RECONCILABLE DIFFERENCES

Gleaning Insight from Conflicting Scientific Studies

by INGRAM OLKIN

To some people the promise of science as the ultimate arbiter of truth dissolved once and for all with the oat bran affair. As the events played out in the pages of the daily newspapers, the public caught an unusual glimpse of a scientific community in excruciating indecision, seesawing wildly over what appeared to be a simple question: Does oat bran lower cholesterol? The episode began quietly with a small study done in 1984 by the endocrinologist James W. Anderson of the University of Kentucky. Following hints from previous studies that the water-soluble fiber in oat bran and beans might reduce cholesterol, Anderson and his coworkers added each of the foodstuffs to the diets of ten men with dangerously high levels of cholesterol: greater than 260 milligrams per deciliter. To match the amount of fiber taken in by the men who ate beans, the ten men on the oat bran plan ate a bowl of oatmeal every morning and five oat bran muffins throughout each day. The results were startling: in only three weeks the average cholesterol levels for each group dropped by 19 percent.

A few years later the oat bran spark caught fire. In 1987 Robert E. Kowalski’s book *The Eight-Week Cholesterol Cure* proclaimed the health-giving properties of oat bran. The following year Anderson wrote a review of high-fiber-diet studies for the *American Journal of Clinical Nutrition*, highlighting the studies he had done with beans and oat bran. A short time later an article in the *Journal of the American Medical Association* reported that oat bran is a more cost-effective treatment for high cholesterol than any available drug. Oat bran had arrived.

At the end of 1988 the Quaker Oats Company reported that sales of oat bran had increased six times and that the company was having trouble keeping up with demand. Suddenly oat bran muffins were everywhere. David’s Cookies alone was selling 100,000 of the muffins a week and could have tripled that number, the owner lamented,
if not for a shortage of oat bran. Even though most people never came close to eating the quantities consumed in Anderson's studies, oat bran had become "the elixir of choice for the health-conscious consumer" and a "soaring success," according to The New York Times. Cheers, after being advertised as an excellent source of oat bran, overtook Kellogg's Frosted Flakes as the number-one cereal in the United States, and the makers of snack bars and even beer began touting oat ingredients.

But it was not long before a wet blanket was thrown on the oat bran explosion. A study from Brigham and Women's Hospital in Boston, appearing in the New England Journal of Medicine in 1990, showed that refined wheat has the same effect as oat bran on cholesterol levels, even though it has none of the soluble fiber that oat bran has. The authors of the study proposed that both oats and refined wheat reduce cholesterol merely by replacing fatty foods that would be eaten otherwise. Furthermore, they noted that the refined wheat does not produce the nasty symptoms of gastrointestinal distress that oat bran does. Anderson and others still dispute those conclusions, proposing that oat bran had little effect in the Brigham study because the subjects began with normal cholesterol levels.

The irony is that all available oat bran studies have been meticulously analyzed before the furor of the late 1980s, the evidence for a health effect would have looked flimsy indeed. A handful of small trials studying only eight to fifteen people—minute samples by statisticians' standards—did suggest oat bran has a positive effect. But the results of other studies were marginal.

The oat bran fiasco is just one in a long string of frustrations the public has suffered with scientific findings. In 1981 a study alarmed coffee drinkers with its conclusion that coffee may cause slightly more than half of all pancreatic cancer, but later studies failed to support the findings. Such notorious incidents raise a number of questions for science: How can people be persuaded not to become exasperated and discount all the claims of science? How can science, in the face of conflicting evidence, get to the truth on matters of such importance as medicine, health and education? More to the point, how does one extract the truth in a timely manner, thereby perhaps saving lives and discouraging the premature reporting of findings such as those that set off the oat bran roller coaster?

