the major international journals, therefore restricting language to English tends to exclude negative studies done in non-English speaking countries [19]. Unfortunately, it is not always possible to obtain a reliable translation of these papers. Decisions regarding inclusion of papers in a foreign language must be made before one begins attempting to locate studies. Even if the results of such papers are not included in the meta-analysis, the existence of such papers should be reported.

We recommend that all relevant studies be listed in the material and methods section or in an appendix, even when it was not possible to formally screen and evaluate them in the meta-analysis, so that the reader will be aware of the number of studies not included in the formal meta-analysis.

Screening and Evaluation

A quick review of the abstracts of the papers will eliminate those that are clearly not relevant to the meta-analysis or do not meet other criteria, such as study design, specific population, duration of treatment or date of the study. If the published material is just an abstract, there must be sufficient information to evaluate its quality. There must also be summary statistics to put into the meta-analysis, available either from the written material or in writing from the investigator. It is essential that when the available written information is insufficient for the meta-analysis that strenuous efforts be made to contact the principal investigator to obtain the needed information in order to reduce the effect of publication bias. This becomes even more important for material that has not been formally published, which can only be obtained from the principal investigator.

Assuming adequate information is available, each study should then be subjected to a structured review of the quality of the study. Table 2 (see additional file: Table 2) summarizes the major points that should be addressed in this evaluation. Although Table 2 is a model of an evaluation score sheet, it has not been formally tested or externally validated.

The items in part A, which address sources, require that the authors and institutions, etc., be known. These questions should be answered by raters not involved in the assessment of the methods, who would prepare a score sheet for each study giving only the answers to these questions. To assure an unbiased review, the items in part B should be assessed by raters who are blinded both to the authors and the results of the study. Personnel not involved in this part of the evaluation should prepare copies of the papers with the sources, results and other information that might indicate the authors or outcome removed for this step in the evaluation [7]. Although an investigator might feel that blinding is not feasible because of time or cost, lack of blinding potentially leads to major biases in the evaluation of studies and thus the extra effort is warranted [4,7,20–23]. Failure to blind the review could lead to biases similar to those in a record review when subjects are selected by investigators who are not blinded to the outcomes of interest. However, there may be some studies that are very well known, or the research area may be so small, that any suitable rater will know the authors and/or the results of the study. When this occurs, efforts should still be made to blind the study and have the study methods rated by individuals outside the area of interest, but with expertise in study design issues. Some investigators feel that the evaluation should also address whether the conclusions of the study are consistent with the data. We feel that, since the meta-analysis is based on the data and not the written conclusions, that this comparison would only be meaningful when the conclusions are so far-fetched as to cast doubt on the accuracy of the entire study. Since this case is very rare we have not included this item in Table 2. If conclusions are to be evaluated, the raters should still be blinded as to the identity of the investigators.

The methods used in these 'de-identified' papers should be evaluated by at least two raters, a content expert who is knowledgeable in the subject matter and a biostatistician or epidemiologist who can evaluate the analytic methods. We strongly recommend the use of a numeric quality score to summarize the results of the evaluation [5,20,22,24]. The two blinded raters will create a consensus score for quality which will be combined with the score for sources from the unblinded raters to give a final score for the study. The structure and items for the quality score must be specified in the meta-analysis protocol and should only be modified if some items are missing in all studies. The score should be based on the items in Table 2, but may be modified to suit the particular application. Most of the items in Part A and all of the items in Part B can be answered as 'Yes', 'No' or 'not applicable'. Since

many of the criteria ask whether or not specific items of information are in the paper, if one of these items is not in the publication and is not available from the investigator it is coded as 'No', not 'missing'. Thus, the number of 'not applicable' or 'missing' items should be small. For example, if the demographic information for some or all subject groups is not available, the response to the third question under "Study Subjects" in Table 2 would be 'No'. If this information was not given, you might not be able to answer the first question in Table 2 under "Controls", which would truly be missing information. We recommend that this type of missing data also be coded as 'No'. If critical information is missing, such as summary statistics, the investigators should contact the authors to try to get this information and, if it is not available, then the study should be excluded from the meta-analysis. The total score may be based on the sum of individual items by scoring 1 for yes and 0 for no (reversed when necessary for consistency) then expressing the total as the percent of the maximum possible. The latter will account for items coded as 'not applicable'. Alternately, the investigators may generate a summary score for each group of items and use either the sum or average of these. The former method gives equal weight to each item in the table, while the latter gives equal weight to the categories but the importance of each item varies with the number of items in a category.

