Meta-Analyses of Clinical Trials

Henry S. Sacks

Clinical Trials Unit, Mount Sinai Medical Center, New York, New York, USA

Meta-analysis has become increasingly popular in recent years. Evidence of this is an article in the New York Times in 1994, which attempted to explain meta-analysis to the general public as follows: "A meta-analysis aims at gleaning more information from existing data by pooling the results of many smaller studies and applying one or more statistical techniques. The benefits or hazards that might not be detected in small studies can be found in a meta-analysis that uses data from thousands of patients" (1). Dr. Thomas C. Chalmers, one of the pioneers of meta-analysis, defined it as "the systematic analysis of data gathered in multiple research projects. Applied to the clinical trials field, it is the process of evaluating quality and combining the results of multiple randomized controlled trials" (T.C. Chalmers, personal communication).

The purposes of meta-analysis include the following: (a) to increase statistical power for primary endpoints and for subgroups; (b) to resolve uncertainty when reports disagree; (c) to improve estimates of size of effect; (d) to answer new questions not posed at the start of individual trials; and (e) to bring about improvements in the quality of the primary research. These functions are particularly applicable to randomized controlled trials because they are often undersized (2).

We conducted a survey in 1987 of 86 meta-analyses of randomized controlled trial reports in the English language literature and updated it in 1991 (3,4) (Table 1). The English language medical literature (from January 1966 to October 1986 for the first survey, and updated through July 1990) was searched for papers that pooled the results of controlled clinical trials. Papers were found in Current Contents and by computer searches of the National Library of Medicine (NLM) and Bibliographic Retrieval Services, Inc. (BRS) data bases, by looking for reviews of specific subjects and using the terms "meta-analysis," "pooled" or "pooling," and "combined" or "combining" in title, abstract, or, where available, full text. Other sources of papers included references in papers found by the above methods and by personal communication.

Our criteria for inclusion of papers in this analysis were that data from more than one study must be combined, and at least one of the studies pooled must be a
Table 1. Comparison of quality features among meta-analyses

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number (%)</td>
<td>Number (%)</td>
<td>Number (%)</td>
</tr>
<tr>
<td></td>
<td>adequate</td>
<td>adequate</td>
<td>adequate</td>
</tr>
<tr>
<td>Prospective design</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protocol</td>
<td>5 (13)</td>
<td>2 (3)</td>
<td>13 (22)</td>
</tr>
<tr>
<td>Literature search</td>
<td>10 (25)</td>
<td>24 (36)</td>
<td>40 (69)</td>
</tr>
<tr>
<td>List of trials analyzed</td>
<td>37 (93)</td>
<td>56 (85)</td>
<td>54 (93)</td>
</tr>
<tr>
<td>Log of rejected trials</td>
<td>4 (10)</td>
<td>11 (17)</td>
<td>24 (41)</td>
</tr>
<tr>
<td>Treatment assignment</td>
<td>38 (95)</td>
<td>17 (26)</td>
<td>46 (79)</td>
</tr>
<tr>
<td>Ranges of patients</td>
<td>13 (33)</td>
<td>12 (18)</td>
<td>36 (62)</td>
</tr>
<tr>
<td>Ranges of treatment</td>
<td>20 (50)</td>
<td>25 (38)</td>
<td>39 (67)</td>
</tr>
<tr>
<td>Ranges of diagnosis</td>
<td>18 (45)</td>
<td>21 (32)</td>
<td>34 (59)</td>
</tr>
<tr>
<td>Combinability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Criteria</td>
<td>17 (43)</td>
<td>26 (39)</td>
<td>39 (67)</td>
</tr>
<tr>
<td>Measurement</td>
<td>5 (13)</td>
<td>17 (26)</td>
<td>27 (47)</td>
</tr>
<tr>
<td>Control of bias</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selection bias</td>
<td>0</td>
<td>0</td>
<td>7 (12)</td>
</tr>
<tr>
<td>Data-extraction bias</td>
<td>0</td>
<td>0</td>
<td>7 (12)</td>
</tr>
<tr>
<td>Interobserver bias</td>
<td>0</td>
<td>7 (11)</td>
<td>11 (19)</td>
</tr>
<tr>
<td>Source of support</td>
<td>17 (43)</td>
<td>14 (21)</td>
<td>16 (28)</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical method</td>
<td>22 (55)</td>
<td>40 (61)</td>
<td>45 (78)</td>
</tr>
<tr>
<td>Statistical errors</td>
<td>15 (38)</td>
<td>31 (47)</td>
<td>38 (66)</td>
</tr>
<tr>
<td>Confidence intervals</td>
<td>14 (35)</td>
<td>27 (41)</td>
<td>49 (84)</td>
</tr>
<tr>
<td>Subgroup analysis</td>
<td>28 (70)</td>
<td>39 (59)</td>
<td>45 (78)</td>
</tr>
<tr>
<td>Sensitivity analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality assessment</td>
<td>9 (23)</td>
<td>10 (15)</td>
<td>15 (26)</td>
</tr>
<tr>
<td>Varying methods</td>
<td>6 (15)</td>
<td>12 (18)</td>
<td>25 (43)</td>
</tr>
<tr>
<td>Publication bias</td>
<td>3 (8)</td>
<td>11 (17)</td>
<td>24 (41)</td>
</tr>
<tr>
<td>Application of results</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caveats</td>
<td>37 (93)</td>
<td>31 (47)</td>
<td>41 (71)</td>
</tr>
<tr>
<td>Economic impact</td>
<td>0</td>
<td>2 (3)</td>
<td>3 (5)</td>
</tr>
</tbody>
</table>

