Meta-Analysis

Background Paper 4

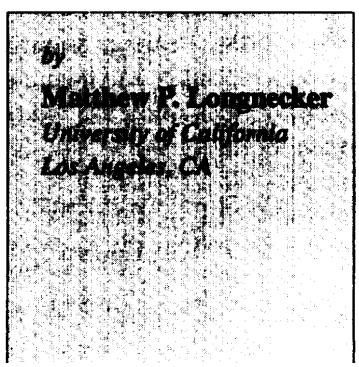
SUMMARY

A meta-analysis is a systematic, quantitative review of a subject. Using very explicit procedures, the analyst reviews the existing studies of a subject and re-analyzes their results to arrive at a more robust and comprehensive result. Three major features distinguish this method from a traditional narrative literature review:

- *the formal and comprehensive search for relevant data;*
- *the explicit, objective criteria for selecting studies to be included; and*
- *the quantitative statistical analysis of the studies' results.*

The justification for analyzing studies' results together in a meta-analysis is that all the component studies provide results that address the same research question.[1] Where all the results come from similar randomized controlled experiments, meta-analysis is widely recognized as a powerful technique for evaluating the effectiveness of health technologies. Where the existing studies are less ideal for a meta-analysis, using the principles of this technique (e.g., making one's criteria for selecting and reviewing studies formal and explicit) can still improve the analyst's ability to undertake an objective, comprehensive review.

Several issues regarding the appropriateness and methodological rigor of meta-analyses are still matters ofdiscussion and



[1] The definition of the term ***same*** in this context depends on the goal of the specific investigation. It may mean that the studies were virtually identical but carried out indifferent places or that the studies were quite different but addressed a similar problem.

debate. These include:

1. *Issues relating to the combinability of studies. Whether to use meta-analysis for nonexperimental or dissimilar studies is controversial and best evaluated on a case-by-case basis. The approach can, however, sometimes provide important insights that might not be evident with traditional narrative review methods.*
2. *Issues relating to publication bias. Results from unpublished studies can be different from published study results. Thus, not including all available studies can lead to bias. Meta-ana - lysts differ in how they attempt to overcome this problem.*
3. *Issues relating to the procedure for conducting a meta-analysis. Meta-analysts also differ in the specific procedures they follow to try to ensure that the review is unbiased and to recognize differences in the quality of the studies being reviewed.*

Despite the continuing discussion of these issues, and their importance for readers to consider when evaluating the quality and validity of any particular meta-analysis, the general technique is now well-established, and its applications continue to grow.

The practice of combining the results of different studies to obtain a more powerful and conclusive result has along history. In 1904, Pearson summarized the relation between mortality and inoculation against enteric fever by calculating the average correlation between mortality and inoculation across five communities (75). Statistical methods for combining the results of agricultural experiments were developed in the 1930s (42). Several applications of statistical methods for combining results across studies appeared in the medical literature in the 1950s (2,64), but it was the application of meta-analysis in the social sciences in the 1970s (37,59) that led to its frequent use in medicine today.

Applied to medical care, meta-analysis can be used to evaluate a treatment effect on any sort of outcome (e.g., to assess a treatment that is supposed to reduce the level of serum cholesterol) or to describe other characteristics when no treatment is involved (e.g., to calculate across studies the average sensitivity of a screening test, the average level of cholesterol indifferent populations, or the average correlation between sex and height). This paper focuses on the use of meta-analysis for assessing the effect of a treatment[2] on a health outcome, such as the risk of death, and assumes that the effect of a treatment is compared with an alternative-no treatment, a placebo, or an accepted treatment.

Rationale

The traditional method of combining the results of previous studies is the narrative review of a subject. Narrative reviews have generally been considered an acceptable evaluation and synthesis of data, but they have several well-known drawbacks. The method of identifying and selecting information is rarely defined, the information may be reviewed haphazardly, and the quality of the data is rarely assessed systematically (71). Because the narrative review approach is not quantitative, nor formally and explicitly systematic in its procedures, a traditional literature review may fail to include important studies and (because of its nonquantitative approach) may fail to make full use of the available data (91). The reviewer's biases may influence the assessment of the data, and directly comparing results across studies can

[2] *Treatment* is used here in a broad sense to include not only medical therapies aimed at improving one's condition (e.g., a drug or a surgical procedure) but also other health interventions and health behaviors (e.g., alcoholic-beverage consumption or cigarette smoking).

be difficult when the treatment effects are expressed differently.[3] In addition, in traditional reviews, authors often assess evidence by "vote counting" (tallying the number of studies that provide evidence for and against the presence of a given treatment effect (40)) without considering that some studies are larger or better than others.

In contrast, in a meta-analysis, the existing studies of the subject of interest are reviewed systematically and quantitatively, using formal and explicit procedures (box 4-1). The advantages of meta-analysis stem from two factors:

1. The use of explicit procedures for identifying and processing the study results. The comprehensive search for relevant studies minimizes the possibility y that available data are omitted. Explicit procedures for evaluating and handling the study results assure, to the extent possible, an unbiased assessment of the data in each study. These explicit procedures also help the reader to assess the competency and appropriateness of the meta-analysis.
2. The expression of the results of individual studies in comparable quantitative terms. A meta-analysis expresses the results of each study in a uniform way, facilitating comparisons of the results and their relation to the size of the individual studies. The uniform expression of results in a meta-analysis allows the analyst to calculate a summary number representing the average effect of a treatment across studies (if such a summary is of interest). A treatment effect is more easily detected when the results of several studies are considered together than when the results are examined individually; a related benefit is that a treatment effect within a subgroup of participants may become clear in the huge sample that is formed when study results are combined. The meta-analytic method facilitates objectivity and reliability, and the use of statistical methods can help researchers identify reasons for any variation in the studies' results (39,54,62,63,84). Identification of patterns in the variation of the treatment effect may contribute to the understanding of the generalizability of the result (31) and may suggest new hypotheses (34).

The astute traditional narrative reviewer may take the size and quality of studies into account, but such steps are key features of the meta-analytic approach. In a recent comparison of conclusions from traditional reviews and meta-analyses, Antman and his associates found that traditional reviews had often failed to recognize important treatment effects that were clearly evident from meta-analyses (l).

Although some authors have asserted that traditional narrative reviews are no longer useful (90), that view seems extreme. If the resources to support a meta-analysis are unavailable or the data are too different for statistical combination, a well-conducted review may be the only alternative. The review, however, will be most useful if the principles of meta-analysis are incorporated to the extent feasible.

The theoretical justification of meta-analysis rests on the assumption that the component studies all address the same research question. If the populations, the treatment, the study design, and the outcomes measured in each study are virtually identical, the studies essentially replicate the same protocol. Consequently, any differences in the treatment effect across studies can be presumed to occur by chance. Under these circumstances, a meta-analysis and an analysis of data from a multicenter clinical trial differ only slightly, and even a skeptic is likely to view a meta-analysis as appropriate.

In practice, however, the studies being combined in a meta-analysis are seldom virtually identical. As the component studies of a meta-analysis become less similar, the appropriateness of ana-

[3] If, for example, one author expressed the treatment effect as the difference in the observed and expected number of cases in the treatment group, and another author expressed the treatment effect as the ratio of the mortality rate in the treated and untreated groups, the results would not be directly comparable.

BOX 4-1: Definitions of Meta-Analysis

The term *meta-anatysis* **was** coined in 1976 by Glass, a social scientist (37), who defined it as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings."1 Other broad definitions in frequent use in the medical and health care literature include:

- "the practice of using statistical methods to combine the outcome of a series of different experiments or investigations" (54);
- "a quantitative summary of research in a particular area" (31); and
- "[the use of the results of collections of research papers to answer specific questions, usually in a quantitative manner" (63).

The National Library of Medicine has developed a detailed definition that actually specifies the procedures to be followed in a meta-analysis (92). It defines this analytic tool as:

> "aquantitative metho dof combining the results of independent studies (usually drawn from the published literature) and synthesizing summaries and conclusions which may be used to evaluate therapeutic effectiveness, plan new studies, etc., with application chiefly in the areas of research and medicine. The method consists of four steps: a thorough literature review, calculation of an effect size for each study, determination of a composite effect size from the weighted combination of individual effect sizes, and calculation of a fail-safe number (number ofunpublished studies with opposing conclusions needed to negate the published literature) to assess the certainty of the composite size."

The elements common to most definitions of meta-analysis are that the analysis is quantitative, that it is based on observations in independent studies, and that the results of the independent observations are summarized across studies.[2] The most prominent difference among the definitions of meta-analysis relates to the kinds of studies that may be included in the analysis. Some definitions stipulate that only results from randomized trials should be analyzed (1 1,98).

The exact systematic procedures used (such as literature searches and quantitative analyses) vary somewhat among published meta-analyses. Thus, the meta-analytic approach is more a set of general principles than a set of standard rules invariably followed. Nonetheless, it is note worthy that at least for the meta-analysis of randomized clinical trials, "meta-analysis has matured as a scientific discipline, with well-documented standards and methods" (57).

Some authors (78) use the term *meta-analysis* to refer to a combined analysis (47) or a pooled analysis (17,60), Unlike other meta-analyses, however, in a combined or pooled analysis the data for the individual participants in different studies are combined into one data set and analyzed as if they were from a multicenter study with a common protocol.[3] In contrast, in most meta-analyses, the studies' results—rather than the original data—are combined. In practice, the results of a combined or pooled analysis usually are virtually identical to those from other meta-analyses. Pooling can be more difficult to perform than other meta-analyses, because It often requires the cooperation of many scientists (to obtain their raw data), but it has the advantage of facilitating the analysis of treatment effects in subgroups of participants.

---

[1] For a scholarty discussion of the etymology of the term, see Dickersi**(24).**

2 other terms sometimes used to describe meta-anafyses include systematic overviews, pooling, data syntheses, and quantitative syntheses **(24) and, less frequently, integrative research reviews, research integrations, research consolidations, research syntheses, quantitative assessments, surveys, re-analyses, and quantitative** reviews (50,69).

3 A study **protocol** defines the **characteristics of** people who are eligible to be in the study, describes the nature and duration of the treatment, discusses how the effect of treatment is assessed, and provides other details of how the study is conducted.

SOURCE: Matthew Longnecker, **1995**

lyzing them jointly becomes a matter of judgment and, therefore, subject to debate. Yet even when the results that are combined come from somewhat dissimilar studies, meta-analysis may be useful—not so much for calculating a summary treatment effect as for allowing the analyst to examine how the treatment effect varies according to study characteristics or across subgroups of participants (72).

Current Applications: An Example

A good illustration of the manner in which meta-analyses are being used can be found in the work of Yusuf and his associates (101), who assessed whether a drug that dissolves blood clots (fibrinolytic therapy) decreased mortality in patients who had heart attacks (myocardial infarctions). The motivation for their assessment was that the results of individual clinical trials addressing this topic appeared contradictory and unreliable. The analysts examined data from studies in which patients were assigned at random to receive either treatment or no treatment (randomized clinical trials).

Using a computerized literature search, reviewing abstracts from scientific meetings, and contacting investigators who had completed trials but not published the results, the analysts located relevant studies and identified 24 eligible trials. For each trial, the number of patients treated with the fibrinolytic therapy, the number not treated with the therapy (the control group), and the number of deaths occurring in each of these two groups were noted. The analysts then used a relatively simple statistical method to calculate across all 24 studies the average effect of the treatment on mortality.

When the data from all the studies were considered together, 51 fewer deaths were found in the group treated with the fibrinolytic therapy than would have been expected if the treatment had no effect on mortality rates. This reduction was found to be statistically significant (see box 4-2). In contrast, just five of the 24 studies had individually shown a statistically significant beneficial effect of treatment. Using the quantitative methods of meta-analysis to consider the results of all the trials simultaneously demonstrated that the treatment was effective in reducing mortality, in a way that a simple narrative review of the results of individual studies would not. (For a detailed discussion of the quantitative methods used in this and other meta-analyses, see appendix 4-A.)