Many people—and I am among them—argue that the answer is to apply the rigorous methods of science to the analysis of the studies themselves. The state of modern science virtually demands it. Each year more than 23,000 scientific journals are published, and investigators carry out more than 9,000 studies. The tools for making sense of that deluge of data are already available in the practice of meta-analysis. By applying statistical techniques, meta-analysis provides a quantitative synthesis of data from a group of studies on a given question and thereby arrives at a conclusion based on a much larger sample than any single study. If the process is carried out properly, it often exposes the weaknesses of individual studies, and it can elicit findings that were not intended by the original investigators. Unfortunately, although meta-analysis was established in its modern form more than thirty years ago, its promise has not yet been fully realized. Many investigators have been reluctant to combine the results of disparate studies because of differences in, say, protocol and treatments. To be sure, meticulous care is needed to take account of such differences. But with powerful contemporary techniques meta-analysis can address scholarly scruples and still offer a clear view through the obscuring mass of scientific information.

As one might guess, the idea of combining the results from a number of small studies in the hope of arriving at a stronger conclusion is far from a recent revelation. In 1904 the English statistician Karl Pearson grouped the statistics from British mil-
itary and medical installations in South Africa and India and concluded that a vaccination against intestinal fever was not sufficiently effective to warrant its routine use. Later, in the 1920s, the English geneticist and statistician Ronald A. Fisher began applying statistical analysis to various experiments conducted during sixty years of agricultural fieldwork at the Rothamstead Experimental Station. But large-scale meta-analysis was not practiced until the 1970s, when social scientists began to evaluate the effects of President Johnson’s Great Society programs. The term meta-analysis was coined in 1976 by the educational psychologist Gene V. Glass of Arizona State University in Tempe, though the practice of combining information from independent studies also went by various other names well into the 1980s. Some called it integration, synthesis or the pooling of results; others referred to it as an overview, but now meta-analysis is more firmly established, especially in medicine.

Meta-analysis got its start in the social and behavioral sciences, because there are so many confounding factors capable of invalidating a tentative conclusion. Yet the technique has also been applied for decades in physics and chemistry. There most applications have been relatively straightforward: to make increasingly refined estimates of physical constants, for instance, the measurements and error estimates from a number of experiments are combined. But the physical sciences are not immune to the problem of contradictory and inconclusive studies. After more than 750 studies meteorologists are still uncertain about which methods of cloud seeding, under which conditions, work best to induce rain.

Careful, quantitative meta-analysis remains an unusual practice in virtually every discipline. What is far more common is the scientific review article, in which a specialist discusses a number of studies pertaining to some question of interest. Such articles are not unlike film reviews, and each literary form is highly vulnerable to the reviewer’s biases and weaknesses. In his 1988 review article on oat bran Anderson put the findings of previous studies in a starkly different light than did the team from Brigham and Women’s Hospital.

Even when massive funds are available for synthesizing research findings, the syntheses often take the form of qualitative expert judgments instead of quantitative meta-analyses. The National Acid Precipitation Assessment Program—a $500 million project mandated by Congress in 1980 to assemble all the government research on acid rain—relied in large part on a panel of experts to assess the conclusions of various studies. The report of the program, published late last year, included star ratings next to individual statements to signal the level of confidence in the finding. Unsupported statements had little or no weight in the final report; statements backed up by more substantial evidence were assigned between one and four stars, to reflect increasing confidence by the panel. Although the panel members were encouraged to use meta-analysis wherever possible, they based most of their ratings primarily on judgment.

When direct comparisons are made, quantitative meta-analysis can offer clear advantages over narrative reviews. In 1980 the psychologists Harris Cooper, now at the University of Missouri, and Robert Rosenthal of Harvard University asked forty-one investigators to consolidate the results of seven studies that assessed how a person’s tenacity of effort might vary by sex. The investigators who conducted a traditional review concluded that men and women exert no significant difference in effort. But the group that performed a meta-analysis found that women exert slightly greater efforts than men. Narrative reviews do have value, but meta-analysis is undeniably superior when small effects are being gauged.