The choice of method may depend on the proposed use of the quality score, which must also be specified in the protocol [5,20,22]. There is no consensus on this issue in the meta-analysis literature. Quality scores can be used in several ways: as a cutoff, with the meta-analysis including only studies above some minimum score; as a weighting value, with studies with higher quality scores being given more weight in the analysis; or as a descriptive characteristic of the study, used in explaining study variability and heterogeneity. We recommend that both the score based on items and the summary score be computed. The latter should be used to define a minimum value below which a study would be excluded from the analysis; the former can then be used to rank studies into three quality groups as a means of assessing heterogeneity between studies. If the number of studies is very large, then more than 3 groups may be used. If the number of studies is small then creating quality groups is not possible and only the summary score need be computed. We do not recommend using the quality score as a weighting variable because we feel it is too subjective. The distribution of quality scores may be addressed in the discussion section of the publication, as noted below.

There are other examples of quality scores in the literature. Chalmers [25] gives a comprehensive instrument for scoring a randomized clinical trial with detailed questions on every aspect of a trial, and assigns weights to the different sections. Although it is specifically aimed at RCT's, it could be adapted to other types of studies. Other authors use greatly reduced versions [11,22,26]. A bibliography of scales and checklists is given by Moher [27].

Data Abstraction

Data should be abstracted onto structured forms designed to capture relevant information in a concise, focused fashion. The protocol should specify the items, the information to be collected for each item and the format for collecting the items. Detailed instructions for data extraction and completion of the form should be prepared. For example, will age be recorded as the range, mean and standard error, or both? If recoding or estimation is required, e.g., estimates of the standard deviation from the range, the algorithm should be specified. Since there is not much consistency with respect to use of the standard deviation versus the standard error, the protocol should specify which should be used and how to convert from one to the other. Other criteria include whether rates will be entered as proportions (less than or equal 1) or as percents, whether natural logarithms or base 10 logarithms are used, etc. If data are incomplete, this should be noted on the form, although often if too much data are missing the study will be excluded from the analysis. Some texts give formulas for converting one test statistic to another [28]; if these are to be used they should be clearly defined in the instructions for data extraction. Ideally two individuals should independently abstract the results from every study and differences resolved by consensus. Some investigators recommend that these individuals be blinded to the authors of the paper [21], however, if the criteria for data collection are objective, blinding of abstractors, although not essential, is still desirable.

The data abstraction form should be headed with a study number, if blinding is to be preserved, or with the name of the study, the publication or source of data, the name and affiliation of the investigators, and the type of design. There should be descriptions of the study groups, including number of groups, size of group, age, gender distribution, diagnoses, treatments (including placebo), other treatment or descriptive variables, and length of treatment. The summary of the results can be quite extensive, including descriptive statistics for all groups and all outcome measures. Differences between groups in time, dosage, etc. should be included in the information on the data abstraction form. The test statistics should be identified by type, and given along with the p value, the sample size and the degrees of freedom when appropriate. Details of statistical models need to be given, listing other variables included in the model.

Space for comments should be included. Table 3 (see additional file: Table 3) is an example of a data collection

form for a meta-analysis of a clinical trial with 2 groups where one outcome is a continuous variable and a second outcome is a proportion. We strongly recommend pilot testing the data abstraction form on a few studies before defining a final format.

