Annu. Rev. Public Health 2006. 27:81–102
doi: 10.1146/annurev.publhealth.27.021405.102239

THE PROMISE AND PITFALLS OF SYSTEMATIC REVIEWS

Patricia Dolan Mullen[1] and Gilbert Ramírez[2]
[1]*Center for Health Promotion and Prevention Research, School of Public Health, University of Texas Health Science Center at Houston, Texas 77030; email: Patricia.D.Mullen@uth.tmc.edu*
[2]*College of Health Sciences, Des Moines University, Des Moines, Iowa 50312-4198; email: Gilbert.Ramirez@dmu.edu*

Key Words meta-analysis, review literature, guidelines, public health, intervention studies

■ **Abstract** The systematic review "movement" that has transformed medical journal reports of clinical trials and reviews of clinical trials has taken hold in public health, with the most recent milestone, the publication of the first edition of *The Guide to Community Health Services* in 2005. In this paper we define and distinguish current terms, point out important resources for systematic reviews, describe the impact of systematic review on the quality of primary studies and summaries of the evidence, and provide perspectives on the promise of systematic reviews for shaping the agenda for public health research. Several pitfalls are discussed, including a false sense of rigor implied by the terms "systematic review" and "meta-analysis" and substantial variation in the validity of claims that a particular intervention is "evidence based," and the difficulty of translating conclusions from systematic reviews into public health advocacy and practice.

INTRODUCTION

> *If you are doing almost anything related to health care today, being "evidence based" is de rigueur. (93, p. 80)*

In the past three decades, the science of systematic review has been an important and growing edge of research methods, with identified statistical techniques and experts (24, 32, 36, 42, 43, 47, 48, 84, 89), production centers (2, 16, 19), software (12, 19, 25, 63), quality standards (66, 69, 95), glossaries (19), and bodies of literature (3, 4, 18, 28, 29, 98, 106).

Caches of reviews, many regularly updated and available on the Internet (see examples above) have the potential to identify and speed the adoption of new practices. News and radio journalists have learned how to describe a systematic review or meta-analysis as a new study with new findings. Nomenclature has evolved over the years, "meta-evaluation," "research synthesis," "meta-analysis," "integrative

review," to what appears to be a relatively stable consensus of "systematic review" in medicine and public health as the generic name for the enterprise and "meta-analysis" for the subset of systematic reviews that combine data from individual primary studies and use specific statistical techniques. The distinctive terms used earlier signaled a break with the traditional literature review. With clearer understanding about the new approach, the simpler term suffices, with the modest modifier, systematic, summing up the difference.

In this paper we aim to provide balanced, if not encyclopedic or systematically derived, comments on the systematic review and public health, its achievements, promise, and pitfalls. Our experience as collaborators and individually is in systematic reviews and meta-analyses in clinical services, counseling and patient education, and complementary medicine; The Cochrane Collaboration; the U.S. Agency for Healthcare Research and Quality clinical guideline panels; and later, with evidence-based practice centers, the U.S. Clinical Preventive Services Task Force; the U.S. Community Preventive Services Task Force; the U.S. Centers for Disease Control and Prevention HIV/AIDS Prevention Research Synthesis Project; the Campbell Collaboration Social Welfare Review Group (which includes public health); and as trainers and teachers.

RATIONALE FOR LITERATURE REVIEWS AS A SCIENTIFIC ENTERPRISE

Recognition of the need to upgrade the standards for literature reviews arose in part because of the increasing difficulty of keeping up with the literature in one of the many fields undergoing a disorderly explosion of studies. For example, the problem presented by the exponential growth of research reports in the social sciences in the middle years of the twentieth century spurred Glass to coin the term "meta-analysis" in 1976 (37) for the work he and colleagues had undertaken on the relationship between class size and school achievement (39) and the effectiveness of various types of psychotherapy (90) and to follow it with the first comprehensive methods book for meta-analysis (38). The importance of "trustworthy accounts of past research . . . [as] a necessary condition for orderly knowledge building" (26, p. 9) was echoed by many other critics of the lack of rigor and scientific method in the traditional subjective literature review (25, 75). Critics noted that in the social sciences, for example, the substantial attention paid to validity issues in primary research was not matched by similar concern for the validity of review outcomes (26). Adding to the call for more orderly methods was the recognition that reviews are among the most frequently cited reports, and they can play a very influential role in the collective understanding of what is known (23, 25, 75).

Stated another way, "scientific subliteratures are cluttered with repeated studies of the same phenomena. Repetitive studies arise because investigators are unaware of what others are doing, because they are skeptical about the results of past studies, and/or because they wish to extend . . . previous findings . . . [yet even when strict

replication is attempted] results across studies are rarely identical at any high level of precision, even in the physical sciences. . .” (23, p. 4). “. . .[L]ocating and integrating separate research projects involves inferences as central to the validity of knowledge as the inferences involved in primary stud[ies]. . .” (26, p. 10).

Meanwhile, critical appraisal was directed to the traditional, idiosyncratic review with its implicit methods. For example, 50 review articles published in the four U.S. medical journals with circulations over 50,000 from June 1985 to June 1986 were rated on eight criteria. Although 40 of the 50 reviews stated a specific purpose, only 1 of the 50 clearly identified the sources and methods of citation search or described inclusion criteria; 43 made explicit reference to the strengths and weaknesses of the included studies, but only 1 described a standardized assessment of study quality; and, although 37 summarized relevant findings, in 12 cases it was not clear whether they had done so (74).

SYSTEMATIC REVIEW AND META-ANALYSIS SHARE SIMILAR METHODS

All types of literature reviews can benefit from a systematic approach, with the exception perhaps of a review that aims to argue for a particular position and that would select evidence accordingly. Our specific focus here is the review with the goal of integrating empiric research for the purpose of generalizing from a group of studies. Implied in this goal is that the reviewer is also seeking to discover the limits of the generalizations. The corollary to this goal is to identify inconsistencies and account for variability in a group of similar-appearing studies (23, pp. 4–5).