The discipline with perhaps the most urgent need for the “resolving power” of meta-analysis is medicine. As a family doctor in the late 1940s and early 1950s, Thomas C. Chalmers, who is now associate director of the technology assessment group at the Harvard School of Public Health, was alarmed at the numerous, sometimes fatal mistakes that could be traced to imprecise or missing information about the benefits and side effects of various treatments. The chief problem, he maintained, was that routinely prescribed drugs were not subjected to adequate clinical trials. In response, Chalmers left family practice after six years to organize trials of his own. Later, as associate director of clinical care at the National Institutes of Health and as dean and president of the Mount Sinai School of Medicine and Hospital, he became frustrated again by the failure of medical administrators to incorporate the latest research findings into their clinical work. As he bluntly puts it, the situation made all administrators incompetent to perform their jobs. By the mid-1970s he was teaching himself statistical techniques that could distill and bring together research findings for clinicians. When a colleague pointed out that what he was doing was an established discipline called meta-analysis, he became one of its foremost champions.
In one of his earliest applications of meta-analysis, Chalmers studied the effectiveness of lidocaine, a drug widely recommended in the 1970s for patients prone to heart attack. Cardiologists had noticed that the drug suppressed irregularities in heart rhythm when it was used as a local anesthetic in heart surgery. General practitioners then began to prescribe lidocaine, because they assumed its calming effect on the heart would prevent attacks. But Chalmers showed no correlation between lidocaine usage and the death rate of heart patients; the drug had no beneficial effect at all. Yet so little heed is paid to Chalmers’s meta-analysis that even today many medical textbooks still recommend lidocaine for treating heart patients.

In other instances truly effective medical treatments have been obscured by a fog of nondefinitive studies. The role of aspirin in preventing the recurrence of heart attack is a case in point. Many small studies in the 1970s and late last year Peto and his Oxford team reported the striking results of a meta-analysis of 133 studies on the effectiveness of various treatments following surgery for breast cancer. Although the studies focused on how chemotherapy or hormone therapy prevents recurrence in the short term, Peto discovered that the hormone tamoxifen has a beneficial impact even years after treatment has ended. Women treated with tamoxifen for even brief periods are more likely to be alive ten years later than are breast cancer patients who do not receive follow-up therapy. Peto estimates that the treatment could save 3,000 lives a year in the U.S. The study also led to a consensus among physicians in favor of prescribing tamoxifen therapy for postmenopausal women with breast cancer.

As Chalmers points out, the days of discovering such medicines as penicillin that have “slam-bang effects” are over. The future of medicine lies in seeking agents with incremental effects. Moreover, because funding is parceled out among many laboratories, most medical research will continue to be conducted on small samples, as it was in the aspirin studies. And, just as it did in those studies, meta-analysis can amass a large sample by drawing on the results of many small studies.

The mechanics of meta-analysis have evolved in the past few decades from the crude to the elegantly sophisticated. One of the earliest, remotely quantitative approaches to analyzing a group of studies was vote counting: each study whose findings are statistically significant gets one vote for or against the hypothesis in question. Vote counting is often the underlying—and unstated—basis of a narrative review, but the method has also been applied overtly. In 1982 British psychologists Peter Warr and Glenys Perry of the University of Sheffield essentially counted votes to evaluate thirty-eight studies of how working outside the home affects the psychological well-being of women. None of the studies detected adverse effects, but the findings were inconclusive, because many of the results were not statistically significant and because the studies had adopted various ways of measuring mental health.

Warr and Perry compared the number of studies that had statistically significant results with the ones that had nonsignificant findings, for each of six measures of mental health: suicide rate, diagnosed psychiatric illness, psychological distress and the like. The investigators also tabulated the findings for various populations of women, including married women with children, married women without children and single women. But as Richard J. Light and David B. Pillemer point out in their 1984 book Summing Up: The Science of Reviewing Research, the analysis was of little value, because it said nothing about the size of the sample or the quality of the studies, much less about the magnitude of any effect. Warr and Perry themselves cautioned that their tables had to be understood in conjunction with information appearing only in the text of the review. Larry V. Hedges and I outlined the serious shortcomings of vote counting in our book Statistical Methods for Meta-analysis. Thus, in spite of its appealing simplicity, the method is rarely used anymore.

Another simplified approach to meta-analysis is to average the results of a number of studies. That technique too is fraught with peril; it is successful only when the

Lisa Milroy, Dresses, 1985

1980s suggested that aspirin exerts a positive effect, but because the sample sizes were too small, the results were inconclusive. In 1983 Richard Peto, codirector of the Clinical Trial Service Unit at the University of Oxford, carried out a meta-analysis of trials on aspirin and other antiplatelet drugs and found that, taken together, the studies provide convincing support for the benefit of taking aspirin. In response, the U.S. Food and Drug Administration approved aspirin in 1985 for survivors of myocardial infarction. By retracing the history of the trials, Chalmers reported that a meta-analysis could have demonstrated the effectiveness of aspirin as long ago as 1976. He estimates that such a meta-analysis might have saved between 10,000 and 20,000 lives in the U.S. alone.