Data Analysis

Specific methods for data analysis in meta-analysis have been developed and are available in many texts and articles. The book by Hedges and Olkin [28] has been considered the standard text since its publication in 1985. The book by Wolf [29] is a more accessible reference at a basic level that gives formulas and procedures for simple studies. There are also many articles on how to do meta-analysis. The review article by Fleiss [30] is reasonably accessible for a non-mathematician who is not frightened of formulas. This section is intended to describe the general approach to the statistical analysis of the summary data that were collected. It is not intended to give instructions on how to do the actual computations. We recommend that you work with a statistician who is knowledgeable about meta-analysis for the formal analysis of the results.

The simplest method is to use a weighted average of the effects of each study. The analysis is usually based on a summary statistic derived from the study, often referred to as the effect size and a weight, which in most cases is the inverse of the variance of the effect size and is usually related to the sample size. The Q statistic [28] is a test of homogeneity between studies. A large value of Q indicates that there is significant heterogeneity between studies. Petitti [31] has observed that this test is conservative, and we recommend that the significance level for this statistic be set to 0.10 rather than the usual 0.05.

Some analysts might try to reduce the heterogeneity by limiting the meta-analysis to a smaller more homogeneous group of studies. However, this limits the scope of the meta-analysis and essentially throws away useful information. Models that incorporate and evaluate sources of heterogeneity are available. The standard approach is the random effects model developed by DerSimonian and Laird [32] and is well described in their paper. Fleiss [30] also includes these methods as alternative methods of analysis in his paper. Berlin [33] and Biggerstaff [34] expand on these methods. There is some controversy about the use of these models. Villar [35], in a study of 84 independent meta-analyses of randomized clinical trials, showed that in meta-analyses where there was a significant value of the Q statistic, the use of random effects models showed wider confidence intervals for the effects in question, but also showed a larger treatment effect. Petitti [31] does not see a clear rationale for choosing between a random or fixed effects model. The choice of analytic methods for any but the simplest situation requires input from a statistician experienced in meta-analysis.

Reporting and Interpretation

The protocol should indicate how the results of the meta-analysis will be presented. We recognize that, like the data analysis, this preliminary plan may be modified during the implementation of the study. The published meta-analysis should include a table containing all relevant descriptive information about each of the papers that are included in the analysis in a table. Ideally, all articles reviewed would be described, but this is not always practical, particularly if the number is large and many of them are irrelevant. Effect sizes, odds ratios, etc are considered results and may be presented in summary form or displayed for individual studies. Graphical displays are very helpful for showing the dispersion of single effects. All the examples described in the next section have good examples of graphical displays.

With respect to interpretation, we wish to emphasize two points. The first is the difference between statistical significance and clinical importance. Most of the techniques for meta-analysis will give a p-value, but these results must be interpreted in light of the other characteristics of the study. The distribution of quality scores for the studies also should be considered when deciding how much emphasis to give to the results. If the quality scores are skewed to low values, then this should at least be mentioned in the discussion as a possible shortcoming of the meta-analysis. If the studies were almost all of high quality, then this gives more credence to the results of the meta-analysis. The second point is the possible effect of unpublished studies on statistically significant results. Formulas for estimating the number of unpublished negative studies that would be necessary to cast doubt on the results of the meta-analysis have been developed [36,37]. Obviously, if this number were small, then the results of the meta-analysis are less credible. On the other hand if this number is large, the results of the meta-analysis are likely to be valid. We recommend that the investigator compute this value and include the results in the interpretation of the findings.

Results

We have selected 4 examples of meta-analyses to illustrate how other investigators have proceeded. The papers were selected to illustrate the differences in objectives and procedures that were described in the preceding sections. The examples are:

Example 1. Maximum androgen blockade in advanced prostate cancer: a meta-analysis of published randomized controlled trials using nonsteroidal antiandrogens [38].

Example 2. Glucosamine and chondroitin for treatment of osteoarthritis [39].

Example 3. Ischemic stroke risk with oral contraceptives [17].

Example 4. An examination of research design effects on the association of testosterone and male aging: results of a meta-analysis [25].

Table 4 (see additional file: Table 4) is a brief overview of the features of these articles as they relate to methods of meta-analysis. Not all articles gave full information on their methods and we may not agree with how some aspects of studies were done. However these papers illustrate the different approaches that can be taken in a meta-analysis.