randomized controlled trial. Each paper was evaluated independently by two investigators using a scoring sheet that lists what we consider to be the important elements of a meta-analysis. The evaluators were blinded as to the name of the journal and the authors of the papers.

RESULTS OF THE SURVEYS

We believe that the important qualities of any meta-analysis can be divided into six major areas: study design, combinability, control of bias, statistical analysis, sensitivity analysis, and problems of applicability. We divided the papers into three time periods, selected to give roughly equal numbers in each: 1955-82, 1983-86, and 1987-90.
Study Design

In meta-analysis, as in any other form of research, it is important to try to make the process as rigorous and as well defined as possible.

Protocol

As with any scientific endeavor, the questions to be answered, the criteria for inclusion in the study, and the methodology to be used should be established beforehand. There has been a slight improvement, but still only a minority of papers gave clear evidence that the study was conducted according to a predetermined protocol or research plan. Many more may have followed protocols, but it was not apparent to the reader. If retrospective meta-analyses are to be converted into prospective research, we believe writing of a protocol is an essential first step.

Literature Search

Because a valid meta-analysis should include as many relevant trials as possible, the authors should provide details of their search procedures. At present, it is insufficient to rely solely on computer literature searches, as they may yield fewer than two-thirds of relevant trials (5,6). A computer search can be supplemented by consulting Current Contents, reviews, textbooks, or experts in the particular field of study and by reviewing the references of the trials found. The proportion of papers clearly using such exhaustive searching methods has increased.

List of Trials Analyzed and Log of Rejected Trials

The report of a meta-analysis should provide a list of the trials analyzed, and most of the reports in all three periods did so. Just as important is an enumeration of the relevant trials excluded and the reasons why. This is analogous to the log of excluded patients in a clinical trial. We believe it is vital for the reader to be aware of any information that was not used, because the meta-analyst may have had a preconception or bias as to how the result should come out. The proportion of papers reporting excluded trials has steadily increased, but even in the latest period fewer than half did so.

Treatment Assignment

The most important question bearing on the validity of the data pooled is the method of treatment assignment in the primary study. It has been shown that results of trials using historical controls are more likely to favor the new treatment than
results of the same treatment tested in randomized controlled trials (7). The proportion of papers using data only from randomized trials appears to have declined in recent years.

Ranges of Patients, Diagnoses, and Treatments

In order for the reader to judge the validity and generalizability of a meta-analysis, data should be provided on the patients, diagnoses, treatments, and endpoints in the original studies. It is not possible to provide more than broad general outlines here, but specific rules can, and should, be developed for each particular meta-analysis. The ranges of patient characteristics in all the trials analyzed—age, sex, relevant socioeconomic data, other diseases, etc.—should be included. Details on these items were given in only a minority of the meta-analyses in the first two time periods. The ranges of treatments should also be defined. Did all patients in all trials receive the same, or similar, treatment? In meta-analyses of drug treatments, were trials combined that used the same drug or the same class of drug? Dosage and route, as well as frequency and duration, of treatment in all the trials to be pooled should be available to the reader. This criterion was variably fulfilled in the papers analyzed. Similarly, data should be presented on the range of diagnoses in the pooled trials. Were diagnostic criteria the same in all trials? What stages or grades of disease were included? Such details on this important question were reported in fewer than half the meta-analyses in the first two time periods but in more than two-thirds in the third time period.

Combinability

A major issue in pooling data is whether the results of the separate trials can be meaningfully combined, and this should be explicitly addressed by the meta-analyst, in sufficient detail for the reader to determine that a useful and clinically relevant result will be obtained.

Criteria

What criteria were used to decide that the studies analyzed were similar enough to be pooled? The meta-analyst should note any differences in the primary studies and discuss how these differences affect the conclusions of the meta-analysis. Fewer than half of the meta-analysts in the first samples and 59% in the third detailed their criteria for pooling.