One way to think about a systematic review is as analogous to a primary study. Thus, the steps are parallel for both a primary study and a systematic review:

1. Specify the study’s aims,
2. Set inclusion criteria for participants/evidence,
3. Design the recruitment/search strategy,
4. Screen potential participants/evidence against inclusion criteria,
5. Decide on measures and design the data collection protocol,
6. Select an appropriate metric to represent the magnitude of the findings and assess the likelihood that these findings could be the result of chance,
7. Collect the data/code the primary studies,
8. Analyze and display the data using appropriate methods, and
9. Draw conclusions based on the data and discuss alternate interpretations in view of the study’s strengths and limitations.

Not only is there similarity in the steps between a primary study and a systematic review, both are expected to report the methods used so that the process is “transparent” (in the language of management) or “replicable” (in the language

of scientific inquiry). Thus, primary studies are to systematic reviews as individual human participants are to primary studies; and the one difference therefore is the primary study's investigators' obligation to assure informed consent and other procedures to protect human participants.

Some individuals newly introduced to current recommendations for systematic review excuse themselves from following and documenting their steps, "because I am not doing a meta-analysis." This is a misunderstanding of the transformation of standards for literature reviews. Although Glass used meta-analysis as a synonym for what we would now call a systematic review, the term has come to be reserved for systematic reviews that use specific statistical analyses (24, 63). Publication of an authoritative text for meta-analysis helped to define meta-analysis as a distinctive specialty within statistics (47). A survey of statistical publications that have defined meta-analysis is available elsewhere, but it is noteworthy that one example, over a century old, was from a public health application: In a 1904 publication, Karl Pearson reported having averaged estimates from five separate samples of the correlation between inoculation for typhoid fever and mortality to improve estimation of the typical effect of inoculation and to compare it with inoculation for other diseases (23, pp. 5–7).

SYSTEMATIC REVIEWS MAY IMPROVE THE QUALITY OF PRIMARY STUDIES

The influence of reviewer frustration over incomplete and misleading abstracts and study reports has led to guidelines increasingly adopted by journal editors to improve the completeness and quality of reporting by authors of primary studies. Indirectly, increased awareness of what will need to be reported when the study is completed may influence choice of study methods, e.g., blinding of data collectors, reporting denominators and participation rates. Readers of the primary studies benefit from increased clarity; reviewers' searches and coding can be more efficient and their classification of primary studies, more accurate. In addition, most of the several dozen quality scoring schemes, including that used by *The Community Guide*, are heavily weighted against incomplete reporting (54, 58, 69, 105, 107). These quality scoring systems also give priority to internal validity or the degree of confidence that should be placed on conclusions about causation. Although clearer reporting about the sample, setting, and time are helpful in determining the generalizability of the study conclusions, more emphasis on generalizability is still needed.

More Informative Abstracts and Study Reports

A maddening aspect of searching databases for study reports that meet a priori criteria is that many study abstracts do not allow a clear assessment of those criteria, for example criteria for type of intervention, study design, and outcome measures. The resources wasted retrieving study reports of unclear relevance gave

rise in 1987 to a proposal from the Ad Hoc Working Group on Critical Appraisal of the Medical Literature for changes in the format and content of abstracts to provide more information to readers and reviewers ("structured abstracts") (1). This proposal was followed by an update with a glossary of methodologic terms, and call for full adoption of not only the headings but also the specific content desired under each heading (44). The Instructions for Authors section of the *JAMA* website is a convenient place to find the full directions (51). At the time of this writing, only the headings are required, with few or no suggestions for content, in the *American Journal of Public Health* and the *American Journal of Preventive Medicine.* Neither the *American Journal of Epidemiology* nor *Health Education and Behavior* requires a structured abstract, however.

Even more resources are spent needlessly by coders charged with determining relevant characteristics of a study from a published report. Completing a coding form such as that used by the U.S. Community Preventive Services Task Force staff for intervention trials requires two to three hours to code a single study (18, 107). Again, medical trials reviewers encountered the same problem and created standards for what is included in various types of journal articles. Chief among them is CONSORT, *Con*solidated *S*tandards *o*f *R*eporting of Controlled Randomized *T*rials (10) and its update (71, 97), followed by a proposed extension to cluster randomized trials (35) and critiques of ethical and analytic issues in such trials (31, 77). A companion document with an explanation of each item on the CONSORT checklist, evidence of its importance, and examples of how to word the text is an excellent resource for editors, reviewers, and authors (5). Public health reviewers working in HIV/AIDS behavioral and social intervention research interpreted CONSORT for reporting non-RCTs—TREND: *T*ransparency in *R*eporting of *E*valuations from Trials with *N*on-randomized *D*esigns (30). A few months later, when the editors of *Annals of Behavioral Medicine* announced they had endorsed TREND, they noted that their journal was one of 153 medical, clinical, and psychological journals to have adopted the CONSORT guidelines (55).

Another group to join the effort to improve reporting encompasses researchers interested in behavioral services and interventions for community health. The evidence-based practice committee of the Society for Behavioral Medicine (SBM) adapted CONSORT for behavioral medicine (28, 56). Important additions focused on the intervention-content/elements (What was the content of the intervention and how was it delivered? e.g., oral communication), provider (Who delivered it?), format (What were the method(s) of administration? e.g., individual, group), setting (Where and when was it delivered?), recipient (e.g., an agent such as a parent versus the ultimate target of the intervention), intensity (number of contacts and total contact time), duration (total time period and spacing), and fidelity of actual delivery and method of monitoring (102a). The SBM authors endorsed the requirement by CONSORT that denominators be reported for recruitment and analysis but they added the requirement that authors include an explicit discussion of the generalizability of the findings. Further, in several articles and in a letter

regarding the TREND statement (31a), authors associated with a dissemination-oriented evaluation framework known as "RE-AIM" (*R*each, *E*fficacy, *A*doption, *I*mplementation, *M*aintenance) (36c) recommended that researchers, funding organizations, and reviewers change the usual sequence of efficacy to effectiveness studies by emphasizing earlier identification of moderating variables and explicit reporting and considerations of external validity (36a, 36b, 36d). Thus, it is expected that reporting requirements will undergo further shaping to respond to the complexities of behavioral and public health interventions.