Meta-analyses can summon unexpected findings from studies that initially addressed a slightly different issue.

July/August 1992 • THE SCIENCES 33
Studies are strictly uniform, gauging the same effects under the same controlled conditions. Such a fortuitous confluence of conditions across studies is rare. Typically, studies investigate diverse combinations of, say, sex and age, or they administer varying doses of drugs on different schedules. Although the single number that comes from averaging is also an attractive simplification, it usually masks underlying effects.

The modern meta-analyst applies a battery of sophisticated statistical tools to make more robust comparisons of results across studies. One important measure is the effect size, which in turn depends on the statistical quantity known as standard deviation. The standard deviation is a measure of how far the individual effects vary from the group mean, or average; the more they scatter about the mean, the higher the standard deviation. If ten fourteen-year-olds in a sample each measured five feet tall, there would be no scatter about the average, and the standard deviation of the heights would be zero. Suppose, on the other hand, the heights ranged a few inches above and below five feet. Then the standard deviation would be the distance, in inches, on each side of the average that encompassed the heights of 68 percent of the teenagers; 95 percent of the teenagers would fall within two standard deviations.

The effect size can then be calculated in a number of ways, depending on the data under scrutiny. One common calculation is to subtract the effect measured in the control group from that in the treated group and then divide the result by the standard deviation of the control group’s effects. Finding the difference gives the raw size of the effect; dividing by the standard deviation puts the value of the measured effect into perspective. For example, if the effects varied greatly, or in other words if the standard deviation were large, a given raw difference between the control group and the treatment group effect would appear comparatively small.

Most review articles rely on statistical significance alone to indicate a valid effect. But statistical significance is merely the probability that a measured effect is a result of chance or error. If the effect is small, it may not be worthwhile even if it is genuine. A drug that leads to a statistically significant reduction in cholesterol levels, for instance, will not be useful if the reduction is only one milligram per deciliter. Minuscule significant effects tend to turn up in studies with large samples, because expanding the sample size in itself mathematically increases statistical significance. In general, effect sizes become noticeable at around 0.5, half a standard deviation. The difference in heights between fourteen-year-old and eighteen-year-old girls is generally noticeable, and the size of the effect of four years of aging on the average height of fourteen-year-old girls is 0.5. In contrast, the difference in heights between fifteen- and sixteen-year-old girls is small, and indeed, the effect size is 0.2.

When the outcome of a study is not a mean value but one of two possibilities, the effect size must be defined in a different way. Studies that investigate whether aspirin can help patients with heart conditions commonly monitor whether the patients live or die over a period. In one study of 615 heart attack victims who took aspirin regularly after their attacks, 8 percent died over a set period, compared with 11 percent of a control group of 624 patients who took placebos.

One way of comparing those results with the results of other studies might be to take note of the raw difference between the two death rates: 3 percentage points. But a difference of 3 points carries more weight in the neighborhood of a 10 percent death rate than it does near an 80 percent death rate. A better measure of the size of the effect is the relative risk ratio, here the ratio of the two death rates, 8 divided by 11, or 0.73. Thus the study concludes that for heart attack patients the risk of dying can be reduced by 27 percent if they take aspirin every day. To define better the neighborhood of the effect, a comparison of the survival rates for each group can be brought into the calculation. In the control group, 89 percent sur-
vived, compared with 92 percent of the treated group. The ratio of 8/92 to 11/89 is 0.70, the odds ratio; it is the most common way of calculating effect size in medicine.

Once effect sizes have been determined for all the studies in a meta-analysis, the next step is to combine them. The meta-analyst may want to take into account study quality and sample size by assigning relative weights to the studies. High-quality studies, which typically involve biostatisticians and several research centers, are tightly controlled. The studies can also be statistically tested for homogeneity: Could the variation in effect sizes be expected from sampling error—the error inherently introduced when only a portion of a larger population is tested—alone? The smaller the sample, the larger the sampling error and the greater the expected variation in the outcome.