Measurement

Related to the problem of combinability is the statistical issue of heterogeneity. In addressing questions of combining estimates from different studies, statisticians
distinguish between two possible models (8). In model 1, each study is considered to be a sample from the same population and provides an estimate of a single underlying true rate, with the differences resulting from experimental error (within study variability). In model 2, each study is considered to be from a different population, the rate varies from study to study, and their differences result from experimental error and from differences in the populations (between-study variability). There are several methods for deciding which of the above models is more appropriate (9). In either case, there are tests for homogeneity that help decide the degree of caution with which pooled results should be interpreted. Because of the heterogeneity of patients treated by practitioners who might be applying the results of meta-analyses, heterogeneity in the trials may not be so bad. Evidence of a statistical test for homogeneity was found in few of the meta-analyses in the first time periods but rose to nearly half in the most recent period.

**Control and Measurement of Potential Bias**

In performing a meta-analysis, as in any scientific endeavor, potential sources of unconscious bias should be controlled for where possible.

**Selection Bias**

To avoid bias in the selection and rejection of papers, the decision to include a paper should be made by looking only at its methods and not at its results, or looking at the two separately under coded conditions. This important source of bias was not reported in any of the meta-analyses reviewed for our first survey and in only 7% in the latest time period.

**Data Extraction Bias**

As with any other data-gathering process that requires interpretation, observers may disagree. When papers list a variety of subgroups, endpoints, exclusions, and so on, it is quite possible that readers may vary in how they extract the data from a particular study. The ideal way to control for this type of bias is to have the data extracted by more than one observer, each of whom is blinded to the various treatment groups through a coded photocopying process, and then measure the interobserver agreement. In none of the meta-analyses in the first time period was such agreement reported, but such reporting appears to be increasing recently. The data extraction process was blinded in none of the early papers and in only 12% of those in the latest period.

**Source of Support**

We feel that it is useful to the reader to know who financed a study when deciding how much credence to give to its conclusions. Potential conflicts of interest do not
necessarily disqualify a study, but they should be clearly acknowledged. The source of support was specified in a minority of papers, and, if anything, the practice appears to have declined over time.

Statistical Analysis

This category deals with questions of statistical methodology.

Methods

We evaluated as "adequate" any recognized method of pooling except simple addition of successes across all trials to give an overall average, which was rated as "partial." An adequate method was used in well over half of the meta-analyses, and that proportion appears to have increased over time.

The most commonly used method was the Mantel–Haenszel test or a modification thereof. Other studies combined data by calculating a standardized average effect size; many also performed various types of regression analyses or significance tests, or a combination. A few papers used various other methods of pooling, and five papers did not specify the methods used. For further discussion of these various methods, see "Remaining Problems" below.

Statistical Errors

Fewer than half of the meta-analyses in the first two time periods showed an awareness of the potential problems of type I statistical errors (concluding that there is a difference when none exists) and type II errors (concluding that there is no difference when there is one). There has been some improvement in the last 3 years.

Confidence Intervals

It is often more useful to the reader to have an estimate with confidence intervals of the difference between the success rates of the treatments being compared than to have only the results of significance tests. Confidence intervals for major outcomes were given in fewer than half (43%) of the meta-analyses in the first periods, but this has risen to 84% recently.

Subgroup Analyses

One of the purposes of meta-analysis, as previously stated, is to increase the statistical power for subgroup analyses. Relevant subgroups were analyzed in the majority of papers.
Sensitivity Analysis

Depending on the test chosen, the same set of data can be combined to give different conclusions (9). Similarly, the results may vary depending on the overall quality of the primary trials and on whether certain trials, subgroups of patients, or other important variables are excluded or changed.

Quality Assessment

In a meta-analysis, the methodologic rigor and scientific quality of the papers to be combined should be assessed and considered in formulating recommendations (11,12). If the original methodology is poor, the resulting conclusion will be less reliable. Such features as the randomization process, the measurement of patient compliance, the blinding of patients and observers, the statistical analyses, and the handling of withdrawals in each primary study should be examined. This issue of quality was fully addressed in only a small proportion of the meta-analyses and was not even mentioned in nearly half. There was no evidence of improvement over time.

Varying Methods

Each meta-analysis should include in a sensitivity analysis data that show how the results vary through the use of different assumptions, tests, and criteria. This type of analysis was performed in only 15% of the meta-analyses reviewed in the first period but increased in use appreciably in the second and third.

Publication Bias

One of the criticisms sometimes made of meta-analysis is that there may be some unpublished studies that would contradict the results of published studies, and there is some evidence that negative studies are less likely to be published than positive ones (13,14). A simple method has been proposed for calculating the number of unpublished negative studies required to refute the published evidence (15), which is one possible measure of the strength of the published evidence. In only 8% of meta-analyses was publication bias considered in the first period, compared to 17% in the second and 41% in the third period.

Application of Results

Caveats

Once the results of the pooling process are available, the meta-analyst should attempt to put them into perspective, based on all of the considerations above. Does
the new treatment seem to be established as more effective than the old one for all patients, for some subgroups, and so on? Or should the conclusions be taken only as suggestions for future study? Such caveats were included in the discussion in varying proportions of the papers, with, if anything, a decline over time.