CONDUCTING A META-ANALYSIS

Before conducting a meta-analysis, the analyst should evaluate its utility and desirability and the combinability of the studies. Questions to be asked include:

> Are there any good studies that address the research question?
> If so, are the study designs similar enough that combining them makes sense?
> Given the available data, are the results of the meta-analysis likely to make an important contribution to knowledge?

If the answers to these questions are positive, the analyst then proceeds.

Conducting a meta-analysis is a systematic process (30,50,53,85) that entails the following steps:

1. defining the research question,
2. defining the admissibility criteria for studies,
3. searching for relevant data,
4. reviewing the retrieved data to determine admissibility,
5. assessing the quality of the eligible studies,
6. correcting for bias,
7. performing the data analysis (including sensitivity analysis and influence analysis),
8. assessing the publication bias, and
9. interpreting the results.

Defining the Research Question

In defining the research question, the analyst specifies the treatment under investigation, the treatment's alternative, the outcome, the study populations, and the quantitative measure of the effect in which the analyst is interested.

BOX 4-2: Statistical Significance

"Statistical significance" is a phrase that traditionally has been used to indicate the researcher's belief that the effect observed in an experiment represents a real phenomenon and is unlikely to be due entirely to chance. It is sometimes contrasted with "clinical significance, " which indicates that the effect is not only real but is large enough or important enough to have a meaningful impact.

In a typical medical experiment to determine whether a new treatment has a beneficial effect (compared with plausible alternatives), the researchers begin by assuming that it does not (the "null hypothesis") and then attempting to disprove that assumption. If the study is a well-designed, randomized trial, it is unlikely that an observed apparent treatment effect will be due to chance alone. In statistics, tradition holds that if an observed effect has less than a 5-percent probability of being observed where no treatment effect exists (i.e., $p<.05$), the treatment effect is most likely not zero. In that case, the treatment effect is said to be statistically different from zero (statistically significant).

The use of the 5-percent cutoff level is a popular scientific convention that is somewhat arbitrary; one could also justify choosing 1 percent, or 0.1 percent, or some other low value, as the cutoff for significance. Thus, whether a result is considered statistically significant depends, in part, on the chosen significance level.

An alternative approach, which is growing in favor with researchers and analysts alike, is to place confidence limits around an observed treatment effect, concerning oneself more with the size of the treatment effect and one's certainty about how big it is than with an absolute answer to whether it exists. A 95-percent confidence interval, for example, indicates that in 95 of 100 hypothetical repetitions of the experiment, the true treatment effect would fall within the range of estimated treatment effects included in the confidence interval. (For a complete discussion of confidence intervals, see Rothman (83)).

SOURCE: Matthew Longnecker, **1995.**

Defining the Admissibility Criteria

The next step is to define formal admissibility criteria for the component studies. The research question is expressed in specific terms that facilitate the decisions about whether potentially eligible studies should be included. The criteria might require, for example, that to be admissible a study must:

- be double-blinded,[4]
- have a placebo as the alternative treatment,
- have the dose of the treatment be in a certain range,
- have study participants whose ages are within a certain range,
- present its results in a manner that permits the relevant effect to be calculated,
- evaluate the effect of treatment on the outcome within a specific length of time, and
- be written in English.[5]

Expertise on the research topic is indispensable at this stage (33).

Searching for Relevant Data

The computerized literature search (45), the most important part of the formal search for study re-

[4] In a double-blinded study, neither the patient nor the clinician administering the treatment know which treatment the patient is actually receiving.

[5] Note that this criterion tight eliminate many otherwise eligible studies.

suits, requires special training and is often done in consultation with a qualified librarian. The librarian performs an over-inclusive search using the admissibility criteria for the meta-analysis. Searching at least two different computer databases for studies to include in the meta-analysis increases the number of eligible studies found (14,86).

Because computerized searches can miss important references (23), meta-analysts usually supplement the searches by perusing the reference lists of the identified articles and by consulting experts in the field, abstracts of conferences where relevant papers are likely to have been presented, and any other informally identified sources.

Reviewing the Data for Admissibility

The articles and papers resulting from the literature search are then reviewed to determine whether they meet the criteria for admissibility. Careful documentation of the rejected studies has been advocated (85). The relevant information is abstracted from the admissible articles, and the study results are re-expressed in a standard fashion, if necessary, for subsequent statistical analysis. The characteristics of the individual studies are recorded for use in the data analysis. Ensuring that the analysis does not include multiple studies based on the same participants prevents the inclusion of redundant data (18).

Some authors recommend that the information in the admissible studies be re-abstracted by a second researcher as well, and the extracted data double-checked (35,100). This process is time-consuming but is believed to improve the quality and objectivity of the analysis.

Assessing the Quality of Studies

Many meta-analysts assess the quality of the eligible studies with the aid of standard, published criteria (10,1 4,21) or with criteria specially tailored to the research question under investigation (6,60). Subjective methods of assessing quality have also been employed (6). Examples of criteria used to assess the quality of a randomized clinical trial are:

- whether the participants knew what they received (the treatment or the placebo),
- whether the investigators knew which participants received the treatment and which received the placebo during the trial,
  whether the presentation of the data was appropriate, and
  whether the statistical analyses were appropriate.

Meta-analysts often try to quantify the quality of the studies by awarding points that reflect how well each study approached the ideal for each criterion; the sum of these points for a given study is then used as its summary quality score (16).

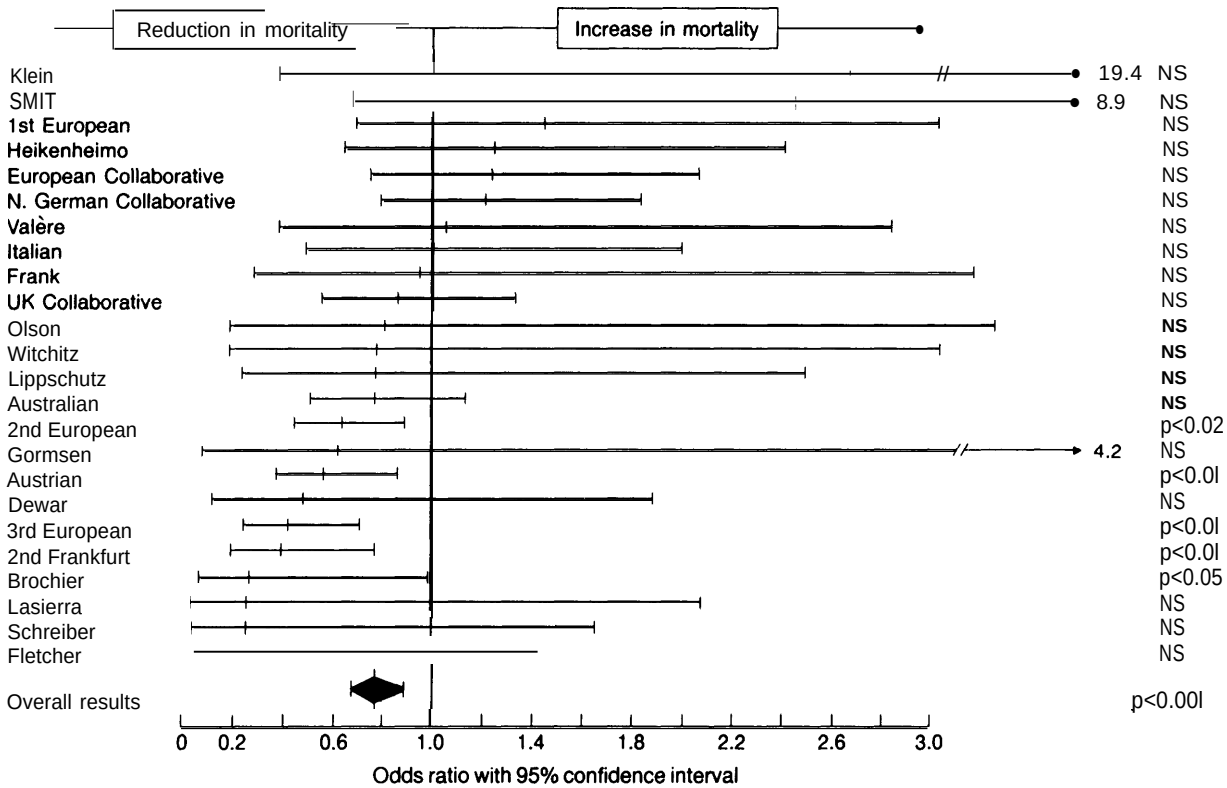
Correcting for Bias

The manner in which a study was designed, conducted, or analyzed can cause the observed effect of the treatment to differ from what would have been observed if the study had been done better. For example, investigators who are aware of what the participants received during a trial tend to find larger treatment effects than do investigators who are blinded to the participants' treatment or lack thereof (19). This may occur because of the investigators' desire to find the new therapy efficacious, which interferes with their ability to make equally accurate assessments of the outcomes in the treatment and control groups. High-quality studies are presumed to provide better estimates of the true effect of the treatment.

If the treatment effect observed in a given study is not an accurate measure of the true effect of the treatment, the result of the study is biased. The amount of bias is reflected by the difference between the observed effect and the true effect.

In some studies, the size of the bias is known with enough certainty that the observed treatment effect can be adjusted for the bias (28,39,93). The adjustment entails taking the treatment effect observed in a study and making it larger or smaller by an amount proportional to the bias (before the study's result is included in the meta-analysis). Correcting the results of studies for bias has not been a frequent practice, however, because it often is nearly impossible to determine whether a given

FIGURE 4-1: Summary of Study Results on Fibrinolytic Agents and Survival After Myocardial Infarction



NOTE: These study results are shown in a different form in tables 4-A-1 and 4-A-2. Here, the studies are listed according to the size of the odds ratio, which has a 95-percent confidence level.

SOURCE: S. Yusuf, R. Collins, R. Peto, et al., "Intravenous and Intracoronary Fibrinolytic Therapy in Acute Myocardial Infarction: Overview of Results on Mortality, Reinfarction, and Side-effects from 33 Randomized Controlled Trials," *European Heart Journal* 6:556-558, 19[illegible]

type of bias occurred in a study or not. Even if a bias is known to have occurred, the degree to which the bias reduced or increased the observed treatment effect is difficult to estimate with certainty.

Analyzing the Data

The quantitative data analysis sets meta-analyses apart from other systematically conducted reviews. In the data analysis, the meta-analyst first examines the results of the component studies. Graphical representation of the studies' results are well-suited to this purpose. For example, figure 4-1 represents the data from the meta-analysis conducted by Yusuf and his associates described above. This figure demonstrates that most of the studies found a beneficial effect of treatment (i.e., an odds ratio less than 1), that the variation in study-specific treatment effects appeared rather large (suggesting heterogeneity), and that many of the individual studies were imprecise.

The reasons for variation among studies' results may be identified by analyzing subgroups of studies separately or by using regression analysis. The degree of variation in the studies' results is assessed with a formal statistical calculation. If a

summary estimate of the treatment effect is appropriate, the effects from the different studies are combined (see appendix 4-A).[6]

Sensitivity analyses are then conducted to determine the extent to which the findings of the meta-analysis depend on assumptions made by the analysts. If, for example, the authors of a meta-analysis excluded several studies on grounds that others might challenge (e.g., the authors excluded studies not published in English), the meta-analysis could be repeated after including those studies to determine whether the overall results were sensitive to those exclusion criteria. If the second meta-analysis yields essentially the same results as the first one did, the authors' findings can more readily withstand criticism.

Influence analyses are another way of testing the robustness of the results. They examine whether the findings of the meta-analysis depend on the inclusion of the results of any particular study, such as a single large study or a study in which the treatment effect is extreme. In an influence analysis, the analyst recalculates the results of the meta-analysis after excluding the particular study of interest (e.g., the study with the unusual treatment effect) to determine whether the new results support the conclusions that were reached when all the data were included. If the results of the meta-analysis do not depend on the inclusion of such studies, the analyst can be more confident of the results.