Emphasis on Effect Magnitude (Effect Size) and Precision (Confidence Intervals)

The most elusive information in intervention study reports is information about the effect size or even the information needed to calculate an effect size and confidence interval. Systematic review in general, and meta-analysis in particular, have been shifting the emphasis in primary studies and in evidence tables in reviews away from statistical significance to effect sizes and confidence intervals. This benefits readers of primary studies as well as reviewers and readers of reviews. Two erroneous conclusions are frequently made based on *p*-values alone; (*a*) that $p \geq 0.05$ indicates no effect and (*b*) $p < 0.05$ indicates an effect of significant magnitude (86). But, of course, statistical significance is influenced by both the magnitude and strength of the relationship and the sample size. Therefore an "effect size" is needed to interpret whether the results from significance tests are meaningful to practice (20).

In writing about statistical analysis in the *American Journal of Public Health*, the editors advised, "Whenever feasible, authors should provide estimates of central tendency . . . along with appropriate indicators of measurement error or uncertainty, such as confidence intervals." Further, they ". . .favor the provision of data in forms that enable comparisons with other studies, for instance, statistics in a form suitable for combining in a meta-analysis . . ." (80). The notions of reporting comparable measures and of enabling one's study to be included in meta-analyses are increasingly found in authors' instructions. Another example of an influential group that has strengthened the call for reporting effect sizes comes from the fifth edition of the Publication Manual of the American Psychological Association, which states, "Neither of the two types of probability value directly reflects the magnitude of an effect or the strength of a relationship. For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship . . ." (6, p. 25). Guidance for authors of primary studies to achieve this aim is increasingly available (17, 27, 36, 41, 85, 99, 100). When study reports do not include effect size measures, various formulas can be used to estimate the effect size from other information provided (e.g., means, standard deviations, cell sizes) (36, 49, 57, 63), and more recently, Web-based or Excel-based calculators have been available from the Internet.

META-ANALYSIS INCREASES POWER TO DETECT AN EFFECT, ALTHOUGH AS YET IT DOES NOT FULLY USE AVAILABLE INFORMATION

Systematic review, and meta-analysis specifically, is generally recognized today as superior to the traditional literature where the review's conclusion was based on a "vote count." In a vote count approach, the reviewer would sort studies into categories—"negative and significant," "positive and significant," and "statistically nonsignificant"—and declare the modal category "the winner" (46). Reviewers have generally adopted a more conservative cut point, such as three fourths majority, but more than 25 years ago, statisticians Hedges & Olkin demonstrated that "the use of any such fraction is a poor practice" (46, p. 360) and that "surprisingly, the power of this procedure decreases as the number of studies increases" (46, p. 359). They suggested alternate methods, such as counting positive results not as a decision procedure ("The intervention is effective or not") but as an estimation procedure for which a confidence interval can be calculated. Meta-analysis has long been cited for increasing statistical power by reducing the standard error of the weighted average effect size (with the resulting effect of producing a narrower confidence interval increasing the likelihood of detecting nonzero population effects) (21). This increased statistical power affords a more efficient use of existing studies, and may provide sufficient evidence of an important population effect without having to conduct another study.

The cumulative meta-analysis offers efficient use of studies as they become available. This is a process whereby the systematic review/meta-analysis is updated with each additional study providing a "revised" estimate that is cumulative over time as new studies appear. This is an essential feature of The Cochrane and Campbell Collaborations whereby reviewers are required to keep their reviews current as new studies are identified. Far more important than keeping the review up to date is the notion that once the evidence sufficiently suggests an important population effect, then there is no longer a need to continue to fund similar studies; the example most often cited is one where had a cumulative meta-analysis been in progress, intravenous streptokinase would have been shown to be life saving almost 20 years before its submission to an approval by the U.S. Food and Drug Administration (74, 75).

Early users of meta-analysis alarmed statisticians and others by blithely including many effect sizes from a much smaller number of primary studies in estimating summary effects. Multiple treatment arms compared to a single control or comparison group, multiple measures as endpoints for each study participant, and multiple follow-up times for each measure can create numerous outcomes for a single study. Some reviewers recognized the problem of these correlated effects (72) not addressed in the early statistical texts, e.g., Hedges & Olkin (46, 47). The one case solved relatively early depended on sets of studies testing the same combinations of treatment arms—an ideal case with little or no application

in public health studies. Usual practice has been to select a single study arm (e.g., the one hypothesized to be the most effective), a single time for the follow-up measure (e.g., the modal time for the group of studies or the first follow-up), and a single measure for an outcome construct of interest or to pool the measures of the construct for the particular follow-up time and selected study arm. This left out a lot of potentially useful information contained in a study that had already been located, screened, and coded! By the 1990s some statistical (regression) solutions had been introduced, although these have limitations and have not made their way into everyday use (40).

SYSTEMATIC REVIEWS INSPIRE NEW RESEARCH QUESTIONS

Emphasis on the validity of review methods has given rise to a whole new literature of empiric studies on such questions as the sensitivity and specificity of various search techniques, the reliability and validity of coding procedures, including the assessment of study quality. Statistical methods of meta-analysis have also been advanced. The growth of this literature has been spurred by the availability of datasets of primary studies collected for meta-analysis and other systematic reviews, the newfound importance of knowing whether resources for particular tasks are justified, and the natural curiosity of reviewers and their colleagues from information sciences, measurement, statistics, and other disciplines. In addition, emphasis in meta-analysis on sensitivity analyses has helped to emphasize the comparison of review outcomes under alternative assumptions.