Economic Impact

In today's climate of financial constraints on health care expenditure, it is increasingly important to consider the economic impact of adopting new methods of treatment or diagnosis. Although some may consider this a topic for other studies, we were disappointed that fewer than 5% of the meta-analyses included a thorough analysis of economic impact.

Of the 23 individual items, 7.63 ± 2.84 (mean ± SD) were adequately addressed in the 40 meta-analyses published between 1955 and 1982, 6.80 ± 3.86 were adequately addressed in the 66 published between 1983 and 1986, and 11.91 ± 4.79 were adequately addressed in the 58 published between 1987 and 1990 ($F = 37.3$, $p < 0.001$). Twenty-two of the 58 papers published in the last time period referred to our initial survey or to other similar guides for meta-analysis (which suggested that the guidelines were being used).

Discussion

An important development in the recent history of meta-analysis was the use by Chalmers and colleagues of cumulative meta-analysis (16,17). Cumulative meta-analysis pools the available data at successive points in time to determine when the combined evidence reaches various significance levels. This technique was applied to the question of thrombolytic therapy for myocardial infarction and showed that, by the early 1970s, there was sufficient evidence to conclude that the treatment saved lives. However, many more clinical trials were conducted over the next decade, and it was not until the mid- to late 1980s that textbooks and review articles began to advocate thrombolytic therapy. Thus, if cumulative meta-analysis had been applied to the evolving clinical trial data, the usefulness of this treatment could have been determined much sooner.

REMAINING PROBLEMS

The greatest problem is defining the role of meta-analysis. When should it be attempted, and how should its results be used? We believe that the best way to answer questions about the efficacy of new treatments or diagnostic methods is to perform well-designed, adequately sized randomized controlled trials. Meta-analysis may have a role when definitive trials are impossible or impractical, when trials have been performed but the results are inconclusive or conflicting, or while awaiting
the results of definitive studies. Meta-analysis, like decision analysis, can give quantitative estimates of the weight of available evidence, which can be helpful in making clinical decisions. There is, however, a danger that meta-analysis may be used inappropriately or indiscriminately. As with many other types of analysis, the quality of the results depends on the quality of the input. Therefore, the question posed in each meta-analysis should be explicitly stated and clinically relevant.

Difficulty still exists in locating both meta-analyses and randomized controlled trials in the literature, because present literature-searching and indexing systems do not always distinguish primary studies and reviews from meta-analyses or randomized controlled trials from other clinical trials. Thus, it cannot be claimed that the papers found for these analyses are an exhaustive or representative sample. It is also quite likely that there are unpublished meta-analyses. Investigators will facilitate the process of recovery and integration of important clinical information if they insist on inclusion of the terms "randomized" and "meta-analysis" in titles and abstracts, so that indexing can be improved.

More attention also needs to be paid to statistical issues and to the advantages and disadvantages of the various pooling methods. A variety of statistical techniques have been developed for combining the results of separate studies. For example, there are several methods for combining the probability values or test statistics from individual studies (10). These methods, however, may not distinguish between small studies with large effects and large studies with small effects and do not yield an estimate of the size of the effect. The Mantel–Haenszel method (18) or a modification of that technique for combining separate $2 \times 2$ tables (19) (the Yusuf–Peto method) is becoming increasingly popular. These methods have several useful properties. They compare each treatment only with its own control and weight studies according to their sample size, and they can include a test for homogeneity as well as an estimate of the effect size. However, they may have some undesirable properties (see below). In the psychiatric literature, studies have been combined by computing effect sizes, defined as the mean difference between experimental and control groups, divided by the control group standard deviation (20). This method allows for the pooling of different endpoints because all findings are transformed into common units, but the conclusions may be difficult to interpret clinically; the validity of this process is thus open to question. A few papers have used multivariate methods or log-linear models to attempt to adjust for differences between studies.

Since publication of our first paper, there have been several important contributions to the statistical methods in meta-analysis. The validity of the Yusuf–Peto one-step method has been challenged (21). Simulations have confirmed this (22), and in those experiments the most valid method of determining variances has been that described by Robins et al. (23). When zero observations in one or both groups present a problem, the exact method, as modified and automated by Mehta et al. (24), is optimal. A Bayesian approach has been advocated (25). However, we encountered no published meta-analyses that used the last two methods.

Another important development in meta-analysis was a textbook entitled Effective Care in Pregnancy and Childbirth (26) with over 1000 individual meta-analyses.
These have not been included in our survey because they represent a special case—these studies were written on consignment by the editors and based on a data base of randomized controlled trials that had been collected in Oxford (27). Obviously, this approach needs to be replicated by the conduct of large numbers of meta-analyses in other fields of medicine.