Objectives

The first three articles attempted to estimate effects of treatment or exposure on survival (example 1), group differences on outcome scores (example 2) or relative risk (example 3). The objective of the fourth paper was to explore the effect of different approaches to study design on the study results. We do not address the actual clinical criteria used in these studies, as that is not relevant to our paper, but all papers reported very specific criteria as to population, treatment, laboratory methods, etc.

Study population

These different objectives influenced the selection of studies. Acceptable papers were limited to RCT's in the first and second articles. The third article was based on case-control and cohort studies, since randomized trials could not be done for this topic. All study designs were allowable in example 4 (the design characteristics in the title refer to population and protocol differences, not the type of study). All four meta-analyses used additional inclusion/exclusion criteria. The criteria in paper 1 addressed the subject and control treatments. Paper 2 required double-blind trials of at least 4 weeks duration. Paper 3 required clear differentiation of ischemic stroke from other strokes, and at least 10 subjects with a stroke. Paper 4 included all study types, but required that measures of association between age and T levels be included. All papers required that usable data be available.

Locating studies

Paper identification began with a Medline search in each study and all authors cited review articles as a source of other references. The authors in all articles said that they had contacted investigators for more data. We are sad to note that example 4 reported that of ten authors contacted, only one replied.

Screening and evaluation methods

Blinding was not used consistently. The authors of example 1 were the only ones who used blinded reviewers in selection and evaluation. The question of blinding was not addressed in papers 2 and 3. Paper 4 reported blinding for the quality evaluation. All papers reported developing a quality score, but the quality score in example 4 was only on the laboratory methods, and thus was not describing the overall quality of the study. In examples 1 and 2, the authors used the scoring system of Chalmers [25], while the others used simpler scoring systems developed for the study. Examples 1, 2 and 3 used two raters. Examples 1 and 2 required a consensus evaluation, but example 3 used the average of the two raters.

Paper 1 reported the overall quality scores for each study included in the analyses. The authors used the quality score to divide the studies into three groups and then compared the pooled relative risk in the three groups. In paper 2, the authors tested the effect of quality scores by comparing results in studies with scores above and below the median. Paper 3 did not describe any quality score and in paper 4 a score was developed for laboratory methods which was used for descriptive information only.

Data abstraction

Although all the articles reported or implied standardized data extraction, none of them provided examples or description of the format. Paper 1 reported using two blinded raters for data abstraction, but the others did not mention blinding at this point.

Data analysis

Examples 1–3 used weighted estimates of effect size to estimate an overall effect. In example 1, the investigators computed survival statistics and odds ratios for response and used the random effects models [32] to analyze the data. We note that in this paper the actual survival statistics had to be estimated from summary data for many of the studies. The simplest analysis was in example 2, where weighted estimates of the differences between treated and untreated groups were computed for pain and disability outcomes. In example 2, tests for heterogeneity were significant when all the studies were included, but became non-significant when one study was excluded. Subanalyses in example 3 looked at the contributions of other factors, such as smoking and alcohol to heterogeneity. The analysis in example 4 focused on differences between the testosterone-age correlation in different patient groups using visual display and multiple regression methods.

Reporting and interpretation

All papers gave details of all the studies that were included. All examples made good use of graphic display. Examples 1–3 included plots of the effect size for the selected

studies, while example 4 presented comparison plots of the regression lines for testosterone on age for different breakdowns of population characteristics or sampling times.

Conclusion

In this paper we have addressed the procedures for performing a meta-analysis, focusing primarily on the steps before data analysis. Meta-analysis cannot be thought of as a quick and easy way to pull a lot of studies together and come up with a publication, but like any other study, requires an appreciable investment of time in planning and implementation. As with any scientific procedure, there are areas of controversy. A primary one is the inclusion of studies other than RCT's. Although this has become more common in the past few years, it is still controversial. In example 3, there were no RCT's for this subject, although there had been many observational studies. It is important that, however you decide, you have a good reason for your choice.