Empiric Work on Review Methods

Numerous questions have been addressed, an illustration of what is popularly known as "publication bias," referring to the publication (or nonpublication) of research findings, depending on the nature and direction of the results (33, 34). Investigators researching this within The Cochrane Collaboration have provided finer definitions in which publication bias is viewed as being a member of a group of biases labeled "reporting bias." These are characterized as follows: Statistically significant, "positive" results are (*a*) more likely to be published (publication bias), (*b*) more likely to be published rapidly (time lag bias), (*c*) more likely to be published in English (language bias), (*d*) more likely to be published more than once (multiple publication bias), and (*e*) more likely to be cited by others (citation bias) (19, 96). Concern for such bias in systematic reviews has led to more sophisticated search standards and attention to search techniques for the "fugitive literature," e.g., Rosenthal (82) and the development of statistical and visual tools for estimating bias (9, 11, 83). Calls for registration of drug and other clinical trials using clinical trials registries such as The Cochrane Central Registry of Controlled Trials, which began as a registry

of obstetrics trials (19), and other biomedical registries that have been developed, for example, http://www.clinicaltrials.gov/ by the National Library of Science and Federal Drug Administration (78). The International Committee of Medical Journal Editors now excludes unregistered trials from consideration for publication (29a). This should help provide a denominator of trials conducted on particular research topics and therefore help offset publication bias. Having main outcomes and study questions available when reviewing a study report would also reduce "fishing" and reporting of secondary outcomes as primary—a problem in several areas of behavioral research related to public health (for example, see 73).

Another topic pertinent to searches is language of the primary study. Thus, questions included, What is the reporting quality of primary studies reported in languages other than English compared to English-language publications? Answer: equally poor (67), and Do study outcomes of language restricted trials differ from language inclusive trials? Answer: no difference in some cases (65, 69), and in other cases, it depends on the topic of study (34). For example, authors of a meta-analysis of trials of an herbal treatment noted, "If we had restricted our literature search to English language publications, as often done in meta-analyses, we would not have identified a single trial in our initial search . . ." (61, p. 256). In public health applications restriction of language to English and setting to the United States would more often be justified because of the confounding of language with setting and population, and indeed, it is the case with *Community Guide* reviews. It would be hoped, however, that such questions might be approached empirically as well as conceptually.

Blinding of coders of primary studies to the journal and authors' identities has been suggested as a precaution against coder bias, largely on the basis of its tradition in clinical trials, and then based on results of experiments with blinded and unblended coders. Such an experiment reported in 1996 with medical trials indicated that blinded assessments produced significantly lower and more consistent scores on ratings of study quality indicators (50). Thus, *JAMA* instructions to authors of systematic reviews and meta-analyses require describing whether coding was blind or open (51).

Empiric Work on the Impact of Study Features

Studies such as those discussed here as examples of empiric research on the relative influence of study features on study outcomes are producing valuable information for the design of primary studies as well as advice for meta-analysis. To illustrate, much discussion among meta-analysts and other systematic reviewers has concerned the wisdom of adhering rigidly to randomized designs as the unique indicator of study quality. The medical trials community has been relatively rigid in this regard, whereas public health reviewers and their advisory groups have argued for more inclusive criteria for study design, such as categorizing nonrandomized trials with concurrent comparison groups and cluster designs together with individual and cluster randomized designs, in the highest quality design category (14).

For *The Community Guide* reviews, study design is further evaluated using a score for "quality of execution," a multi-indicator score representing several types of validity (22, 88) and with few design elements having been subjected to empiric verification. In the mid-1990s, a review of schema in use for scoring the quality of randomized clinical trials identified 25 scales and 9 checklists with diverse emphases, lack of empiric support for the scoring components, and general lack of psychometric assessment (68). Further, application of the various scoring schemes resulted in different conclusions (54). From this research emerged a second generation of scoring instruments, e.g., the Jadad score (50) now widely used for scoring medical RCTs based on the presence and conduct of random assignment, presence and conduct of double blinding, and whether withdrawals and dropouts were described. (An intention-to-treat analysis is not a scoring point, although it is often an additional study feature—intention-to-treat analysis, yes or no—in looking at the influence of study quality on outcomes.)

This attention to scoring has not been duplicated in public health applications, although it is hoped that wider discussion of methods of the *Community Guide* (14) and the HIV/AIDS Prevention Research Synthesis (92) may lead to such advances. In the meantime, the other approach has been to treat design features as individual variables to be investigated through stratified analyses or ideally, through meta-regression that allows focus on the often highly correlated study features most relevant to the particular set of studies. A stratified approach was used in the HIV/AIDS Prevention Research Synthesis (53) because of small numbers of studies on sex behavior risk reduction in some population strata [drug users, 33 studies (87); sexually experienced youth, 16 studies (73); heterosexual adults, 14 studies (79); and men who have sex with men, 9 studies (52)], for the overall study methods of interest (e.g., random versus nonrandom assignment in these controlled studies), as well as methods unique to the population stratum (e.g., cluster assignment with analysis at the individual level in the youth studies).

Elsewhere, in a secondary analysis of over 100 meta-analyses from psychology, behavioral interventions, and educational treatments, investigators found substantial variability in outcome across the primary studies within a typical meta-analysis (103). They estimated and compared the relative effects of treatment, respondent, measurement, and study design characteristics on this variance and found that the percent associated with substantive features of the intervention (treatment type, intensity/duration of treatment, respondent features, and the outcome construct) was roughly the same as the percent of variance resulting from method characteristics (design type, comparison group type, sample size, and operationalization of the outcome construct). Only about 50% of the variance in effect size was attributable to the former, and it must compete with bias associated with study method (21%) as well as noise from sampling error (26%). As the authors put it, "the signal is relatively small and the random and nonrandom noise is relatively large" (103, p. 424). They went on to interpret their findings to mean that a "single study will not typically provide a trustworthy indication of the effectiveness of a particular treatment" (103, p. 424) and a relatively underexamined

study design feature—operationalization of the outcome variable—is as important as study design (controlled studies, whether randomized or not, had comparable impact compared to single group studies).