Some meta-analysts believe that the only valid pooling of results should include the outcomes for all randomized patients regardless of how long (or even whether) they received the assigned treatment (the intention-to-treat method) (19). The authors of at least one paper apparently felt that dropouts and withdrawals should be excluded (exclusion method) (28). However, the data on withdrawals, dropouts, and so on are not always available. We believe that the results should be reported both according to the intention to treat and according to exclusion rules to facilitate evaluation of the differences. If there are none, there is no problem. If clinically important differences exist, the study may be difficult to interpret (29).

Because of the problem of publication bias, some meta-analysts choose to supplement their published data with unpublished trials or data. Unpublished results may be less reliable because they have not been found acceptable by peer reviewers and may not be collected with the same rigor or accuracy as published results (30). However, the potential problems inherent in unpublished results make it unclear whether both types of data should be given equal weight.

Greater uniformity in reporting meta-analyses would be helpful to readers. Many of the meta-analyses found were written in the standard format of scientific papers, with detailed methods and results sections, but several were editorials, leading articles, or letters to the editor with little detail on methodology. We believe that meta-analyses should be presented with sufficient information for readers to draw their own conclusions about the validity of the results.

With the growth in the number of published meta-analyses, there has been growing concern about their quality (31). Several national and international conferences have been held to develop and refine standards for meta-analysis (32,33). The Potsdam International Consultation on Meta-Analysis brought together proponents and critics of meta-analysis for a lively discussion, which was published in 1995 (33). It concluded with methodologic guidelines that should be consulted by anyone wishing to perform a meta-analysis. I quote here their guiding principles (34), which I endorse:

1. A systematic review must address a specific health care question. The question will determine which studies and data are relevant and how they should be synthesized.
2. Methodology must serve biology and the users and providers of health care. Therefore, a team with expertise in both the content area and methodology is ideally suited to conduct valid, useful systematic reviews.
3. A systematic review requires collaboration with the investigators who conducted the primary studies.
4. Systematic reviews are retrospective research and are potentially subject to many of the same biases that affect other retrospective studies; therefore, a systematic
META-ANALYSES OF CLINICAL TRIALS

review has to rely on both good randomized controlled trial methodology and good review methodology.

5. For several reasons, review methods may vary (for example, scarce resources may limit search strategies). Thus, the review methods actually employed must be described in detail.

6. The existence of unsatisfactory randomized trials, case-control studies, and cohort studies does not mean that any of these study designs should be abandoned; it means that they should be critically appraised, empirically studied, and improved. Overviews of observational studies require a great deal of methodologic development.

VALIDITY

A major unresolved issue is the validity of meta-analysis. Ingram Olkin, a statistician who has developed methodologies for meta-analysis, has proposed the following ranking of the validity of evidence from various study designs (35):

1. Anecdotal case reports.
2. Case series without controls.
3. Series with literature controls.
4. Analyses using computer data bases.
5. Case-control observational studies.
6. Series based on historical control groups.
7. Single randomized controlled trials.
8. Confirmed randomized controlled clinical trials including meta-analysis.
9. Meta-analysis with original data.

Others are much less optimistic. Alvan Feinstein complained that meta-analyses frequently violate scientific principles of precision, homogeneity, and consistency while focusing on the big process of aggregation but not on the overall small processes that produced the primary data (36).

We agree that a meta-analysis must be carefully done, according to the criteria discussed above, and that careful assessment of the primary data is a critical step. However, we believe meta-analysis can provide important and useful information on which to base clinical decisions. No question can ever be considered completely closed, and thoughtful physicians will always be searching for new data and revising their strategies.

META-ANALYSIS OF DIAGNOSTIC TESTS

Irwig et al. proposed methodologic standards for meta-analyses of diagnostic tests (37). They suggest that the process should include six steps:

1. Determine the objective and size of the meta-analysis. This should include a clear statement of the test of interest, the disease of interest and the reference standard
by which it is measured, and the clinical question and context; that is, is the objective to evaluate a single test or to compare the accuracy of different tests?

2. Retrieve the relevant literature. The literature retrieval procedure should be described with search and link terms given and inclusion and exclusion criteria stated.

3. Extract and display the data. Studies should be assessed by two or more readers. The authors should explain how disagreements between readers were resolved. The meta-analysis should give a full listing of diagnostic accuracy and study characteristic for each primary study.

4. Estimate diagnostic accuracy. The method of pooling sensitivity and specificity should take account of their independence. When multiple test categories are available, they should be used in the summary.

5. Assess the effect of variation in study validity on estimates of diagnostic accuracy. The relation between estimates of diagnostic accuracy and study validity of the primary studies should be examined for each of the following design characteristics: appropriate reference standard, independent assessment of the test or tests and reference standard, avoidance of verification bias. In comparative studies, were either all of the tests of interest applied to each patient, or were patients randomly allocated to the tests? Are analytic methods used that estimated whether study design flaws affect diagnostic accuracy rather than just test threshold?