Another area of controversy is in the homogeneity of the studies. We stated previously that when the purpose of a meta-analysis is to provide estimates of specific effects, then the criteria for inclusion would be more restrictive than if the objective were to model sources of variability. In practice, most meta-analyses combine both objectives. Even if the primary objective is simple estimates, there are population effects that should be investigated and discussed. One of the benefits of meta-analysis is that it may be used to extend conclusions beyond the frequently limited populations that are included in a single study. Moreover, given the effort that goes into identifying and evaluating papers, ignoring or rejecting valuable information is wasteful. Dickersin and others [18,34] point out that heterogeneity is not all bad. It improves the generalizability of the results of the meta-analysis. It may help to point out factors that influence the results of the outcome that were not observable in individual trials. If the effect is consistent even with discrepant studies, it strengthens the case for the causality of the treatment. If a meta-analysis is performed prior to beginning a new study, then heterogeneity may help the investigator improve his design by incorporating an understanding of these other factors.

We mention here that there is some leeway to modify your goals once original studies are reviewed, since only then will you know the extent of the data and which variables, other than the primary effect, have been measured. For example, you may find that there is an obvious grouping of study populations by age or ethnicity, and decide to investigate those effects. Such modifications should, of course, be restricted to observation of the available data and should not be based on the results of preliminary analysis.

In summary, a meta-analysis is an important and valuable tool for summarizing data from multiple studies. However, it is not an easy task and requires careful thought and planning to provide accurate and useful information.

Competing Interests

None declared.

Authors' Contributions

Nancy G. Berman and Robert A. Parker collaborated on the conceptualization and writing of this paper. Both authors read and approved the final manuscript.

Additional material

Additional Table 1

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2288-2-10-S1.pdf]

Additional Table 2

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2288-2-10-S2.pdf]

Additional Table 3

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2288-2-10-S3.pdf]

Additional Table 4

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2288-2-10-S4.pdf]

Acknowledgements

Supported in part by grant RR 00425 to the General Clinical Research Center at Harbor – University of California at Los Angeles Medical Center (Dr. Berman) and grant RR 01032 to the Beth Israel Deaconess Medical Center General Clinical Research Center (Dr. Parker) from the National Institutes of Health.

References

1. Glass GV: **Primary, secondary and meta-analysis of research.** *Educ Res* 1976, **5:**3-8
2. Lelorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F: **Discrepancies between meta-analyses and subsequent large randomized, controlled trials.** *N Engl J Med* 1997, **337:**536-542
3. Liberati A: **"Meta-Analysis: Statistical Alchemy for the 21st Century": discussion. A plea for a more balanced view of meta-analysis and systematic overviews of the effect of health care interventions.** *J Clin Epidemiol* 1995, **48:**81-86
4. Bailar JC 3rd: **The promise and problems of meta-analysis.** *N Engl J Med* 1997, **337:**559-561
5. Dickersin K, Berlin JA: **Meta-analysis: state-of-the-science.** *Epidemiol Rev* 1992, **14:**154-176