Statistical Advances in Meta-Analysis

With funding and attention to the needs of reviewers tackling difficult (i.e., most) sets of studies with disparate outcomes, new statistical techniques are making the task easier. For example, a problem affecting many behavior change studies is identifying an effect measure that can be compared across studies and combined. In some HIV/AIDS prevention studies, outcomes were measured on a continuous scale, e.g., percent of time a condom is used, whereas in others, a dichotomous measure was used, e.g., whether a condom was used in the respondent's most recent episode of sex. In the former studies, the most natural effect size for the mean and standard deviations is a standardized mean difference, and in the latter studies, the natural effect size for a 2×2 contingency table is the odds ratio (72, 73). Although the rates in the treatment and control groups "depend strongly on the cutpoint used to dichotomize the outcome, the odds ratio is almost independent of the cutpoint" (45, 46). Statisticians working on meta-analyses of screening and diagnostic tests had developed a useful connection between the two metrics, making it possible to enlarge the group of studies with comparable metrics (43).

SYSTEMATIC REVIEW HOLDS PROMISE TO INFORM POLICY, PRACTICE, AND RESEARCH, ALTHOUGH PITFALLS AND CHALLENGES REMAIN

Varying Legitimacy of Claims of Being "Evidence-Based"

Even when it is not obligatory to do so, claiming to be "evidence based" conveys a measure of credibility nowadays that is valuable to have, observed two health care policy authors in a recent paper, "Evidence Based? Caveat Emptor!" (93, p. 80). Indeed, such claims for medicine, diagnostic tests, dentistry, mental health, nursing, physical therapy, alternative/complementary medicine, and numerous other health-care related maneuveurs as well as discussions of methods appear to make up most of the 54 million hits using the Google search engine; one day later the number had grown to 54.1 million. With the current emphasis on methods of systematic reviews and meta-analysis, their growing importance in health policy and medical decision making, and the burgeoning empiric literature on review methods described above, tools for assessing the quality of systematic reviews and meta-analyses also have appeared, including proposals for more informative abstracts of review articles (76), caution about quality in meta-analyses of RCTs and the tendency of lower-quality meta-analyses to produce positive findings (70), guidelines for reading review articles (81), and standards for reporting reviews—QUOROM or *Qu*ality *o*f *R*eporting *o*f *M*eta-analyses of RCTs (66).

Completeness of reporting and quality of the review, although important, are not the only basis for judging the strength of a body of evidence, however (93). A recent review of 40 approaches to rating the strength of an overall body of evidence found only 8 met criteria set by experienced investigators at an evidence-based practice center (64) (101). These criteria are a mixture of quality indicators for conducting a systematic review (whether the approach used to identify potentially pertinent studies was comprehensive and unbiased and whether bias was avoided in evaluating, synthesizing, and interpreting available evidence) and the quality (internal validity of each study), quantity (e.g., the number of studies or aggregate sample size), consistency (the extent to which similar findings are reported in studies with comparable designs, including outcome measures and interventions), and coherence of evidence (do the findings make sense as a whole?). Nevertheless, subjectivity is still a factor in the decision about the relevance and quality of individual studies—because various questions underlie a practice recommendation, with each a link in a chain of evidence; and "opinion often fills in gaps in the evidence base related to a chain of reasoning that underlies a clinical guideline" (93, p. 84).

Thus, the systematic reviews conducted for *The Community Guide* can be compared using these criteria. To a varying extent, all of the considerations listed above are taken into account in assembling data for a recommendation. To begin with, areas for study are broad and not based on the usual "silos" found in public health funding. Thus, *Guide* topic areas have spanned risk behaviors (tobacco, nutrition, physical activity, and sexual behavior), specific health conditions (diabetes, vaccine-preventable diseases, motor vehicle occupant injury, cancer, violence, and oral health), and the environment (socio-cultural environment). A diverse group of experts is assembled to develop a logic framework for the broad area that is designed to represent all the ways in which the problem might be addressed [see, for example, interventions in the social environment to improve community health (7, p. 115)]. All *Guide* reviews of interventions within a topic area also produce specific analytic frameworks, reflecting a priori decisions about the outcomes that will be considered and how they are related to health outcomes (if they are not themselves health outcomes), including important benefits and harms, for example, the analytic framework used to conduct the systematic reviews of tenant-based rental assistance programs (7, p. 123). These frameworks were adopted from those used by the U.S. Preventive Services Task Force as the principal organizational approach in developing recommendations for the *Guide to Clinical Preventive Services* (104). In reviews of both clinical and community interventions, the available studies arrayed along the links to display the chain of evidence typically have different studies and evidence of differing strength for the various links. It is here that more subjective judgments enter the process in making an overall recommendation, as described above (93).

The Community Guide reviews meet most best practices for systematic review, with the limitation of restricting included studies to published reports, but with unusual breadth in the sense of not restricting searches to the health and public

health literature. Because topics are inclusive of a broader view of public health, the review of the effects of tenant-based rental assistance programs (sufficient evidence of effectiveness) and mixed income housing (insufficient evidence of effectiveness), for example, included studies funded by the Department of Housing and Urban Development (7). *Guide* reviews use quantitative estimates of the effects, usually percentage-point improvement, but without confidence intervals or weighting study findings based on the precision of the point estimates. Thus, median percentage-point or percent change is the typical summary effect and without a confidence interval for the summary effect. Techniques for looking at consistency are also less sophisticated than meta-analysis. The tradeoff has been the ability to produce literally hundreds of reviews, despite serious deficiencies in reporting in the primary studies. The next generation of reviews will undoubtedly include more meta-analyses.