6. Assess the effect of variation in the characteristics of patients and test on estimates of diagnostic accuracy (generalizability). Is the relationship between estimates of diagnostic accuracy and characteristics of the patients and test examined? Are analytic methods used that differentiate whether characteristics affect diagnostic accuracy or test threshold?

Irwig et al. (37) applied these standards to 11 published meta-analyses and found deficits both in the meta-analyses and in the primary studies on which they were based.

META-ANALYSES OF OBSERVATIONAL STUDIES

This is an area of much debate, where again a critical problem is the methodologic quality of the primary studies. Meta-analyses of observational studies have been done in many areas and have produced intriguing but controversial findings, including that chlorination of water supplies may slightly but significantly increase the incidence of malignancies of several organs (38) and that increased fat in the diet does not increase the incidence of breast cancer (39). It is clear that the danger of incorrect conclusions is higher than when results of randomized controlled trials are combined, but a distinguished group of researchers concluded that meta-analysis must not be limited to such trials (40). Nonrandomized controlled trial data make up the bulk of what is known regarding medical care. It is imperative that technology assessors, policy makers, clinicians, and patients have clear and explicit means for summarizing all sorts of published data. Policy makers cannot afford to ignore the vast majority
of information available today merely because it does not fit the simple model of classical meta-analysis. Researchers must develop new methods and new models of analysis to assist those who must use today’s imperfect data to make difficult decisions today.

META-ANALYSIS SOFTWARE

At least three computer software packages for meta-analysis are available. A detailed review and comparison was recently published (41).

CONCLUSIONS

In conclusion, if meta-analysis is to be accepted as a scientific tool, each meta-analysis should be conducted as a scientific experiment, beginning with a clear plan of the question to be answered and the methodology to be employed. Attention needs to be paid to intraobserver and interobserver variability, and attempts should be made to identify and minimize bias. Concerns have been expressed about the validity of pooling (42,43), but the process is increasingly used and frequently defended (44–47). We feel that a quantitative synthesis of the data in similar randomized controlled trials is more useful to the practicing physician than a traditional narrative review article, but such syntheses must be properly performed to warrant serious attention. We hope that the points raised here will stimulate discussion that will ultimately lead to better meta-analyses.

REFERENCES

DISCUSSION

Dr. Perman: You pointed out that the principal reason for doing meta-analysis is to be able to pool data from many smaller studies. Is the advantage of meta-analysis over multicenter clinical trials simply that somebody else has already done the work and sponsors have borne the cost, or is there something else that is inherently advantageous about meta-analysis over multicenter trials?

Dr. Sacks: I think the converse, that the best answers come from large-scale multicenter trials. However, it is not always practical to do that, so meta-analysis can be helpful either in the planning of definitive studies or in the interim, while awaiting the results of definitive studies when clinicians still need to make decisions about what they should do now. I think that any studies that are being done now should be designed and conducted in such a way that the results can be pooled with those of other studies.

Dr. Guesry: I feel very uncomfortable with the exponential trend to do meta-analysis rather than to do original work, for many reasons. The first and probably not the most important is the ethical reason: you take other people’s work, usually without their permission, and make a paper of your own from it. But the most important reasons are really methodologic. There is bound to be inhomogeneity in the data: these studies have been done in different countries, with different investigators, different laboratories, and at different times, and many things could have changed in between. And of course, there are all the data that were not published because they were negative, which are not taken into account, so you introduce a very important bias when you do meta-analysis. Another factor that seems very important to me is that we pay more attention to studies that are made with large numbers of subjects, so by pooling together a lot of small studies, each of which would not be credible in itself, we create something that becomes credible simply by virtue of the numbers involved. Then, with the multiplication of meta-analysis, the same incredible study could be used many times. I think that we have to be very careful in the interpretation of meta-analysis, and it would be nice if the authors of the original study, if still living, could have a say in the interpretation.

Dr. Sacks: I agree with a lot of what you said. The strongest evidence comes from meta-analysis that combines the individual patient data, and that can usually be done only with the active collaboration of the original investigators. So that is probably the best way to synthesize the available information. I am less concerned than you are about the ethical issue. I believe there is also an important ethical obligation to the patients, and I think this is at least as important as the obligation to the authors of the papers. I certainly have no objection to trying to get the cooperation and collaboration of the primary authors, but if I thought that
one could make it easier for some physicians to treat their patients, then I think that that would be at least as important for me as what was owed to the authors. The point about publication bias is a big issue, but it is an issue whether or not you are doing meta-analysis. Whether you are doing meta-analysis or looking at a single report of a clinical trial or reading or writing a traditional review, you still have to be concerned about publication bias. So I think that it is an issue not only for meta-analysis but for the medical literature in general.