Information regarding the quality of the studies may be considered in the data analysis. For example, summary quality scores may be included in the calculation of the overall result, in a regression analysis, or in the sensitivity or influence analyses.

Assessing Publication Bias

In the data analysis, the effect of publication bias on the result of the meta-analysis is assessed. Publication bias occurs when the published studies are not representative of the results of all the studies that have been conducted on the research question. Publication bias reflects the preference for publishing studies that have statistically significant findings or that support popular ideas (23). The analyst evaluates the potential effect of publication bias when graphically representing the individual studies' results (3,58) or after estimating the summary treatment effect (48,81).

Interpreting the Results

Like other analysts, meta-analysts conclude the process by interpreting the results so that their generalizability and their implications for practitioners and researchers are clear.

RELIABILITY AND VALIDITY

Reliability

A reliable meta-analysis is one that gives the same result when used again to assess the same research question using the same set of studies. If the identical studies were available to two meta-analysts and both were addressing precisely the same research question, the comparison of the results of the two meta-analyses would be a direct reflection of the reliability of the method. In practice, however, the research questions in replicate meta-analyses usually differ slightly, or the meta-analyses are done at different times, when different studies are available. Thus, some differences between replicate meta-analyses are expected.

[6] Meta-analyses often require specialized statistical methods, because the units of observation in meta-analyses differ from those used in traditional statistical analyses. The units of observation in meta-analyses are the results of independent studies, whereas the units of observation in clinical trials are the data for individual participants. Several reviews of statistical methods in meta-analysis are available (39,42,54) for readers interested in the technical details.

Meta-analyses appear to be at least moderately reliable. In an investigation of 20 replicated meta-analyses done by others, Chalmers and his associates (13) found that the differences among meta-analyses of the same research question were "almost always of degree rather than direction." The authors' interpretations of the findings in the replicate meta-analyses differed more than did the estimates of the summary treatment effects.

In a more recent examination of the reliability of meta-analyses (44), 20 more research questions were examined in replicate meta-analyses. These meta-analyses appeared to be more reliable than the ones studied by Chalmers and his colleagues (13). Henry and Wilson attributed the disagreements between the meta-analyses to differences in the research questions addressed by the analysts. In meta-analyses of progestins to prevent early pregnancy failure, for example, one group found no effect, whereas another group-which focused on studies whose subjects were women with histories of recurrent miscarriages-found evidence that the treatment was effective.

Inasmuch as the quality of meta-analyses before 1987 was found to be highly variable (85), the greater reliability of recent meta-analyses may reflect their improved quality.

Validity

Assessing the validity of meta-analysis requires the comparison of the results of applying this technique with the treatment's true effect, which is rarely known. As a substitute, investigators often use the results from a large clinical trial (one that is not part of the meta-analysis) as an estimate of the true effect. Where the true effect is unknown, this practice may be the most reasonable method for assessing validity.

Two teams of investigators have compared the results of meta-analyses with the results of single, large randomized clinical trials (15,44). Comparing three meta-analyses with the results of their respective clinical trials, Chalmers and his associates found that only one pair clearly agreed on the treatment effect (15). The researchers offered no explanation for the disagreement between another meta-analysis and its large trial but suggested that the third meta-analysis was based on such a small number of subjects that its result might have been greatly influenced by publication bias. Henry and Wilson (44) compared a large trial of oral anticoagulants with a meta-analysis addressing the same question and" found their results comparable. Although similar examples are easily identified (57), their importance is unclear. There has been no comprehensive survey to test the validity of meta-analysis. (Box 4-3 discusses what such a survey might look like if it were conducted.)

The reliability and validity of meta-analyses are likely to improve as the quality of meta-analyses improve. Several investigators have proposed guidelines for assessing the quality of meta-analyses (69,74,79,85). Whether these guidelines succeed in identifying meta-analyses of greater reliability and validity has not been established.

ISSUES AND CONTROVERSIES

Meta-analytic results can be controversial (66) because of concerns regarding the combinability of results, publication bias, or the meta-analytic protocol.

Combinability

Important questions regarding the combinability of studies (72) include the following:

- What types of studies should be included in a meta-analysis?
- Are the study protocols similar enough to warrant combining the results?
- What should be done if the treatment effects vary widely across studies?

Including Studies with Different Designs

Although evidence from good randomized clinical trials is widely accepted as valid, the validity of results from nonrandomized trials is less clear. Because the quality of randomized studies is related to the size of the treatment effect that is observed (19), some researchers believe that nonexperimental data—which presumably are particularly susceptible to bias—have no place in a

BOX 4-3: Designing a Survey To Test the Validity of Meta-analysis

Existing studies that have examined the validity of meta-analyses on particular subjects are few and have somewhat conflicting results. One possible way to examine the question more conclusively would be to conduct a comprehensive survey to address this topic.

Such a survey would begin with the definition of the broad research area about which an investigation of validity is desired. If the area were defined as, for example, the effects of all drugs on total mortality, the investigator would enumerate all the specific drugs for which meta-analyses of the effects on mortality have been done, select at random several specific drug-mortality meta-analyses to evaluate. For each drug-mortality meta-analysis, the investigation would perform a new meta-analysis in the following way: the investigator would order the original studies according to their dates of publication, then take the first five studies and compare the inverse variance of their combined meta-analytic estimate of treatment effect to the inverse variance of the treatment effect of the next published report. If the inverse variance of the next report is at least 50 percent of the inverse variance of the meta-analytic estimate, a comparison of the estimates of the treatment effect will provide information about validity If the inverse variance is less than 50 percent, the meta-analytic estimate should be recalculated to include the result of the sixth study. This result should be compared with that of the seventh study, and the process should be repeated, if necessary.

Some refinement might be required to make this approach work. If it could be carried out, however, its results would enable users of meta-analyses to be more confident of the validity of their results.

SOURCE Matthew Longnecker, 1995.

meta-analysis, and these results are excluded from many meta-analyses (76). Some researchers even define meta-analysis as including only data from randomized trials (1 1,98).

For many research questions, however, data from such trials either are unavailable (4) or not currently possible to collect (e.g., for logistical or ethical reasons). If data from randomized clinical trials (experimental data) are not available, a meta-analysis of observational (nonexperimental data) may still provide a more useful summary of data than a traditional narrative review would provide. The analyst's interpretation of the results of a meta-analysis of observational data should be appropriately conservative, as should the interpretations of the underlying individual observational studies.

Some types of observational studies are more susceptible to bias than others (box 4-4), a fact which meta-analysts must take into account. In a meta-analysis of alcohol consumption in relation to the risk of breast cancer, for example, the analysts examined the results of followup studies and case-control studies separately (60). The treatment effect in the followup studies was found to be larger than that in the case-control studies, and the results of the two types of studies were not combined, because the analysts felt that the results of the followup studies were more likely to represent an unbiased estimate of the treatment effect. This represents an empirical approach to decisions regarding combinability. A consensus regarding the appropriateness of combining the results of observational studies with different designs (regardless of their results) has not been reached (85). At this time, it is common in meta-analyses of observational studies to present the results separately according to the types of study design.

BOX 4-4: Bias in Study Designs

Study designs can be classified according to the methods of obtaining data: simple observation (nonexperimental studies) or observation after some type of intervention (experimental studies). The designs can be further classified according to the units upon which the observations are made: a population (for example, ecologic studies) or individuals (e.g., followup and case-control studies). Study designs vary with respect to the type of bias most likely to occur and the likelihood that the bias would materially affect the result.

In theory, experimental studies (randomized controlled trials) are the best method by which to assess the effect of a treatment. Data obtained from clinical trials are the most powerful for making causal inferences and are the least likely to be biased. Unfortunately, clinical trials are sometimes infeasible for practical, financial, or ethical reasons, When clinical trials cannot be performed, followup and case-control studies are the two nonexperimental study designs most commonly used. In a followup study, the occurrence of disease among the individuals who have and have not undergone the treatment of interest is compared. In case-control studies, the prior treatment experience of persons who already have the disease is contrasted with that in nondiseased (control) subjects, who represent samples of the population in which the cases occur. In general, the results of followup studies are considered less likely to be biased than are the results of case-control studies.

An example of an ecologic study is an examination of the rate of death from breast cancer in relation to per capita sales of fat in different countries, Ecologic studies such as this provide only weak evidence for causal inference, because it is not known whether the subjects who ate fat are the ones who got breast cancer, and because some factor (other than fat intake) correlated with fat sales may be the true reason for the variation in rates of breast cancer across countries.

SOURCE: Matthew Lortgnecker, 1995.

Different Protocols

Another combinability issue is whether the protocols of the studies (e.g., a group of randomized trials) are similar enough to warrant the combining of the studies' results. The answer depends on the particular research question, and the decision to combine the results depends on the judgment of the meta-analysts and their audience.

Many of the criticisms of meta-analysis have been related to decisions regarding combinability in a specific meta-analysis rather than to the method itself (9,34,36,46,52). In conducting a meta-analysis of nonmedical treatments for chronic pain, for example, Malone and Strube (65) calculated the average effect of one treatment on several different kinds of pain, including headache pain and cancer pain. Holyrod and Penizen (46) criticized the meta-analysis because the treatment effect might have been very different for headache and cancer pain, and the summarization across different types of pain might have obscured a treatment effect.

In another example, Held and associates (43) performed a meta-analysis in which they sought to summarize the effect of a general class of drugs (calcium antagonists) on preventing death among persons who had had heart attacks. The meta-analysis was criticized (9) because specific drugs within this class differed in their treatment efficacy. One subclass of drugs (which lowered the heart rate) reduced morbidity and mortality, whereas another subclass increased these outcomes. By analyzing all these drugs together, Held's team had come to the potentially misleading conclusion that the general class of drugs was not effective in reducing mortality and morbidity.

BOX 4-5: Fixed-Effects vs. Random-Effects Models—A Hypothetical Example

Fleiss and Gross (34) have presented an interesting hypothetical example that illustrates some of the issues regarding the choice between the fixed-effects model and the random-effects model (described in greater detail in appendix 4-A). One meta-analysis includes two published studies with odds ratios' of 1.0 and 6.0, and another includes two published studies with odds ratios of 2.0 and 3.0. All four odds ratios have the same variance (of the logarithm of the odds ratio): 0.01. When a fixed-effects model is used for both meta-analyses, the summary odds ratio is 2.45, and the 95-percent confidence intervals extend from 2.13 to 2.81. In both of these meta-analyses, the fixed-effects model's confidence intervals do not even include the values upon which they are based.

If a random-effects model is used to analyze the same data, however, the 95-percent confidence intervals are 0.5 to 10.0 in the first meta-analysis and 1.65 to 3.64 in the second. In both cases, the random-effects confidence intervals include the values upon which they are based, and the width of the confidence intervals reflects the amount of variation in treatment effect between the studies.

To take the example further, note that the random-effects summary in the first meta-analysis (2.45) gives the impression that, on average, a treatment effect exists, even though one of the two studies showed no treatment effect at all (an odds ratio of 1). Although the confidence interval is wide, an observer might look at the summary estimate and fail to appreciate that some studies showed no effect. Despite the disadvantages of summaries from random-effects models, however, the disadvantages of fixed-effects models are often even greater. As a result, random-effects models are gaining widespread acceptance (57).

[1] The odds ratio is the ratio of the odds of an event occurring under one set of circumstances to the odds of the event occurring under another set of circumstances

SOURCE Matthew Longnecker, 1995

Heterogeneous Results

A third combinability issue arises where the treatment effects in the component studies vary markedly—for example, when several studies show a large beneficial effect but other studies show a harmful effect. Large variation (heterogeneity) in the results of individual studies, when present, is usually evident when the study-specific results are represented graphically (see figure 4-1). The analyst can also assess the degree of variability to applying formal statistical tests of this characteristic. Summarizing a treatment effect across studies even when the study results are heterogeneous has been common in meta-analysis, although the practice is a subject of debate (39,77).