Additional considerations in translating a body of evidence to practice or policy are the applicability to diverse populations and settings, the balance between benefit and cost and benefit and harms (13, 14, 93, 104). In the case of *The Community Guide*, evidence for these considerations is frequently scarce. Most *Guide* recommendations are thought to apply to a broad range of populations and settings, but restricted samples and contexts among the primary studies cause some recommendations to be cast more narrowly, with further research recommended for other groups. Economic data are frequently missing from primary studies (13, 94), and postulated harms are understudied, with the result that some of these go on the research agenda and in other cases, recommendations are not directed to settings or populations where a plausible harm is more likely to occur (13).

Evidence-Based Public Health Policy and Practice

The promise of the evidence-based medicine and now evidence-based public health movements may not have the impact on practice and policy many believed they would or should—at least not initially. In the language of epidemiology, the generation of evidence (via primary studies and systematic reviews) may be necessary, but it is not sufficient to result in large-scale changes to practice and policy. For example, analysts commenting on the failure of adequate guideline uptake in medicine suggested that the traditional professional perspective of the clinician as sole decisionmaker has stood in the way of multifaceted implementation strategies that take the collaborative nature of medical work into consideration (97a).

Another response to the question, Why would evidence-based practice/policy resulting from an accumulation of evidence-based knowledge appear to have stalled? is that "researchers and policy makers operate in different contexts, motivated and constrained by different imperatives, differing world views and different priorities" (60, p.14). Health policy differs from evidence-based knowledge in that it is a set of competing rationalities (cultural, political and technical) with research (evidence) comprising only the technical rationality. As with systematic reviews, evidence-based public health policy and practice has considerable promise but

it will require hard work, conceptually and practically, if it is to realize its full potential.

The Agency for Healthcare Research and Quality's (AHRQ) evidence-based practice centers (2) were created to facilitate evidence-based health policy through their internal "review methods" that require "expert review panels" representing a range of expertise and interests (policy makers, consumers) in addition to the scientific/technical expertise. This approach, over time, could provide "nonscientists" a greater influence into the relevance factor of systematic reviews. The evidence report on chronic fatigue syndrome, for example, particularly in the decision "not to combine data," was influenced greatly by the members of the expert panels who represented consumer and practitioner stakeholders (102). Even the process for selecting evidence review topics for the AHRQ evidence-based centers is strongly influenced by those outside the evidence-based sphere–policy makers, practitioners and consumer advocacy groups. More inclusive input also has been suggested for *The Community Guide*, one possibility for which would be expansion of the membership on consultation teams (12a).

Other perspectives are based on experience with the complex behavioral, educational, and social interventions that play an important role in public health. Such interventions are far more context dependent than those examined in medicine, and they do not lend themselves readily to currently used methods of systematic reviews (12a, 86a). Even thoughtful critics considering transferability of principles of evidence-based medicine to the design of an online course for medical education suggested that evidence be expanded to include "informal knowledge, practical wisdom, and shared representations of practice" (40b, p. 142). Green, a program planning methodologist with a strong community focus, pointed out gaps between best practices from systematic reviews and local practice, suggesting alternatives to or variations on evidence-based practice, including reviews that address questions about "best *processes*" (e.g., 71a); primary studies that are more participatory and focused on adaptation, implementation, and maintenance; more encouragement and guidance for users of reviews to do more local evaluation and self-monitoring; more systematic study of place, setting, and culture; and more use of tailoring processes and new technologies (40a). Several of these points are consistent with the RE-AIM recommendations discussed above under reporting guidelines.

More fundamentally skeptical observations on the gap between systematic reviews and practice outside of medicine were made by Lipsey, who pointed out that initiatives such as *The Community Guide* assume that programs conducted as demonstration projects under conditions that facilitate the research and frequently evaluated by the program developer will provide the same benefits in practice (62). And, he noted that as yet, we have "little evidence about the effects of taking research-based programs to scale in public health and related areas of mental health, education, welfare, and criminal justice" (62, p. 3). Research evaluating the impact on public health of programs based on *Guide* recommendations is needed to resolve these concerns.

Clarifying the Research Agenda

From the beginning of serious discussion about systematic review, the method was seen as having a major role to play in clarifying the gaps in available studies to define the next important research questions (59). This would be expected, in view of the chief use of literature reviews in introductions to research papers and grant proposals. The difference expected to come from a systematic review, however, was not only that it would be based on a more comprehensive view of the relevant literature and be more objective, but also that it would result in fewer statements about the findings being "mixed"—as a conclusion. Instead, systematic review is better suited to suggesting or testing hypotheses about the sources of variability in study findings. Thus, a gap in research based on a systematic review would be more likely to be introduced with the phrase, "findings are mixed." Systematic review helps clarify the research agenda in several ways, including description of available studies; efforts to estimate the consistency, size, and precision of an intervention's effect; and analyses of the relative effects of alternative interventions and combinations of interventions and interactions with settings and population groups.

Presentation of characteristics of studies assembled for the review using clearly described and thoughtful search methods and eligibility criteria can sometimes overturn "common wisdom" about what has been studied. In the case of the HIV/AIDS Prevention Research Synthesis, the team expected that intervention studies of men who have sex with men (MSMs) would be far more plentiful than studies of populations affected later in the epidemic, e.g., heterosexual adults or youth. Compared with 14 studies with injecting drug users and 16 studies with youth, however, the 9 studies of MSMs was, in the reviewers' words, "striking," and they called for more rigorous evaluations of HIV prevention efforts with MSM (52).