Dr. Lucas: Let's suppose that you are interested in a nutritional intervention such as whether iron influences neurodevelopment, and there are five small trials in the literature with 50 subjects in them, all just about meeting the criteria for a meta-analysis but none particularly good in its own right. Investigator Bloggs comes along and decides that he is going to do this properly and does a study of 1000 subjects, a big study supported by a proper funding body, well designed, and so forth. Now, which would you rather rely on in your final conclusion of therapeutic efficacy, the stand-alone study on 1000 subjects or a meta-analysis that includes with those 1000 subjects the five small trials, so you now have 1250 subjects but with the noise of the less well conducted smaller studies? Olkin would suggest that the latter is more powerful than the former, but I intuitively feel that the 1000-patient study would be better.

Dr. Sacks: I am not advocating that a meta-analysis is superior to an adequately sized and well-designed single study. In fact, I emphasize that a lot of the work we have been trying to do is to establish criteria for meta-analysis so that it is not done sloppily. I think the critical issue is what the weight of the total evidence is. Meta-analysis will be most useful when a definitive study does not exist, either in planning a definitive study or in making an estimate where there is no definitive study.

Dr. Lucas: The problem is that in evidence-based medicine, people are now in the process of evaluating the literature to determine how they should best proceed, and it is a relevant question whether you should go for the big definitive study or the meta-analysis that includes the big definitive study with all the little ones as well. Olkin's categorization would suggest that anybody evaluating medical practice should in fact always give top priority to the meta-analysis, and I am just challenging that.

Dr. Sacks: I think it is more often the case that meta-analysis of small studies will agree with the definitive single trial than disagree, so there is not necessarily a conflict.

Dr. van't Hof: Also, if you are thinking in terms of evidence-based medicine, you have to read through all the literature. It is a good thing to have a structured review, as with meta-analyses—it saves you a lot of time.

Dr. Klish: I want to know whether or not meta-analysis can make discoveries that the original studies were not intended to discover. The particular meta-analysis I am referring to is one that you are probably aware of that was done several years ago (1,2), when all the clinical trials of dietary treatment of hypercholesterolemia were analyzed. It was found that the incidence of heart disease decreased, as was shown in the individual studies, but the overall mortality increased, and the increased mortality was from violent death, murder, suicide, and other things. I have always been disturbed by that study, and I would like your opinion.

Dr. Sacks: That is an interesting area. Meta-analysis can certainly be used to generate a hypothesis and identify areas that need further research; as to whether you can conclusively answer a question that was not intended to be answered by the original studies. I am not sure that you can. In terms of cholesterol, I think a recent analysis showed that the mortality benefit, not surprisingly, turned out to be greatest in the patients at highest risk (3), and now with the more active drugs, it is becoming clear that you can save lives overall with treatment
of hypercholesterolemia. What was useful about those early studies was to raise the issue of the overuse of some of those drugs in low-risk patients, and maybe a little caution was appropriate for physicians prescribing those drugs. Another example of the value of meta-analysis to try to answer other questions was a study undertaken by a gastroenterologist who used studies of corticosteroid versus placebo in the treatment of a variety of illnesses to try to answer the question of whether the incidence of peptic ulcer was increased by corticosteroid treatment. There were very small numbers in all the studies, but by putting them together, he was able to show that there was a statistically significant increase in peptic ulcer in people who were given steroids. So I think useful information can come from that sort of exercise.

Dr. Lozoff: Are there any techniques within meta-analysis for giving differential weight to better or worse studies by some standard set of methodologic criteria.

Dr. Sacks: There is not a clear answer to that. There have been some papers that have examined the weighting of studies by quality criteria, but it gets complicated for a variety of reasons, mostly because no one has agreed on what are the most important criteria of quality of the primary studies or on how to weight them. None of these criteria has been validated, so although there has been some work in that area, we don’t really have an answer.

Dr. Lozoff: In one of your very early slides of magnitude of effects, you had a real outlier. What is the effect of having such an outlier in a meta-analysis?

Dr. Sacks: This is also a question that has received a lot of attention from statisticians and others, but without a definitive answer. There is, however, an argument for including such outliers: you could argue that if you include a variety of patient populations and still found a fairly consistent overall trend, that is actually stronger evidence than just including people between the ages of 35 and 37 studied in England, for example.

Dr. Ferry: When you put together a multicenter clinical trial, you design a variety of things into it that make the study reliable; you try to get centers that are comparable in many ways, and there are many techniques to try to produce a first-class study. When you look at meta-analyses, you do not have any direct information about biases and about the way some of those individual studies were conducted, so it seems to me it is a little hard to interpret their quality. On the other hand, accepting what you say, that most of these studies really do match good clinical trials, then in general, do meta-analyses make the data more generalizable? In other words, do they interpret across a broader range?

Dr. Sacks: I think they do. In response to the early part of your question about the standards of multicenter studies versus the standards of individual studies, I think there is a lot of variability in the former too. I don’t think you can really say that multicenter studies are always done better than smaller studies. In general, that may be true, but there have been some famous examples, e.g., among breast cancer studies, where patients were included who should not have been, and various things were done that cast a great deal of doubt on the results on these multicenter studies. So I think it is important, whether in small studies, large studies, or multicenter single studies, to do as careful an assessment as possible of the quality of that study, and there are various criteria for doing this.