If the studies are too heterogeneous to be combined, all is not lost. The distribution of effect sizes often reveals hidden trends. Variations noted in the effectiveness of Head Start programs across a number of studies could lead to the conclusion that some settings or circumstances are more conducive to effective programs than others—small towns versus large urban centers, for instance. Different hospitals may have different death rates associated with a given treatment. An analysis of the effect sizes of population subgroups across many studies often turns up unexpected findings.

One way to bring out hidden patterns is to plot the number of studies whose effect sizes fall into various ranges: a histogram of the effect sizes. In 1982 a meta-analysis was conducted on fifty-one studies that each had investigated

Lisa Milroy, Shoes, 1986

July/August 1992 • THE SCIENCES 35
META-ANALYSIS is not without its detractors. The most common complaint is a philosophical one: data from different studies should not be combined because of unavoidable variations in study conditions, both stated and unstated. It is the primal scientific fear of combining “apples and oranges.” And to be sure, the uneven quality of scientific studies in all fields today can be enough to send the budding meta-analyst back to the independent life of the laboratory investigator.

But workers have been forced to confront such problems in single studies and narrative reviews as well. And some investigators, in such disciplines as epidemiology, must also deal with the uncertainties of secondhand observation, just as the meta-analyst does. Epidemiologists must take a great deal on faith when they scour the data banks of hospitals or scrutinize the answers to questionnaires to establish a correlation between disease factors.

Critics are also quick to point out that bias can intrude whenever the meta-analyst decides to reject a study or to assign quality ratings and weightings. Undeniably, those procedures must be handled with care. But experienced meta-analysts in various fields have drawn up strict guidelines for rating study quality and for rejecting poor studies. Studies that randomly assign patients to a treatment or that use double blinding usually get high ratings, and studies that monitor “hard” endpoints such as death or disease rates are preferred to those with “soft” endpoints such as depression or insomnia. While good statistical techniques are not likely to correct the conclusions of bad studies, they can ensure that valid conclusions from good studies do not go unnoticed.

ONE REAL DANGER of meta-analysis is the injudicious use of subgroup analysis. When no overall effect results from a meta-analysis, investigators sometimes indulge in so-called data dredging, in which they search all conceivable subgroups for an effect, often with computer assistance. That approach almost inevitably turns up a spurious effect. Another recurring criticism of meta-analysis is that it incorporates what is known as publication bias: because journals tend to publish only studies that find statistically significant effects, analyses that culled studies from published reports are likely biased toward finding an effect. But a recent proposal could eliminate the ill effects of publication bias altogether. If experimental results, published or otherwise, were filed with a national registry on a computer network, computerized meta-analyses could be done continuously and could flag statistically significant outcomes. Conclusions could be reached in a timely way with such a plan, and the plan could also prevent research results from being reported prematurely, thereby avoiding repeats of the oat bran fad. The plan is still in the conceptual stages, but its proponents envision a network sponsored by a government agency such as the National Institutes of Health.

Whatever the weaknesses of meta-analysis may be, the technique is a clear improvement over the narrative review for most scientific questions. Indeed, for many issues, meta-analysis may bring scientists as close as they can hope to come to the absolute answers they seek. As the statistician John W. Tukey, professor emeritus of Princeton University, once wrote, science “does not begin with a tidy question, nor does it end with a tidy answer.” Medicine, psychology, environmental health and social science will benefit the most from meta-analysis, but every discipline stands to gain. If investigators can resist the temptation to regard meta-analysis as an absolutely objective solution, they will find the method a valuable tool in discovering the truths hidden in the masses of data that inundate science.

INGRAM OLKIN is professor of statistics and education at Stanford University. He is the coauthor, with Larry V. Hedges, of one of the earliest books on meta-analysis, STATISTICAL METHODS FOR META-ANALYSIS, published by Academic Press in 1985. He and Hedges are working with Thomas C. Chalmers and Joseph Lau on a book about meta-analysis in the health sciences.