A more explicit and helpful standard of comparison than "common wisdom" can be found in *The Community Guide* methods, using the logic and analytic frameworks. *Guide* reviews provide both specific observations about public health research. Importantly, they also have made general statements about biases, stating that high-quality public health intervention studies have favored clinical treatment and individually oriented approaches with simpler and shorter-term interventions (15). *The Community Guide* approach leads to a research recommendation when findings about an intervention from primary studies are not sufficient in number and quality, are not consistent, or of sufficient magnitude. In the case of interventions with sufficient magnitude, but which have been tested on narrow populations and particular settings, research is often recommended to answer the generalizability questions. And, where there are plausible harms but no evidence, this becomes another recommendation for the research agenda. Although *Guide* methods do not provide the precision and sophistication of meta-analysis (62), there is attention directed to these issues using acceptable tools. Research recommendations are thus relatively consistent and useful. *Guide* methods do pay considerable attention to the basis for differences in study findings, and suggest hypotheses for exploration in

subsequent research. Formal meta-analyses also provide findings from exploratory analyses that suggest directions for future research.

CONCLUSIONS

Systematic reviews and meta-analysis have come to public health, and public health research and practice are already beginning to reflect new standards, new knowledge, and new opportunity. Much that has been developed in medicine has helped public health move ahead relatively quickly, and the first generation of systematic reviews and recommendations for public health are more holistic with regard to the information attempted to be gathered and with more attention to the practice community in public health, including population-based medical care.

Reporting standards, standards for conducting reviews, and the beginning of bodies of evidence on review methods and indicators of study quality are becoming available in public health, although a shortage of capacity—so far based primarily at the Centers for Disease Control and Prevention—and training in the methods of systematic review and meta-analysis is lagging. Resources for conducting reviews are not yet available on a scale to attract public health scientists. Use of evidence-based recommendations from *The Community Guide* is beginning to be encouraged in some requests for proposals, and also, use of study methods that would fit the "best evidence subset of the *Community Guide.*" In areas of community preventive services in which the *Guide* has completed a review, respondents are asked to justify the intervention they plan to develop or test against the *Guide* findings.

More clearly reported primary studies will benefit practitioners directly and indirectly, through more completed systematic reviews and meta-analyses—the result of less time spent coding individual primary studies. This protection against intentional or unintentional obfuscation by the authors of study reports should clarify the study's conclusions and make it more likely that readers take away straightforward messages about the effectiveness, applicability, and cost-benefit ratio of particular interventions. Emphasis on effect sizes will begin to make the magnitude of study outcomes clearer and more easily compared. Further discussion of reporting standards by experts in behavioral sciences, public health, and dissemination is pushing the science further in the direction of more relevant evidence.

Claims of being "evidence-based" will continue to have variable legitimacy, but the quality of reviews, particularly from sources such as the CDC, should be improving. Application of standards for reviews will help to curb some of the weakest methods. In time, with more evidence and further use of meta-analysis, reviews should be able to test more specific interventions. Best practice or model programs "off the shelf" will continue to be needed by practitioners, even in the context of a systematic review (62) as, for example, the HIV Prevention Research Synthesis Project with meta-analyses and compendia of best practice (91). And, systematic intervention development techniques such as "intervention mapping"

(8) should address selecting and adapting *Guide* recommendations and results from other systematic reviews.

The *Annual Review of Public Health* is online at http://publhealth.annualreviews.org

LITERATURE CITED

1. Ad Hoc Work Group Crit. Apprais. Med. Lit. 1987. Academia and the clinic: a proposal for more informative abstracts of clinical articles. *Ann. Intern. Med.* 106:598–604
2. AHRQ. 2005. *Evidence-based practice centers*. http://www.ahrq.gov/clinic/epc/
3. AHRQ. 2005. *Guide to clinical preventive services*. http://www.ahrq.gov/clinic/cpsix.htm
4. AHRQ Clinical Practice Guidelines. 2005. *Agency for Healthcare Research and Quality, clinical practice guidelines online*. http://www.ahrq.gov/clinic/cpgonline.htm
5. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, et al. 2001. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann. Intern. Med*. 134:663–94
6. Am. Psychol. Assoc. 2001. *Publication Manual*. Washington, DC: Am. Psychol. Assoc. 5th ed.
7. Anderson L, Fullilove MT, Scrimshaw SC, Fielding JE, Normand J, et al. 2005. The social environment. In *The Guide to Community Preventive Services: What Works to Promote Health?* (*Task Force on Community Preventive Services*), ed. S Zaza, PA Briss, KW Harris, pp. 114–40. New York: Oxford Univ. Press
8. Bartholomew LK, Parcel GS, Kok G, Gottlieb NH. 2001. *Intervention Mapping: Designing Theory- and Evidence-based Health Promotion Programs*. Thousand Oaks, CA: Mayfield
9. Begg CB. 1994. Publication bias. See Ref. 24, pp. 399–409
10. Begg CB, Cho M, Eastwood S, Horton R, Moher D, et al. 1996. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 276:637–39
11. Begg CB, Mazumdar M. 1994. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 50:1088–101
12. Biostat. 2005. *Comprehensive Meta-Analysis*. http://www.meta-analysis.com/index_.html

12a. Briss PA. 2005. Evidence-based: US road and public-health side of the street. *Lancet* 365:828–30

13. Briss PA, Brownson RC, Fielding JE, Zaza S. 2004. Developing and using *The Guide to Community Preventive Services*: lessons learned about evidence-based public health. *Annu. Rev. Public Health* 25:281–302
14. Briss PA, Mullen PD, Hopkins DP. 2005. Methods used for reviewing evidence and linking evidence to recommendations. See Ref. 106, pp. 431–48
15. Briss PA, Portnoy B, Vogel-Taylor M, Zaza S. 2005. Continuing research needs. See Ref. 106, pp. 464–75
16. Campbell Collaboration. 2005. C2: The Campbell Collaboration. http://www.campbellcollaboration.org/
17. Capraro RM, Capraro MM. 2002. Treatments of effect sizes and statistical significance tests in textbooks. *Educ. Psychol. Meas.* 62:771–82
18. Cent. Dis. Control Prev., Task Force Commun. Prev. Serv. 2005. *Guide to Community Preventive Services:*