Dr. Walter: I was interested in the results in Table I comparing the quality of meta-analyses in three different time periods. Looking at those rather quickly, it seems to me that there was no obvious trend to improvement in the quality of reported meta-analyses, although some of the variables seemed to show a pattern in that the middle time period had a rather poor quality: for instance, the description of treatment assignment was 95% in the first time period, 26% in the second, and back up to 79% in the third. Is there perhaps a U.S. renaissance going on in the most recent time period?
Dr. Sacks: Our overall conclusion was that perhaps there was a slight trend toward improvement, but it was over a relatively short period of time, and it is something that needs to be looked at again. There seems to be a tremendous amount of variability in the quality of meta-analysis, and although there are some that I think have been done very well, there are many that have been done very poorly. So with the proliferation of meta-analysis has come a greater diversity in the quality of the individual meta-analyses.

Dr. Hamburger: Most meta-analyses are retrospective, but the studies that we generally prefer are prospective.

Dr. Sacks: I think the issue there is not so much the point in time at which the study starts but the way we try to address the issue. The closer a meta-analysis is done to a prospective study, the better it is going to be—in other words, the question must be spelled out clearly, and the inclusion and exclusion criteria for the meta-analysis properly defined at the outset.

Dr. Whitehead: Where people ignore these quality features, do they do it deliberately, or do they do it because the data do not allow them to include these features? In other words, does this reveal a primary problem with the raw data? If so, doesn’t it cast a question mark over the whole procedure?

Dr. Sacks: We tried to define criteria that will be applicable to the meta-analysis, not just to the primary studies, but it is certainly true that if there are major deficiencies in the primary studies, that will weaken the conclusions of the meta-analysis. But that is a different issue. All those 23 standards that we thought were important were things that can be done by the meta-analysis, not by the primary studies. Most of those meta-analyses were done before we proposed the standards, and, as I tried to show, some of the standards have met with some acceptance, but many of them have not. So those are our opinions as to what the standards should be; we are not saying that those are gold standards.

Dr. Berlin: I have heard an argument that ethical committees ought to look at summaries of existing data before they approve new studies. In this context, my feeling is that we are really obligated to do some kind of quantitative summary if possible before we start new studies. The answer may be that the existing data are not conclusive for a variety of reasons, either because existing studies are of poor quality or because there is a lot of heterogeneity in previous results, but all this is important to know before we undertake a new study, in addition to getting estimates of the likely effect size that we are looking for. Another point relates to multicenter trials versus meta-analysis. There are several examples, but I will mention one. In general, we think that bigger is better, but a study being large does not necessarily make it more reliable. Magnesium after myocardial infarction is an example (ISIS-4): because of issues regarding the timing of administration of magnesium relative to thrombolytic therapy, it seems that the study was not done under optimal conditions. The fact that we can look at a meta-analysis of previous studies before ISIS and then at ISIS itself gives us a contrast that we could not necessarily have picked up from ISIS alone or from the other studies alone. So here is an example in which we may have learned something from the meta-analysis and also from a large clinical trial about how to design the definitive study.

Dr. Rey: I was interested in two of your examples. One was the comparison of rice-based antidiarrhea treatment and WHO oral solution, and the conclusion of the meta-analysis was that there is no advantage to rice-based treatment. My opinion is that it is cleverer to say that rice-based solutions are as effective and less costly than the WHO solution. So we can have a different approach to the same data. The second example is vitamin A. There are two meta-analyses, as you say, and practically all the data come from Southeast Asia; their conclusion is that vitamin A protects against infectious diseases. In other studies, it has also been said that vitamin A is protective against AIDS. But two recent studies in well-nourished children
(5,6) have shown that vitamin A supplementation has a deleterious effect, not only by causing intoxication (25,000 units of vitamin A causes a 10% incidence of side effects) but because the incidence of infectious disease or respiratory disease was found to be increased in well-nourished children treated with vitamin A. So it appears that we are using mathematical approaches to try to solve unsolved problems that may be unsolvable in this way. We don’t need meta-analyses when we look at the effect of penicillin or to prove that ascorbic acid prevents scurvy or that vitamin D prevents rickets. We use meta-analysis only when we have no idea of the result. I think in such cases, it would be more useful to conclude that we don’t know and that we should organize prospective multicenter studies to try to prove that there is an effect or no effect.

Dr. Sacks: I agree with much of what you say. Clearly, one does not need to be a meta-analyst to show that well-nourished children are more likely to be harmed than to be benefited by megadoses of vitamins. However, as I said, I still believe that there is a role for well-done meta-analyses in the planning of definitive trials, in providing interim answers, and in suggesting new areas for further study.

REFERENCES