Also debated is the appropriate statistical procedure for summarizing the treatment effect in such cases (68,77). The two methods used most frequently to summarize treatment effects across studies are the fixed-effect model and the random-effects model (described in greater detail in appendix 4-A). In practice, if the treatment effects found in the component studies vary greatly, the results of meta-analyses using the two approaches maybe somewhat different (box 4-5). If the results of the studies are homogeneous, however, the two approaches give the same result.

The assumptions underlying the fixed-effects model are that all studies are estimating the same treatment effect and that the difference in effect observed across studies results by chance. The assumptions underlying the random-effects model are that the treatment effect truly differs across studies and that the goal is to determine the aver-

age of the different effects. Fixed-effects models have been used frequently in the past and still have some strong advocates (39,77), although the use of random-effects models to summarize the treatment effect has been favored recently (20,34,72).

Critics of random-effects models (39) question the assumption underlying the model: that the studies were sampled from a hypothetical universe of studies where the true treatment effect varies. They also note that the meaning of random-effects summaries are often misinterpreted. (The correct interpretation is that the random-effects treatment effect is an estimate of the average treatment effect in the universe of hypothetical studies with differing treatment effects.) Proponents of random-effects models argue that they are appropriately imprecise when heterogeneity is present (34).

Publication Bias

Publication bias refers to the fact that results are more likely to be published if they are statistically significant than if they are not (3,23,27). The likelihood of publication is also greater for results from large studies or results that are perceived as important (27). The exclusion of unpublished study results thus can cause the results of a meta-analysis (or any literature review) to be misleading (3).

Informal graphic methods of detecting publication bias have been proposed (58). These methods are easy to use and are widely employed, although their sensitivity and specificity are unknown.

Formal statistical methods for detecting and assessing publication bias have also been proposed (5,48,81), but experts disagree about which formal statistical approach is best (49). Some of these methods are more easily implemented (81) than others (5,48). An advantage of the more computationally intensive methods (5,48) is that they can be used to estimate the true effects of treatment (what would have been observed if there were no publication bias). Estimating the true effects of treatment or determining the number of negative studies necessary for canceling out a positive finding in a meta-analysis is possible, however, only if assumptions are made, and these assumptions may be untestable or not entirely reasonable. The Iyengar and Greenhouse (48) approach, for example, relies on the assumption that the often inappropriate fixed-effects model is used for summarizing treatment effect and that only results significant in one direction[7] are published, which apparently is not entirely true (82). Berlin's approach relies on the assumption that a study's size is unrelated to the size of the treatment effect, which may be incorrect, as well (5). Thus, the formal approaches to assessing publication bias are useful but imperfect solutions to the problem.

The definitive method of correcting publication bias is to include all unpublished results in a meta-analysis, subjecting them to the same inclusion criteria and quality scoring methods as published studies. Accordingly, some authors suggest that analysts routinely attempt to include all the relevant unpublished data in meta-analyses (35, 100).

However, tracing unpublished studies can be difficult. Registries of studies undertaken in a given field are becoming more common (26), but where registries are unavailable, the inclusion of every unpublished study may be impossible for lack of information (22) or may be infeasible on practical grounds (97). When unpublished results can be easily obtained, their inclusion in a meta-analysis, at least in a sensitivity analysis, is reasonable.

Because of a concern that unpublished results may be less reliable than published ones, including unpublished results to combat publication bias is not universally accepted (4).[8] This concern is

[7] Treatments, if they have an effect, can work in two directions: they can either benefit or harm patients. "Statistically significant in one direction" means, for example, that only these studies are published that show the treatment decreased deaths.

[8] This would cause publication bias if unpublished studies tended not only to show no treatment effect but also gave biased estimates of the treatment effect.

probably unwarranted: there is no strong evidence that a study's quality is related to its publication (22,27), and the quality of the component studies may be assessed and considered in a meta-analysis.

Protocol Controversies

The third major category of issues regarding meta-analysis concerns the details of the meta-analytic protocol—the specific procedures followed when conducting the meta-analysis.

Variations in the Standard Meta-Analytic Protocol

Chalmers (12) has long advocated the use of blinding in evaluating the studies for a meta-analysis: to minimize bias in the evaluation process, identifying information is obscured on each article that is potentially eligible for inclusion in the meta-analysis. Thus, the names of the authors, where they did their study, whether they found an effect of treatment, and other pieces of information that might bias an assessment are not available to the person who determines whether to include the study in the meta-analysis. Blinding also helps ensure an unbiased evaluation of the study's quality. Nonetheless, the added assurance that studies are chosen and evaluated in an unbiased fashion comes at a price. If a large number of studies must be reviewed, the blinding process can substantially increase the cost of the meta-analysis. Although the theoretical justification for such blinding is understandable, its actual benefit has not been studied.

Chalmers also recommends that two persons independently evaluate the quality of the studies in a meta-analysis (12). This practice serves as a form of quality control, but its cost-effectiveness has not been documented.

Some statisticians who have spent considerable time thinking about the analytic issues in meta-analysis have recommended that procedures to correct bias be implemented (39,70,84). In practice, however, few investigators have done so. This fact stems partly from a lack of the information necessary for correcting the bias. More important, as mentioned previously, the validity of bias-correction procedures depends on strong assumptions that may be untestable, wrong, or controversial. Furthermore, bias-correction procedures complicate the analysis and may decrease the understandability of its results. Nonetheless, such procedures may be the only sensible approach when optimal analysis of flawed data is a priority.

How best to incorporate assessments of the quality of the component studies in a meta-analysis has not been resolved. One issue is whether a summary measure of a study's quality should be used. Another issue is how to use the summary measure.

The debate as to whether a summary measure of a study's quality has a use in meta-analysis stems from uncertainty about whether such measures reliably and validly identify biased studies. The specific weaknesses that bias a study's observation of the treatment effect are often unknown. A summary score that reflects what the meta-analyst suspects are specific flaws in a study thus may exclude information that is, in fact, related to the findings. If certain studies in a meta-analysis are given less weight because of the analyst's impression that an irrelevant aspect of the study was not ideally conducted, the results of the meta-analysis may be misleading (92). Also, the summary score of quality may contain too much "noise" to adequately reflect the problem of interest.[9] Furthermore, because of space limitations in publications, authors may not present enough information for the quality of the study to be fully assessed; this problem is particularly acute for those attempting to assess observational studies.

Rubin (84) has suggested that individual features of the quality of the studies be examined in relation to the treatment effect, so that important features can be identified. One possible compo-

[9] Noise refers (to the random variation that may obscure the general trend or characteristics of the Item of interest.

nent of a summary measure of quality, for example, is whether the study participants were blinded to the treatment they received. Analysts could examine whether the treatment effect found in studies with blinding differed from that found in studies where blinding was not used. Although this approach is sensible, it is of limited use in practice: the attributes of different studies are often so highly correlated and the numbers of studies so limited that analysts have dificulty linking specific aspects of the quality of a study to the size of the treatment effect observed.

How best to incorporate a summary measure of the quality of a study into a meta-analysis is also unclear. Detsky (21) has outlined the major options for using the information. The first option is to exclude poor studies from the analysis. The second option is to weight a given study not only according to its statistical precision, but also according to its quality. Finally, the quality scores may be included as terms in statistical models or serve as the basis for sensitivity or subgroup analyses. A consensus about which of these methods is theoretically preferable has not emerged in the literature on meta-analysis.

Bayesian Meta-Analysis

A Bayesian approach to meta-analysis (box 4-6) is strongly supported by some investigators (29). Few meta-analysts have used Bayesian methods and few empirical comparisons between the results from the Bayesian and traditional methods have been presented (79).

Bayesian methods have three potential advantages. First, the statistical results are more easily interpreted than are those from the traditional fre-

BOX 4-6: Frequentist and Bayesian Approaches to Data Analysis

The frequentist and Bayesian approaches to data analysis are two different ways to use data to make inferences about the treatment effect, The frequentist approach is more prevalent throughout the sciences, though the use of the Bayesian approach is growing.1

The frequentist approach assumes that a given study could hypothetically be repeated an infinite number of times, and that the particular treatment effect observed in the study actually done is just one of all possible observations, selected at random.

In the Bayesian approach to statistics, the analyst specifies in quantitative terms his or her belief (and certainty in that belief) about the size of the treatment effect under investigation, and the observations made in a particular study are used to modify the analyst's belief.

A frequentist, at the end of the data analysis, specifies an estimate of the size of the treatment effect (based only on the data in the study performed) and also presents a p-value or an equivalent 95-percent confidence interval (see box 4-2). This confidence interval describes statistically the interval within which the true effect of the treatment would lie in 95 of 100 hypothetical repetitions of the experiment. Because most studies cannot be repeated multiple times, the assumptions upon which the statistics are based cannot be verified directly,

A Bayesian, at the end of the data analysis, specifies an estimate of the size of the treatment effect and an interval in which he or she believes with 95-percent certainty the true treatment effect lies. This approach gives validity to the analyst's subjective beliefs, a controversial issue behind some of the resistance to broader use of Bayesian statistics in the sciences,

[1] For an in-depth discussion of these two approaches, see **Oakes (73).**

SOURCE: Matthew Longnecker, 1995.

quentist approach (box 4-6) (38). Second, when adjusting treatment effects for bias, the analysts may incorporate their degrees of certainty or uncertainty about the adjustment into the analysis. Third, greater flexibility is possible when combining different types of information about the treatment effect.

The Bayesian approach also has three disadvantages. First, even fewer people understand Bayesian methods than understand the frequentist approach. Second, performing Bayesian analyses requires specialized computer software that has only recently become widely available (29). Third, because Bayesian analyses can be based on even more assumptions than can frequentist analyses, the Bayesian results maybe subject to more debate. Once empirical comparisons of the two methods are available and more investigators have experience with Bayesian methods, the relative merits of the approach will be easier to assess.

FUTURE APPLICATIONS

Meta-analysis is gaining in popularity, especially in the medical field. The tool has been used frequently for assessing technology and promises to be useful for improving assessments of risk and, by strengthening the estimates of the effects of treatments, for increasing the accuracy of cost-effectiveness analyses.

The number of meta-analyses conducted each year is growing. Dickersin and her associates (24), in their examination of the literature, found three meta-analyses published between 1966 and 1969, nine published between 1976 and 1978, and 44 published between 1985 and 1987. Seventy percent of these meta-analyses were on medical topics. The computerized database of the National Library of Medicine began formally identifying meta-analyses and related work in 1989. A computerized search for articles relating to meta-analysis in that database resulted in 232 articles for the year 1989,297 articles for 1990, and 368 articles for 1991. Although only a portion of these articles are themselves meta-analyses (many are merely *about* meta-analysis), the increasing prominence of this tool in the medical literature is evident.

The use of meta-analyses is also growing. For example:

- Influential medical professionals use evidence from meta-analyses to evaluate treatment efficacy (57,67,91).
- The Food and Drug Administration allows the results of meta-analyses to support New Drug Applications (34).
- The U.S. General Accounting Office (GAO) has endorsed meta-analysis as a method of assessing treatment efficacy (93).
- The Agency for Health Care Policy and Research is using meta-analyses to guide policy regarding medical procedures that will be reimbursable under Medicare (23).

GAO (93) has proposed that meta-analyses combining results from randomized clinical trials and "database analyses" be conducted. 10 The justification for combining results among studies conducted using different designs is that randomized clinical trials tend to measure the treatment effects in only small subsets of all the types of subjects who might receive the treatments in practice. Database analyses provide an estimate of the treatment effect in a much more diverse group of subjects, and they reflect the effect of treatment as administered by physicians in general, not just those specialists conducting clinical trials. They are also observational studies, however, and an estimated treatment effect based on observational data alone is often not reliable (see J. Whittle, "Analysis of Large Administrative Databases," background paper #2).