6. Thompson SG, Pocock SJ: **Can meta-analyses be trusted?** *Lancet* 1991, **338:**1127-1130
7. Chalmers TC: **Problems induced by meta-analyses.** *Stat Med* 1991, **10:**971-979
8. Kassirer JP: **Clinical trials and meta-analysis. What do they do for us?** *N Engl J Med* 1992, **327:**273-274
9. Spector TD, Thompson SG: **The potential and limitations of meta-analysis.** *J Epidemiol Community Health* 1991, **45:**89-92
10. Spitzer WO: **Meta-meta-analysis: unanswered questions about aggregating data.** *J Clin Epidemiol* 1991, **44:**103-107
11. Thacker SB: **Meta-analysis. A quantitative approach to research integration.** *JAMA* 1988, **259:**1685-1689
12. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB: **Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis of observational studies in epidemiology (MOOSE) group.** *JAMA* 2000, **283:**2008-2012
13. Moher D, Olkin I: **Meta-analysis of randomized controlled trials. A concern for standards.** *JAMA* 1995, **274:**1962-1964
14. Olkin I: **Meta-analysis: reconciling the results of independent studies.** *Stat Med* 1995, **14:**457-472
15. Stein RA: **Meta-analysis from one FDA reviewer's perspective.** *1988 Proceedings of the American Statistical Association Biopharmaceutical Section* 1988, 34-38
16. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR: **Publication bias in clinical research.** *Lancet* 1991, **337:**867-872
17. Gillum LA, Mamidipudi SK, Johnston SC: **Ischemic stroke risk with oral contraceptives: a meta-analysis.** *JAMA* 2000, **284:**72-78
18. Dickersin K, Min YI, Meinert CL: **Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards.** *JAMA* 1992, **267:**374-378
19. Gregoire G, Derderian F, Le Lorier J: **Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias?** *J Clin Epidemiol* 1995, **48:**159-163
20. L'Abbe KA, Detsky AS, O'Rouke K: **Meta-analysis in clinical research.** *Ann Intern Med* 1987, **107:**224-233
21. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC: **Meta-analyses of randomized controlled trials.** *N Engl J Med* 1987, **316:**450-455
22. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA: **Incorporating variations in the quality of individual randomized trials into meta-analysis.** *J Clin Epidem* 1992, **45:**225-265
23. Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R: **Meta-analysis of clinical trials as a scientific discipline I: control of bias and comparison with large co-operative trials.** *Stat Med* 1987, **6:**315-328
24. LaValley M: **A consumer's guide to meta-analysis.** *Arthritis Care Res* 1997, **10:**208-213
25. Chalmers TC, Smith H, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A: **A method for assessing the quality of a randomized control trial.** *Control Clin Trials* 1981, **2:**31-49
26. Gray A, Berlin JA, McKinlay JB, Longcope C: **An examination of research design effects on the association of testosterone and male aging: results of a meta-analysis.** *J Clin Epidemiol* 1991, **44:**671-684
27. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S: **Assessing the quality of randomized controlled trials: an annoted bibliography of scales and checklists.** *Control Clin Trials* 1995, **16:**62-73
28. Hedges LV, Olkin I: *Statistical Methods for Meta-Analysis. London: Academic Press* 1985
29. Wolf FM: *Meta-Analysis: Quantitative Methods for Research Synthesis. Newbury Park: Sage Publications* 1986
30. Fleiss JL: **The statistical basis of meta-analysis.** *Stat Methods Med Res* 1993, **2:**121-145
31. Petitti DB: **Approaches to heterogeneity in meta-analysis.** *Stat Med* 2001, **20:**3625-3633
32. DerSimonian R, Laird N: **Meta-analysis in clinical trials.** *Control Clin Trials* 1986, **7:**177-188
33. Berlin JA, Laird NM, Sacks HS, Chalmers TC: **A comparison of statistical methods for combining event rates from clinical trials.** *Stat Med* 1989, **8:**141-151
34. Biggerstaff BJ, Tweedie RL: **Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis.** *Stat Med* 1997, **16:**753-768
35. Villar J, Mackey ME, Carroli G, Donner A: **Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: comparison of fixed and random effects models.** *Stat Med* 2001, **20:**3635-3647
36. Gleser LJ, Olkin I: **Models for estimating the number of unpublished studies.** *Stat Med* 1996, **15:**2493-2507
37. Rosenthal R: **The "file drawer problem" and tolerance for null results.** *Psychol Bul* 1979, **86:**638-641
38. Caubet JF, Tosteson TD, Dong EW, Naylon EM, Whiting GW, Ernstoff MS, Ross SD: **Maximum androgen blockade in advanced prostate cancer: a meta-analysis of published randomized controlled trials using nonsteroidal antiandrogens.** *Urology* 1997, **49:**71-78
39. McAlindon TE, LaValley MP, Gulin JP, Felson DT: **Glucosamine and chondriotin for treatment of osteoarthritis: a systematic quality assessment and meta-analysis.** *JAMA* 2000, **283:**1469-1475

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/2/10/prepub