GAO has named this type of data synthesis "cross-design synthesis." The technique entails combining the results from studies with different

[10] In "database analyses" (as the term is used by GAO (91)), information about patients—including the treatments they have received and he outcomes of those treatments—from computer records routinely kept for accounting purposes is analyzed to provide estimates of the treatnent effects.

designs and analyzing raw data from the database(s) as part of the analysis, whereas meta-analysis entails simply analyzing the results of studies. The validity of cross-design synthesis will be even more difficult to establish than the validity of traditional meta-analysis has been. Still, like meta-analysis in general, cross-design synthesis might sometimes facilitate more efficient use of existing data than is possible with the traditional narrative approach to evaluating the effects of treatments.

The potential for meta-analysis to improve risk assessments [11] has been recognized by several observers (32,87). Meta-analysis may improve the accuracy of cost-effectiveness analyses and can identify effective therapies, gauge the treatment effect, and estimate other quantities that influence cost-effectiveness (88).

Because the meta-analytic approach can be applied to virtually any problem in the evaluation of medical technology that has been previously studied (28), its use in the health care field can be expected to increase, but the benefit of a given meta-analysis in relation to its cost deserves critical evaluation. The cost of a meta-analysis depends on the number of potentially eligible studies, the number of admissible studies, the use of blinding (or lack there of), the usability of the format in which the data have been presented, the experience of the analysts, the number of decisions that the analysts must make, and other factors.

For meta-analysis to be beneficial, its results must be persuasive. The results of a meta-analysis are most likely to be persuasive where there is little controversy about how it was done or how its results should be interpreted. The credibility of a meta-analysis is likely to be greatest when the approach is applied to clearly combinable, homogeneous results from methodologically strong randomized clinical trials that were identified through a registry of all trials conducted on a given research question. As the circumstances of a meta-analysis depart from this ideal, the validity of its results will be less clear and increasingly difficult to assess. Even when the results of a meta-analysis are controversial, however, they may provide insights into data not attainable with traditional review methods. Several authors have suggested criteria to be used in evaluating the results of a meta-analysis (53,85), and these may assist an evaluation of a meta-analytic result. With or without such guidelines, the evaluator must have subject-matter expertise to fully appreciate the worth of a given meta-analysis.

CONCLUSION

Despite the controversies, meta-analysis appears to be generally accepted as a useful tool for analyzing data from, at least, randomized clinical trials (51,89). Yet unquestioning reliance on the results of meta-analysis (67) has been criticized (55), because despite the advantages of meta-analysis' explicit, formal approach, the results of meta-analyses are still influenced by the sometimes fallible judgments of their authors (93).

The usefulness of meta-analysis may best be considered on a case-by-case basis. Where the setting is ideal for a careful meta-analysis, the method may accelerate and aid the evaluation of health care technologies and practices. Where the setting is less than ideal, the method may help investigators to identify the combination of treatment and participant characteristics where the efficacy is greatest or the circumstances under which more or better data regarding a treatment effect are needed.

Although the results of meta-analyses may reduce the number of randomized controlled trials needed to evaluate a technology (57), meta-analysis should not eclipse the need for randomized trials. In fact, a meta-analysis may clarify the need for a trial when the meta-analytic result suggests that a treatment effect is present but the estimate of the effect is imprecise.

[11] Risk assessment, according to Last (56), is "the qualitative or quantitative estimation of the likelihood of adverse effects that may result from exposure to specified health hazards or from the absence of beneficial influences."

Meta-analysis, properly done, requires significant resources, including access to experts in the specific technique and in the subject being studied. Since identifying relevant studies is one of the most time-consuming steps, the systematic registration of randomized clinical trials (and other studies) could improve the efficiency of this technique.

REFERENCES

1. Antman, E. M., Lau, J., Kupelnick, B., et al., "A Comparison of Results of Meta-Analyses of Randomized Control Trials and Recommendations of Clinical Experts: Treatment for Myocardial Infarction," *Journal of the American Medical Association 268(2):240-248, 1992.*
2. Beecher, H. K., "The Powerful Placebo," *Journal of the American Medical Association 159(17): 1602-1606, 1955.*
3. Begg, C. B., and Berlin, I. A., "Publication Bias: A Problem in Interpreting Medical Data," *Journal of Royal Statistical Society* 151(3):419-463, *1988.*
4. Begg, C. B., and Berlin, J. A., "Publication Bias and Dissemination of Clinical Research," *Journal of the National Cancer Institute 81(2): 107-115, 1989.*
5. Berlin, J.A., Begg, C. B., and Louis, T. A., "An Assessment of Publication Bias Using a Sample of Published Clinical Trials," *Journal of the American Statistical Association* 84(406):381-392, *1989.*
6. Berlin, J. A., and Colditz, G. A., "A Meta-Analysis of Physical Activity in the Prevention of Coronary Heart Disease," *American Journal of Epidemiology* 132(4):612-628, *1990.*
7. Berlin, J. A., Laird, N. M., Sacks, H. S., et al., "A Comparison of Statistical Methods for Combining Event Rates from Clinical Trials," *Statistics in Medicine 8(2): 141-151,* 1989.
8. Berlin, J. A., Longnecker, M. P., and Greenland, S., "Meta-Analysis of Epidemiologic Dose-Response Data," *Epidemiology* 4:218-288, *1993.*
9. Boden, W. E., "Meta-Analysis in Clinical Trials Reporting: Has a Tool Become a Weapon?" (editorial), *American Journal of Cardiology* 69(6):681-686, *1992.*
10. Brown, S. A., "Measurement of Quality of Primary Studies for Meta-Analysis," *Nursing Research* 40(6):352-355, *1991.*
11. Bulpitt, C.J., "Meta-Analysis," *Lancet* 2(8602):93-94, *1988.*
12. Chalmers, T. C., "Problems Induced by Meta-Analyses," *Statistics in Medicine* 10(6):971-979, *1991.*
13. Chalmers, T. C., Berrier, J., Sacks, H. S., et al., "Meta-Analysis of Clinical Trials as a Scientific Discipline. II: Replicate Variability and Comparison of Studies that Agree and Disagree," *Statistics in Medicine* 6:733-744, *1987.*
14. Chalmers T. C., Hewett P., Reitman D., et al., "Selection and Evaluation of Empirical Research in Technology Assessment," *International Journal of Technology Assessment in Health Care* 5(4):521-536, *1989.*
15. Chalmers, T. C., Levin, H., Sacks, H. S., et al., "Meta-Analysis of Clinical Trials as a Scientific Discipline. I: Control of Bias and Comparison with Large Cooperative Trials," *Statistics in Medicine* 6:315-325,1987.
16. Chalmers, T. C., Smith, Jr., H., Blackburn, B., et al., "A Method for Assessing the Quality of a Randomized Control Trial," *Controlled Clinical Trials* 2(l):31-49, *1981.*
17. Checkoway, H., "Data Pooling in Occupational Studies," *Journal of Occupational Medicine* 33(12):1257-1260, *1991.*
18. Choi, B., "Meta-Analysis" (letter), *Annals of Internal Medicine* 109(l):83, *1988.*
19. Colditz, G. A., Miller, J. N., and Mosteller, F., "How Study Design Affects Outcomes in Comparisons of Therapy. I: Medical," *Statistics in Medicine* 8(4):441-454, *1989.*
20. DerSimonian, R., and Laird N., "Meta-Analysis in Clinical Trials," *Controlled Clinical Trials* 7:177-188, *1986.*

21. Detsky, A. S., Naylor, C.D., O'Rourke, K., et al., "Incorporating Variations in the Quality of Individual Randomized Trials into Meta-*Analysis,'' Journal of Clinical Epidemiology* 45(3):255-265, *1992.*
22. Dickersin, K., "The Existence of Publication Bias and Risk Factors for Its Occurrence," *Journal of the American Medical Association 263(10): 1385-1389, 1990.*
23. Dickersin, K., and Berlin, J. A., "Meta-Analysis: State-of-the-Science," *Epidemiologic Reviews* 14:154-176, *1992.*
24. Dickersin, K., Higgins, K., and Meinert, C. L., "Identification of Meta-Analyses: The Need for Standard Terminology," *Controlled Clinical Trials* 11(l):52-66, *1990.*
25. Durlak, J. A., and Lipsey, M. W., "A Practitioner's Guide to Meta-Analysis," *American Journal of Community Psychology* 19(3): *291-332, 1991.*
26. Easterbrook, P.J., "Directory of Registries of Clinical Trials," *Statistics in Medicine* 11(3):345-359, *1992.*

2'7. Easterbrook, P.J., Berlin, J. A., and Gopalan, R., et al., "Publication Bias in Clinical Research," *Lancet* 337(8746):867-872, *1991.*

28. Eddy, D. M., Hasselblad, V., and Shachter, R., "An Introduction to a Bayesian Method for Meta-Analysis: The Confidence Profile Method," *Medical Decision Making* 10(1): 15-23, 1990.
29. Eddy, D. M., Hasselblad, V., and Shachter, R., *Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence* (Boston, MA: Academic Press, 1992).
30. Einarson, T.R., Leeder, J. S., and Koren, G., "A Method for Meta-Analysis of Epidemiological Studies," *Drug Intelligence and Clinical Pharmacy* 22(10):81 *3-824, 1988.*
31. Ellenberg, S. S., "Meta-Analysis: The Quantitative Approach to Research Review," *Seminars in Oncology* 15(5):472-481,1988.
32. Farland, W. H., "The U.S. Environmental Protection Agency's Risk Assessment Guidelines: Current Status and Future Directions," *Toxicology and Industrial Health* 8(3):205-212, *1992.*

33. Felson, D.T., "Bias in Meta-Analytic Research," *Journal of Clinical Epidemiology* 45(8):885-892, *1992.*

34. Fleiss, J.L., and Gross, A.J., "Meta-Analysis in Epidemiology, with Special Reference to Studies of the Association Between Exposure to Environmental Tobacco Smoke and Lung Cancer: A Critique," *Journal of Clinical Epidemiology 44(2): 127-139, 1991.*

35. Furberg, C. D., and Morgan, T. M., "Lessons from Overviews of Cardiovascular Trials," *Statistics in Medicine* 6(3):295-306, *1987.*

36. Gelber, R. D., and Goldhirsch, A., "Meta-Analysis in Clinical Research" (letter), *Annals of Internal Medicine 108(1): 158-159, 1988.*

37. Glass, G. V., "Primary, Secondary, and Meta-Analysis of Research," *Educational Researcher* 5:3-8, *1976.*

38. Goodman, S. N., "Meta-Analysis and Evidence," *Controlled Clinical Trials 10:188-204, 1989.*

39. Greenland, S., "Quantitative Methods in the Review of Epidemiologic Literature," *Epidemiologic Reviews 9: 1-30, 1987.*

40. Greenland, S., and Longnecker, M. P., "Methods for Trend Estimation from Summarized Dose-Response Data, with Applications to Meta-Analysis," *American Journal of Epidemiology* 135: 1301-1309, 1992.

41. Greenland, S., and Salvan, A., "Bias in the One-Step Method for Pooling Study Results," *Statistics in Medicine* 9(3):247-252, *1990.*

42. Hedges, L. V., and Olkin I., *Statistical Methods for Meta-Analysis* (Orlando, FL: Academic Press, 1985).

43. Held, P. H., Yusuf, S., and Furberg, C. D., "Calcium Channel Blockers in Acute Myocardial Infarction and Unstable Angina: An Overview," *British Medical Journal 299* (6709):1187-1192, *1989.*

44. Henry, D.A., and Wilson, A., "Meta-Analysis Part 1: An Assessment of Its Aims, Validity and Reliability," *Medical Journal of Australia* 156(l):31-38, 1992.

45. Hewitt, P., and Chalmers, T.C., "Using MEDLINE To Peruse the Literature," *Controlled Clinical Trials* 6:75-83, *1985.*

46. Holroyd, K. A., and Penzien, D. B., "Meta-Analysis Minus the Analysis: A Prescription for Confusion," *Pain* 39(3):359-363, *1989.*

47. Howe, G., Rohan, T., Decarli, A., et al., "The Association Between Alcohol and Breast Cancer Risk: Evidence from the Combined Analysis of Six Dietary Case-Control Studies," *International Journal of Cancer* 47:707-710, *1991.*

48. Iyengar, S., and Greenhouse, J. B., "Selection Models and the File Drawer Problem," *Statistical Science 3(l): 109-117, 1988.*

49. Iyengar, S., and Greenhouse, J. B., "Rejoinder," *Statistical Science 3(l): 133-135, 1988.*

50. Jenicek, M., "Meta-Analysis in Medicine: Where We Are and Where We Want to Go," *Journal of Clinical Epidemiology* 42(l): *35-44, 1989.*

51. Kassirer, J. P., "Clinical Trials and Meta-Analysis: What Do They Do for Us?" (editorial), New *England Journal of Medicine 327(4): 273-274, 1992.*

52. Kriebel, D., Wegman, D. H., Moure-Erase, R., et al., "Limitations of Meta-Analysis: Cancer in the Petroleum Industry" (letter), *American Journal of Industrial Medicine* 17(2):269-271, *1990.*

53. L'Abbe, K. A., Detsky, A. S., and O'Rourke, K., "Meta-Analysis in Clinical Research," *Annals of Internal Medicine* 107:224-233, *1987.*

54. Laird, N. M., and Mosteller, F., "Some Statistical Methods for Combining Experimental Results," *International Journal of Technology Assessment in Health Care* 6(l):5-30, *1990.*

55. Lancet, "Cross Design Synthesis: A New Strategy for Studying Medical Outcomes" (editorial), *Lancet* 340:944-946, *1992.*

56. Last, J. M., *A Dictionary of Epidemiology (New* York, NY: Oxford University Press, 1988).

57. Lau, J., Antman, E. M., Jimenez-Silva, J., et al., "Cumulative Meta-Analysis of Therapeutic Trials for Myocardial Infarction," New *England Journal of Medicine 327(4): 248-254, 1992.*

58. Light R.J., and Pillemer D. B., *Summing Up* (Cambridge, MA: Harvard University Press, 1984).

59. Light, R. J., and Smith, P. V., "Accumulating Evidence: Procedures for Resolving Contradictions Among Different Research Studies," *Harvard Education Review 41(4):429-471, 1971.*

60. Longnecker, M. P., Berlin, J. A., Orza, M.J., et al., "A Meta-Analysis of Alcohol Consumption in Relation to Risk of Breast Cancer," *Journal of the American Medical Association* 260(5):652-665, *1988.*

61. Longnecker, M.P., Martin-Moreno, J. M., Knekt, P., et al., "Serum a-Tocopherol Concentration in Relation to Subsequent Colorectal Cancer: Pooled Data from Five Cohorts," *Journal of the National Cancer Institute* 84:430-435, *1992.*

62. Louis, T. A., "Assessing, Accommodating, and Interpreting the Influences of Heterogeneity," *Environmental Health Perspectives* 90:215-222, *1991.*

63. Louis, T.A., Fineberg, H. V., Mosteller, F., "Findings for Public Health from Meta-Analysis," *Annual Review of Public Health* 6:1-20, *1985.*

64. MacMahon, B., and Hutchison, G. B., "Prenatal X-ray and Children's Cancer: A Review," *Acta Unio Internationalism Contra Cancrum 20: 1172-1174, 1964.*

65. Malone, M. D., and Strube, M.J., "Meta-Analysis of Non-Medical Treatments for Chronic Pain," *Pain* 34(3):231-244, *1988.*

66. Mann, C., "Meta-Analysis in the Breech," *Science* 249:476-480, *1990.*

67. Manson, J. E., Tosteson, H., Ridker, P. M., et al., "The Primary Prevention of Myocardial

In farction," New *England Journal of* Medicine 326(21): 1406-1416, 1992.

68. Meier, P., "Commentary," *Statistics in Medicine* 6:329-331, *1987.*
69. Meinert, C. L., "Meta-Analysis: Science or Religion?," *Controlled Clinical Trials 10(4 suppl.):257S-263S, 1989.*
70. Mosteller, F., "Summing Up," *The Future of Meta-Analysis,* K.W. Wachter, and M.L. Straf (eds.) (New York, NY: Russell Sage Foundation, 1990).
71. Mulrow, C. D., "The Medical Review Article: State of the Science," *Annals of Internal Medicine* 106(3):485-488, *1987.*
72. National Research Council, *Combining Information: Statistical Issues and Opportunities for Research* (Washington, DC: National Academy Press, 1992).
73. Oakes, M., *Statistical Inference* (Chestnut Hill, MA: Epidemiology Resources, Inc., 1990).
74. Oxman A. D., and Guyatt G. H., "Validation of an Index of the Quality of Review Articles," *Journal of Clinical Epidemiology* 44(1 1):1271-1278, 1991.
75. Pearson, K., "Report on Certain Enteric Fever Inoculation Statistics," *British Medical Journal* 2:1243-1246, *1904.*
76. Pete, R., "Why Do We Need Systematic Overviews of Randomized Trials?" *Statistics in Medicine* 6(3):233-240, *1987.*
77. Pete, R., "Discussion," *Statistics in Medicine* 6(3):241-244, *1987.*
78. Pignon, J. P., Arriagada, R., Ihde, D. C., et al., "A Meta-Analysis of Thoracic Radiotherapy for Small-Cell Lung Cancer," New *England Journal of Medicine 327(23): 1618-1624, 1992.*
79. Pollard, W. E., *Bayesian Statistics for Evaluation Research: An Introduction* (Beverly Hills, CA: Sage Publications, 1986).
80. Robins, J., Breslow, N., and Greenland, S., "A Mantel-Haenszel Variance Consistent Under Both Large Strata and Sparse Data Limiting Models," *Biometrics* 42:311-323, *1986.*
81. Rosenthal, R., "The File Drawer Problem, and Tolerance for Null Results," *Psychological Bulletin* 86(3):638-641, *1979.*
82. Rosenthal, R., and Rubin, D. B., "Comment: Assumptions and Procedures in the File Drawer Problem," *Statistical Science 3(l): 120-125, 1988.*
83. Rothman, K.J., *Modern Epidemiology* (Boston, MA: Little, Brown and Company, 1986).
84. Rubin, D. B., "A New Perspective," *The Future of Meta-Analysis,* K.W. Wachter, and M.L. Straf (eds.) (New York, NY: Russell Sage Foundation, 1990).
85. Sacks, H. S., Berrier, J., Reitman, D., et al., "Meta-Analysis of Randomized Controlled Trials," New *England Journal of Medicine* 316:450-455, *1987.*
86. Schoones, J.W., "Searching Publication Data Bases" (letter), *Lancet* 335(8687):481, 1990.
87. Shore, R. E., Lyer, V., Altshuler, B., et al., "Use of Human Data in Quantitative Risk Assessment of Carcinogens: Impact on Epidemiologic Practice and the Regulatory Process," *Regulatory Toxicology and Pharmacology 15: 180-221, 1992.*
88. Simes, J., "Meta-Analysis: Its Importance in Cost-Effectiveness Studies," *Medical Journal of Australia 153* (suppl.):S13-S16, *1990.*
89. Spitzer, W.O., "Meta-Analysis: Unanswered Questions About Aggregating Data" (editorial), *Journal of Clinical Epidemiology* 44(2):103-107, *1991.*
90. Teagarden, J. R., "Meta-Analysis: Whither Narrative Review?" *Pharmacotherapy 9(5): 274-281, 1989.*
91. Thacker, S. B., "Meta-Analysis. A Quantitative Approach to Research Integration," *Journal of the American Medical Association* 259(11):1685-1689, *1988.*
92. Thompson, S.G., and Pocock, S.J., "Can Meta-Analyses Be Trusted?" *Lancet 338* (8775):1 *127-1130, 1991.*
93. U.S. General Accounting Office, *Cross Design Synthesis: A New Strategy for Medical Effectiveness Research* (Washington, DC: U.S. Government Printing Office, 1992).

94. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Library of Medicine, *Medical Subject Headings, Annotated Alphabetic List* (Bethesda, MD: 1989).
95. U.S. Department of Health and Human Services, Public Health Service, Office of the Assistant Secretary for Health, Office of Disease Prevention and Health Promotion, Preventive Services Task Force, *Guide to Clinical Preventive Services: An Assessment of the Effectiveness of 169 Interventions* (Baltimore, MD: Williams & Wilkins, 1989).
96. Wachter, K. W., "Disturbed by Meta-Analysis?" *Science* 241(4872):1407-1408, 1988.
97. Wachter, K. W., "Concepts Under Scrutiny: Discussion," *The Future of Meta-Analysis*, K.W. Wachter and M.L. Straf (eds.) (New York, NY: Russell Sage Foundation, 1990).
98. Whitehead, A., and Whitehead, J., "A General Parametric Approach to the Meta-Analysis of Randomized Clinical Trial s," *Statistics in Medicine* 10(1 1):1665-1677, 1991.
99. Wilson, A., and Henry, D. A., "Meta-Analysis Part 2: Assessing the Quality of Published Meta-Analyses," *Medical Journal of Australia 156(3): 173-174, 1992.*
100. Yusuf, S., "Obtaining Medically Meaningful Answers from an Overview of Randomized Clinical Trials," *Statistics in Medicine* 6(3):281-294, *1987.*
101. Yusuf, S., Collins, R., and Pete, R., et al., "Intravenous and Intracoronary Fibrinolytic Therapy in Acute Myocardial Infarction: Overview of Results on Mortality, Reinfarction and Side-Effects from 33 Randomized Controlled Trials," *European Heart Journal* 6:556-558, *1985.*

APPENDIX 4-A: QUANTITATIVE METHODS IN META-ANALYSIS

The quantitative methods appropriate for any analysis, including a meta-analysis, depend on the research question to be addressed. Although meta-analysis has many uses in health care-e. g., calculations across studies of the average value of a laboratory result, a disease rate, a population characteristic, or the sensitivity of a diagnostic test—the discussion here focuses on meta-analytic techniques for evaluating the effect of a medical treatment on an outcome.

Determining the Treatment Effect in a Single Study

Calculating the Size of the Treatment Effect

In evaluating a controlled trial of a medical treatment, the outcome measure in the treatment and control groups is usually expressed as a proportion, rate, or mean. One might be interested in a drug's effect, for example, on the proportion of participants who develop side effects, on the rate at which subjects die or develop a disease, * or on their mean level of cholesterol. In the meta-analysis of fibrolytic therapy performed by Yusuf and his associates, the outcome measure for the treatment and control groups in each of the component studies was the proportion of patients with myocardial infarction who died within a specified period of time.

The treatment effect in a study is the outcome measure in the treated group compared with that in the control group. The comparison between outcome measures may be a difference, a ratio, or a related measure. Where the outcome is a proportion, one might be interested, for example, in the difference in the proportion who died in the treated and control groups (the proportion dead in the treatment group minus the proportion dead in the control group); in the ratio of these proportions (the proportion dead in the treatment group divided by the proportion dead in the control group); or in the difference in the rates (rate difference) or the ratio of the rates (rate ratio) between the treatment and control groups. Where the outcome is a mean, the treatment effect usually examined is the difference of the means in the treated and control groups.

In practice, the treatment effect is often expressed as: 1) the difference between the observed and expected number of deaths in the treatment group, and 2) the odds of death in the treatment group divided by the odds of death in the control group (odds ratio). (Odds are related to proportions in that a proportion divided by one minus the proportion is the odds.) Neither expression is a simple example of a difference or a ratio of outcome measures, but they are worth explaining in detail because they are commonly used in the evaluation of medical therapies.

The first treatment effect commonly measured is the difference between the observed and expected numbers of deaths in the treatment group. The outcome in the Yusuf meta-analysis of fibrinolytic therapy was a proportion (the number of deaths in a treatment or control group divided by the number of subjects in that group) (table 4-A-l). The observed number of deaths is the numerator in the proportion. Thus, for the Fletcher study (the first study shown in table 4-A-l), the observed number of deaths in the treatment group was one, and the total number of study participants in the treatment group was 12. Thus, the proportion of observed deaths in the treated group was 1/12, or 8.3 percent. If the treatment had no effect, the proportions of deaths in the treatment and control groups could be expected to be the same.

The authors of the meta-analysis computed the proportion of deaths that would be expected in the treatment group if the treatment had no effect. This was accomplished by combining the number of deaths in the treatment and control groups, then dividing by the number of participants in both

---

1 A death rate is calculated as the number of deaths per unit of person-time. If 100 study participants are observed for 2 years, for example, and one of them dies during that period, the death rate is 1/200 person-years. The proportion of participants who die in the 2-year period is 1/100.

TABLE 4-A-1: Summary of Study Results on Fibrinolytic Agents and Survival After Myocardial Infarction

study No. (1)	Author (2)	Treated Group Deaths (3)	Treated Group Total (4)	Control Group Deaths (5)	Control Group Total (6)	O - E[a] (7)	V(O - E)[b] (8)
1	Fletcher	1	12	4	11	-1.6	1.0
2	Dewar	4	21	7	21	-1.5	2.1
3	1st European	20	83	15	84	2.6	7.0
4	Heikenheimo	22	219	17	207	2.0	8.9
5	Austrian	37	352	65	376	-12.3[c]	21.9
6	Italian	19	164	18	157	0.1	8.2
7	Australian	51	376	63	371	-6.4	24.2
8	NHLBI SMIT	7	53	3	54	2.0	2.3
9	Frank	6	55	6	53	-0.1	2.7
10	Valere	11	49	9	42	0.2	3.9
11	UK	48	302	52	293	-2.8	20.8
12	Witchitz	5	32	5	26	-0.5	2.1
13	Lasierra	1	13	3	11	-1.2	0.9
14	3rd European	25	156	50	159	-12.1[c]	14.3
15	Olson	5	28	5	24	-0.4	2.0
16	Schreiber	1	19	4	19	-1.5	1.1
17	2nd European	69	373	94	357	-14.3[c]	31.7
18	2nd Frankfurt	13	102	29	104	-7.8[c]	8.4
19	Klein	4	14	1	9	1.0	1.0
20	N. German	63	249	51	234	4.2	21.8
21	Lipshultz	6	43	7	41	-0.7	2.8
22	Gormsen	2	14	3	14	-0.5	1.1
23	Brochier	2	60	8	60	-3.0[c]	2.3
24	European	41	172	34	169	3.2	14.7
	Totals	463	2,961	553	2.896	-51.4[c]	207.1

[a] $O - E$ refers to the difference between the observed and the expected number of deaths in the treated group (see main text).
[b] $V(O - E)$ refers to the variance of $O - E$.
[c] $p < 0.05$

SOURCE: Adapted from S. Yusuf, R. Collins, R. Pete, et al., "Intravenous and Intracoronary Fibrinolytic Therapy in Acute Myocardial Infarction: Overview of Results on Mortality and Side-effects from 33 Randomized Controlled Trials," **European** *Heart Journal* 6:556-558, **1985. Only data from studies of the effect of intravenous streptokinase are shown,**

groups combined. For the Fletcher study, this number is 5/23 (i.e., (1 +4)/(12 + 1 l)), or 21.7 percent. If the treatment had no effect, 21.7 percent of the subjects in the treatment and control groups should have died. The expected number of deaths in the treated group is 21.7 percent of 12, or 2.6. Expressed as a formula,

$$E = rid/N,$$

where:

- E is the expected number of deaths in the treatment group if there were no treatment effect,
- N is the total number of participants in the trial,
- n is the number of treated participants, and
- d is the number of deaths in the treated and control groups combined.

Thus, among the treated participants, one case was observed and 2.6 were expected. The difference, -1.6 (i.e., 1- 2.6), is the treatment effect for the Fletcher study (see table 4-A-l); it suggests that treatment reduced (by 1.6)[13] the number of deaths that occurred in the treatment group. Calculating the difference in the proportions of deaths

[13] One cannot, of course, actually reduce a fraction of a real death. Statistically, however, one is assuming that the 12 treated patients in the study are representative of a larger population. If that population were sampled many times, drawing a sample of 12 people each time, on average the deaths in each sample would be reduced by 1.6.

TABLE 4-A-2: Results on Fibrinolytic Agents and Survival After Myocardial Infarction with Intermediate Calculations for Homogeneity Chi-Square and Fixed-Effects Model

Study No. (1)	Author (2)	Treated Deaths (3)	Treated Total (4)	Control Deaths (5)	Control Total (6)	OR_i (7)	V_i (8)	W_i (9)	$(\ln OR_i - \ln OR_s)^2$ (10)	$(\ln OR_i - \ln OR_s)^2 * W_i$ (11)	$\ln OR_i * W_i$ (12)
1	Fletcher	1	12	4	11	0.16	1.48	0.67	2.55	1.72	-1.24
2	Dewar	4	21	7	21	0.47	0.52	1.91	0.26	0.50	-1.44
3	1st European	20	83	15	84	1.46	0.15	6.80	0.38	2.61	2.58
4	Heikenheimo	22	219	17	207	1.25	0.11	8.72	0.21	1.87	1.93
5	Austrian	37	352	65	376	0.56	0.05	20.49	0.11	2.30	-11.81
6	Italian	19	164	18	157	1.01	**0.12**	8.18	0.06	0.52	0.10
7	Australian	51	376	63	371	0.77	**0.04**	23.92	0.00	0.01	-6.34
8	NHLBI SMIT	7	53	3	54	2.59	**0.52**	1.93	1.42	2.74	1.84
9	Frank	6	55	6	53	0.96	**0.38**	2.67	0.04	0.11	-0.11
10	Valere	11	49	9	42	1.06	**0.26**	3.87	0.09	0.35	0.23
11	UK	48	302	52	293	0.88	**0.05**	20.77	0.01	0.25	-2.75
12	Witchitz	5	32	5	26	0.78	**0.48**	2.06	0.00	0.00	-0.52
13	Lasierra	1	13	3	11	0.22	1.54	0.65	1.59	1.03	-0.98
14	3rd European	25	156	50	159	0.42	0.08	13.02	0.40	5.26	-11.42
15	Olson	5	28	5	24	0.83	0.50	2.02	0.00	0.01	-0.39
16	Schreiber	1	19	4	19	0.21	1.37	0.73	1.76	1.28	-1.14
17	2nd European	69	373	94	357	0.64	0.03	31.03	0.05	1.41	-14.09
18	2nd Frankfurt	13	102	29	104	0.38	0.14	7.35	0.54	3.94	-7.16
19	Klein	4	14	1	9	3.20	1.48	0.68	1.97	1.34	0.79
20	N. German	63	249	51	234	1.22	0.05	21.59	0.19	4.11	4.21
21	Lipschultz	6	43	7	41	0.79	0.37	2.73	0.00	0.00	-0.65
22	Gormsen	2	14	3	14	0.61	1.01	0.99	0.06	0.06	-0.49
23	Brochier	2	60	8	60	0.22	0.66	1.51	1.57	2.38	-2.26
24	European	41	172	34	169	1.24	0.07	14.53	0.21	3.05	3.16
	Totals	463	2,961	553	2,896			198.83		36.86	-47.96

Summary Treatment Effect
Fixed-Effects Model, odds ratio 0.79 (95-percent confidence interval, 0.69-0.91)
Random-Effects Model, odds ratio 0.79 (95 percent confidence interval, 0.64-1.00)

NOTE: *OR* is the odds ratio; *V* is the variance of the logarithm of the odds ratio; *W* is the weight (= 1/*V*); *lnOR* is the natural logarithm of the odds ratio; subscript *i* indexes the study; and OR_s is the summary treatment effect from the fixed-effect model. See text for details.

SOURCE: Adapted from S. Yusuf, R. Collins, R. Peto, et al., "Intravenous and Intracoronary Fibrinolytic Therapy in Acute Myocardial Infarction: Overview of Results on Mortality, Reinfarction and Side-Effects from 33 Randomized Controlled Trials," *European Heart Journal* 6:556-558, 1985. Only data from studies of the effect of intravenous streptokinase are shown.

in the treatment and control groups, -28.1 percent (i.e., 8.3 to 36.3 percent) would have yielded a similar conclusion.

The second measure of treatment effect in common use is the odds of death in the treatment group divided by the odds of death in the control group (odds ratio). An odds ratio is, under usual circumstances, an approximation of the ratio of the rate of disease in the treated group to the rate of disease in the control group (rate ratio or, in more generic terms, the relative risk). In the Fletcher study (see table 4-A-2), the proportion of deaths in the treatment group was 8.3 percent (1/1 2), and the odds of death were (8.3 percent/(100 percent -8.3 percent)), or 0.091. (Note that 1/(12 - 1) is another way to calculate the odds and is equal to 0.091.) The odds of death for the control group were 4/7, or 0.571. The ratio of these two odds is 0.091/0.571, or 0.16, the odds ratio (see table 4-A-2). If the odds of death were the same in the treatment and control groups, the odds ratio would be 1. In the Fletcher study, the odds ratio was much smaller than 1, which suggests that treatment decreased the odds of death. Calculating the proportion ratio-(1/12)/(4/l 1), or 0.23—would show that the proportion of cases in the treated group was about one-quarter of that in the control group. Note again the similarity in conclusion, regardless of the particular method of calculating a treatment effect.

Although there are various methods for expressing treatment effects, the choice of the type of treatment effect calculated is somewhat arbitrary and is often based on tradition and interpretability as well as practical and theoretical statistical considerations. As a general rule, so long as the treatment effect is correctly interpreted, the manner of expressing the treatment effect is not important.

Calculating fhe Precision of the Measured Effect

For each expression of the size of the treatment effect, there is an associated value (a variance) that reflects the precision with which the treatment effect has been measured. This measure of precision is similar to the concept of a standard deviation and is, in fact, calculated as the square of the standard deviation. If the variance of a treatment effect is large, the treatment effect has not been precisely measured.

The variance of a treatment effect reflects the amount of information in the study. The result of a small study is imprecise and offers little information about the treatment effect. Conversely, the result of a large study is precise and conveys much information about the treatment effect. As precision (information) increases, variance decreases, and vice versa. In other words, the inverse of the variance of the treatment effect reflects the informativeness of the study results.

The variance of a treatment effect is calculated from a simple formula. Understanding why the formulas for variances are constructed as they are is not important for understanding the basic concepts of meta-analysis. Nonetheless, the following examples show how variances are calculated.

The treatment effect for the Fletcher study (see table 4-A-1) is -1.6. The variance associated with this value is

$$E(l - n/N)(N - d)/(N - 1),$$

where:

- E is the expected number of deaths in the treatment group if there were no treatment effect,
- N is the total number of participants in the trial,
- *n is* the number of treated participants, and
- d is the number of deaths in the treated and control groups combined.

E is equal to rid/N, as explained earlier. Thus, for the Fletcher data, the variance is 2.6(1 - 12/23)(23 - 5)/(23 - 1), or 1.02 (see table 4-A-1). The square root of the variance (1.02) is the standard error (like a standard deviation) of the treatment effect (O - E), which in this case is 1.01.

Taking the ratio of the treatment effect to its standard error (-1 .6/1 .01) yields -1.58, a statistic that can be used to test the significance of the treatment effect. The ratio of an observed treatment effect to its standard error reflects the probability that an effect as large as the observed treatment effect would have been found if, in fact, no real treatment effect were present. The ratio is compared

with values in a statistical table (of the Z distribution), which shows that if the absolute value of this ratio is <1.96, the probability of an effect of this size being observed by chance if no treatment effect existed is >0.05 (see box 4-2). If the probability of an observed treatment effect is >0.05, the analyst accepts the hypothesis that there was no evidence of a treatment effect in the study.[3] For the Fletcher data, with ratio-1.58, the treatment effect observed was not significantly different from the effect of no treatment.

The same example can be used to illustrate how the variance of an odds ratio is calculated. A popular variance formula for the odds ratio is complicated (78). In this example, because a much simpler formula (81) works nearly as well, it is presented instead. The variance can be estimated by the sum of the inverse of the number of deaths and nondeaths in the treatment and control group: 1/1 + 1/(12- 1)+ 1/4+ 1/(11 - 4), or 1.48 (see table 4-A-2). The standard error is 1.48, or 1.22. The statistic to test the significance of the odds ratio is obtained by dividing the natural logarithm of the odds ratio[4] by its standard error, which is ln(O. 16)/1.22, or -1.50. As before, because the absolute value of this ratio is <1.96, the probability of an effect of this size being observed by chance if no treatment effect existed is >0.05. Thus, the analyst would accept the hypothesis that there was no evidence of a treatment effect in this study. Note that the smallest value of the number of deaths and nondeaths in the treatment and control group-1 in this example—is the most important in determining the variance.

Summarizing the Treatment Effect Across Studies

The two methods used most frequently to summarize treatment effects across studies are the fixed-effects model and the random-effects model. The assumptions underlying the fixed-effects model are that all studies are estimating the same treatment effect and that the difference in the effects observed across studies results by chance. The assumptions underlying the random-effects model are that the treatment effect truly differs across studies and that the goal is to determine the average of the different effects. Although fixed-effects models were used frequently in the past, the use of random-effects models to summarize the treatment effect has been favored recently (70). In practice, if the results of the studies are homogeneous, the two approaches give the same result.

When the results from different studies are ready to be analyzed jointly, the analyst may choose either to search for different characteristics of the study designs or study populations that might account for the variation in results, or to evaluate the homogeneity of the results. If the results prove to be heterogeneous, varying considerably among studies, the analyst has two options:

1. to refrain from summarizing the results across studies (summarizing, instead, within groups that have similar results or not calculating a summary at all, if the results are markedly heterogeneous), or
2. to summarize the results across studies using a random-effects model.

One could also summarize heterogeneous results using a fixed-effects model, but although this has been a common practice in the past it is no longer recommended.

The search for characteristics of the studies or study populations that might account for variation in results can be undertaken by grouping the study results according to the characteristic under study and summarizing results within groups. The sum-

[3] In statistical terms, the null hypothesis (that there is no treatment effect) is being tested. If the probability of a given treatment effect's being observed by chance, if in fact the null hypothesis is true, is >0.05, then in formal statistical terms one would fail to reject the null hypothesis.

[4] The (natural) logarithmic scale is used with ratio measures of effect (such as odds ratios) so that treatment effects of the same size, but in opposite directions (like an odds ratio of 0.5 and 2; 1/0.5=2) will be of equal absolute size arithmetically. The natural logarithm of 0.5 is -0.6931; the natural logarithm of 2 is 0.6931. The sampling theories upon which the relevant statistics are based work only when such symmetry is present.

mary treatment effects are then compared across groups. Regression techniques that effectively accomplish the same goal can also be used.

Evaluating the Homogeneity of Results

Results sometimes vary greatly and inexplicably from study to study, which may influence how a meta-analysis is interpreted. If the study results are markedly heterogeneous, for example, one might have little confidence in one's ability to predict the effect of treatment in any future study.

The evaluation of the homogeneity of results across studies in a meta-analysis is based on the homogeneity chi-square statistic(81). This statistic is a sum across all studies of the square of the difference between the study-specific treatment effects and the summary treatment effect, multiplied by the inverse variance of the study-specific treatment effect. In statistical terms, the homogeneity chi-square statistic is as follows:

$$\chi^2=\Sigma w_i(te_i - te_s)^2,$$

where:

- te_i is a study-specific treatment effect,
- te_s is the summary treatment effect (described below),
- $w_i = 1/v_i$ (v_i is the study-specific variance), and
- i indexes the study.

Thus, the squared difference of each study's result from the overall average is weighted by the precision of the study. In this way, the deviation of a small study from the summary treatment effect contributes little to the homogeneity statistic, whereas the deviation of a large study from the summary treatment effect contributes much more. This makes sense intuitively, because smaller studies are more likely to deviate from an overall mean by virtue of sampling error alone (4). Deviation of a large study from the overall mean suggests that the studies in the meta-analysis may have been samples from populations in which the treatment effects differed. The expected size of the homogeneity chi-square statistic is based on the number of studies in the meta-analysis and is found in a statistical table for values of chi-square.

The statistical evaluation of homogeneity can be illustrated using the meta-analysis of fibrolytic therapy discussed previously. The calculation of the homogeneity chi-square statistic is based on several columns in table 4-A-2. Column 7 contains the study-specific odds ratio, and column 9 contains the study weight (l/vi). Column 10 contains the squared difference of the logarithm of the study-specific odds ratio from the logarithm of the summary odds ratio (described below). Column 11 contains the product of the study weight and the squared difference of the effects. At the bottom of column 11 is the sum of the contribution of each study, which is the homogeneity chi-square statistic. In this example, the chi-square statistic is 36.9. The value expected under homogeneity is 35.2 or less. (This is the value from a chi-square table for 23 degrees of freedom and p=O.05. The degrees of freedom are the number of studies minus 1.) Thus, the variation in study results is greater than expected, suggesting that something other than chance accounts for the differences in the findings. Perhaps the effect of the drug differs depending on the exact dose used, the patient population, the length of time the patients were studied, or some other factor.

One problem with the homogeneity chi-square statistic is that it may not detect a variation that has biologic or practical importance. Therefore, a search for factors related to study results is recommended, regardless of whether statistical homogeneity is present.

Combining Results Across Studies

Once homogeneity has been evaluated, the results across studies may be summarized, if deemed appropriate, using a fixed-effects model or a random-effects model. In a fixed-effects model, the contribution of each study within the meta-analysis to the summary treatment effect is inversely proportional to its variance. Thus, larger studies contribute more to the summary treatment effect because they have smaller variances. Random-effects models, however, weight the contribution of individual studies according to their inverse variance and according to a measure of the variability of results across studies. In random-effects mod-

els, as the degree of heterogeneity increases, the studies tend to be given more equal weight, as in a simple average. An advantage of the random-effects summary for heterogeneous results is that the estimate of the summary effect is less precise than that calculated in a fixed-effects model, reflecting the greater degree of variation in the study results.

The random-effects model is now generally considered preferable where substantial heterogeneity exists (70). The illustration below of the method of calculating a summary treatment effect across studies uses a fixed-effects model, however, because that procedure is more straightforward and reflects the essential points about how the results from different studies are combined in a meta-analysis.

To illustrate the fixed-effects model, assume that the treatment effects for the data from the Yusuf meta-analysis were homogeneous. The most straightforward method of combining results across studies is to calculate the simple average of the treatment effects, in which the results of each study carry equal weight. The fixed-effects model, however, is a weighted average in which each treatment effect is weighted by the inverse of the precision of the estimate (inverse variance weights). The previous section described how to calculate the observed deaths minus the expected number of deaths in the treatment group, O - E, and its variance, v(O - E), for an individual study. Summing the O - E across studies is a form of inverse-variance-weighted summary treatment effect. Note that if there were no treatment effect, the sum of O - E would be zero. If the treatment reduced the number of observed cases by 10 percent, the sum would change accordingly. The 0- E of a study with a large number of treated subjects would be larger than that of a study with a small number of treated subjects. In this way, the larger studies contribute more to the summary treatment effects

The sum of 0- E for all studies of fibrinolytic therapy is -51.4 (seethe bottom line of column 7 in table 4-A-l). (In statistical notation, this is $Z(O_i - E_i)$, where i indexes the study and the summation is across all studies.) In other words, there were 51.4 fewer deaths than expected in all the treatment groups combined. The variance of this summary treatment effect is the sum of the variances of each treatment effect, which is 207.1 (see the bottom line of column 8 in table 4-A- 1). In statistical notation, this is Zvi(Oi - E_i). The square root of 207.1 is 14.4, the standard error; taking the ratio of the summary treatment effect to its standard error (-51.4/14.4) yields -3.57, which has an absolute value greater than 1.96 and thus is statistically significant at the $p<O.05$ level. Therefore, the meta-analysis supports the conclusion that fibrinolytic therapy is effective in reducing death after myocardial infarction.

Using data from the same example, one can calculate the summary treatment effect as an odds ratio. This approach more directly illustrates the principle of the weighted average. The formula for the natural logarithm of the summary odds ratio is

$$Zwiln(ORi)/Zwi,$$

where $W_i = l/v_i$ (the inverse variance), *OR* is the odds ratio, and i indexes across study results. In other words, the weight for each study result is multiplied by the natural logarithm of the study's odds ratio. This quantity is then summed across all i studies (see the bottom line of column 12 in table 4-A-2) and divided by the sum of all the study weights (see the bottom line of column 9 in table 4-A-2). In the example, this yields an answer of -48.0/199, or -0.24, the natural logarithm of the summary odds ratio. Exponentiation of the logarithm of the summary odds ratio (e^{-024}) gives the

[5] Although the O - E method has been very popular in the past for meta-analysis of randomized trial data and probably provides the right answer in most cases, the method has been shown to be misleading in some situations (41). One advantage to the O - E method is that the results are relatively easy to understand and explain. Because of the problems with occasional misleading answers, however, meta-analysts recently have favored methods based on odds ratios.

odds ratio, 0.79. The variance of the summary log odds ratio is l/Xwl, or 0.005 (i.e., 1/199). The square root of the variance is the standard error of the logarithm of the odds ratio, or 0.071. Calculating the ratio of the summary logarithm of the odds ratio to its standard error (-0.24/0.071) yields -3.38, which has an absolute value greater than 1.96 and thus is significant at the p<O.05 level. This significance means that it is unlikely that a treatment effect this large would have been observed by chance if there truly were no treatment effect. Thus, the results suggest a benefit of treatment.

The standard error of the logarithm of the odds ratio can also be used to calculate the confidence limits of an odds ratio. The width of the confidence interval is proportional to the standard error of the odds ratio. Thus, a large confidence interval implies small precision, and vice versa. A confidence limit for an odds ratio that excludes 1 indicates that the treatment effect is statistically significant. The 95-percent confidence interval around the fixed-effects estimate for the example data is 0.69 to 0.91, which excludes 1 (see table 4-A-2).

The details of calculating a random-effects summary are beyond the scope of this document, although the calculation is not markedly more complicated than the procedures illustrated above. In this example, the random-effects model summary treatment effect is an odds ratio of 0.79 (95-percent confidence interval 0.64 to 1.00). Note that the confidence interval around this estimate is wider than the confidence interval around the fixed-effects estimate (see table 4-A-l). The greater width of the confidence interval reflects the fact that the study results were more variable than would be expected if chance were the only reason for variation. The heterogeneity in this example was relatively small; if more marked heterogeneity were present, the difference between the results of the fixed-effects and random-effects models would be greater (7).

Regression methods can also be used to combine studies' results (8,39,40,82). The regression approach allows the shape of dose-response curves to be estimated and provides a convenient method for identifying patterns in study results associated with characteristics of the study populations or study designs. Both fixed-effects and random-effects regression models can be constructed.

The frequentist approach, which is used routinely in medical meta-analysis, has been used for summarizing the treatment effects presented in this appendix. Another method of summarizing treatment effects is the Bayesian approach (see box 4-6 in main text). The Bayesian meta-analyst specifies his or her belief about the size of a treatment effect and the certainty about that belief prior to examining any of the results of the studies in a meta-analysis (28,29,77). In the absence of a strong prior belief, the Bayesian meta-analyst may find all possible values of the treatment effect equally likely (and thus has no certainty about its size). The results of the studies in the meta-analy - sis are then used to modify the analyst's belief about the size of the treatment effect. The result is an expression of the analyst's belief about the size of the treatment effect that primarily reflects the results of the studies and only minimally reflects the prior belief. The contribution of the prior belief (or data) relative to the contribution of the new data (the studies in the meta-analysis) depends on the strength of evidence from each source. For example, when there is much prior information about the size of a treatment effect and only a few small studies are in the meta-analysis, the result of including the new data in the synthesis may not much alter the estimate of the treatment effect. In practice, Bayesian methods give quantitative results that are similar to those from a random-effects model.