Systematic Reviews and Evidence Based Recommendations. http://www.thecommunityguide.org/

19. Cochrane Collaboration. 2005. *Cochrane Collaboration*. http://www.cochrane.org/index0.htm
20. Cohen J. 1992. A power primer. *Psychol. Bull.* 112:155–59
21. Cohn LD, Becker BJ. 2003. How meta-analysis increases statistical power. *Psychol. Methods* 8:243–53
22. Cook TD, Campbell DT. 1979. *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin
23. Cooper H, Hedges LV. 1994. Research synthesis as a scientific exercise. See Ref. 24, pp. 3–14
24. Cooper H, Hedges LV. 1994. *The Handbook of Research Synthesis*. New York: Russell Sage Found.
25. Cooper HM. 1982. Scientific guidelines for conducting integrative research reviews. *Rev. Educ. Res.* 52:291–302
26. Cooper HM. 1984. *The Integrative Research Review: A Systematic Approach*. Beverly Hills, CA: Sage
27. Cortina JM, Nouri H. 2000. *Effect Size for ANOVA Designs*. Thousand Oaks, CA: Sage
28. Davidson KW, Goldstein M, Kaplan RM, Kaufmann PG, Knatterud GL, et al. 2003. Evidence-based behavioral medicine: What is it and how do we achieve it? *Ann. Behav. Med.* 26:161–71
29. Davidson KW, Trudeau KJ, Ockene JK, Orleans CT, Kaplan RM. 2004. A primer on current evidence-based review systems and their implications for behavioral medicine. *Ann. Behav. Med.* 28: 226–38

29a. De Angelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, et al. 2005. Is this clinical trial registered?: a statement from the International Committee of Medical Journal Editors. *JAMA* 293:1927–29

30. DesJarlais DC, Lyles C, Crepaz N. 2004. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *Am. J. Public Health* 94:361–66
31. Donner A, Klar N. 2004. Pitfalls of and controversies in cluster randomization trials. *Am. J. Public Health* 94:416–22

31a. Dzewaltowski DA, Estabrooks PA, Klesges LM, Glasgow RE. 2004. TREND: an important step, but not enough. *Am. J. Public Health* 94:1474–75

32. Eddy DM, Hasselblad V, Shachter RD. 1992. *The Statistical Synthesis of Evidence: Meta-analysis by the Confidence Profile Method*. Boston, MA: Academic
33. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. 2003. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technol. Assess.* 7:1–76
34. Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G. 1997. Language bias in randomised controlled trials published in English and German. *Lancet* 350:326–29
35. Elbourne DR, Campbell MK. 2001. Extending the CONSORT statement to cluster randomized trials: for discussion. *Stat. Med.* 20:489–96
36. Fleiss JL. 1994. Measures of effect size for categorical data. See Ref. 24, pp. 245–60

36a. Glasgow RE, Bull SS, Gillette C, Klesges LM, Dzewaltowski DA. 2002. Behavior change intervention research in healthcare settings: a review of recent reports with emphasis on external validity. *Am. J. Prev. Med.* 23:62–69

36b. Glasgow RE, Klesges LM, Dzewaltowski DA, Bull SS, Estabrooks P. 2004. The future of health behavior change research: what is needed to improve translation of research into health promotion practice? *Ann. Behav. Med.* 27:3–12

36c. Glasgow RE, Vogt TM, Boles SM. 1999. Evaluating the public health impact of

health promotion interventions: the RE-AIM framework. *Am. J. Public Health* 89:1322–27

36d. Glasgow RE, Lichtenstein E, Marcus AC. 2003. Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. *Am. J. Public Health* 93:1261–77

37. Glass GV. 1976. Primary, secondary, and meta-analysis of research. *Educ. Res.* 5:3–8

38. Glass GV, McGraw B, Smith ML. 1981. *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage

39. Glass GV, Smith ML. 1978. Meta-analysis of research on the relationship of class size and achievement. *Educ. Eval. Policy Anal.* 1:2–16

40. Gleser LJ, Olkin I. 1994. Stochastically dependent effect sizes. See Ref. 24, pp. 339–55

40a. Green LW. 2001. From research to "best practices" in other settings and populations. *Am. J. Health Behav.* 25:165–78

40b. Greenhalgh T, Toon P, Russell J, Wong G, Plumb L, Macfarlane F. 2003. Transferability of principles of evidence based medicine to improve educational quality: systematic review and case study of an online course in primary health care. *BMJ* 326:142–45

41. Grissom RJ, Kim JJ. 2005. *Effect Size for Research: A Broad Practical Approach*. Mahwah, NJ: Erlbaum

42. Hasselblad V, Hedges LV. 1995. Meta-analysis of screening and diagnostic tests. *Psychol. Bull.* 117:167–78

43. Hasselblad V, Mosteller F, Littenberg B, Chalmers TC, Hunink MGM, et al. 1995. A survey of current problems in meta-analysis: discussion from the Agency for Health Care Policy and Research Inter-PORT Work Group on literature review/meta-analysis. *Med. Care* 33:202–20

44. Haynes RB, Mulrow CD, Huth EJ, Altman DG, Garewal H. 1990. More informative abstracts revisited. *Ann. Intern. Med.* 113:69–76

45. Hedges LV, Johnson WD, Semaan S, Sogolow E. 2002. Theoretical issues in the synthesis of HIV prevention research. *J. Acquir. Immune Defic. Syndr.* 30:S8–14

46. Hedges LV, Olkin I. 1980. Vote-counting methods in research synthesis. *Psychol. Bull.* 88:359–69

47. Hedges LV, Olkin I. 1985. *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic

48. Hunter JE, Schmidt FL. 1994. Correcting for sources of artificial variation across studies. See Ref. 24, pp. 323–36

49. Hunter JE, Schmidt FL. 1990. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Newbury Park, CA: Sage

50. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, et al. 1996. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control. Clin. Trials* 17:1–12

51. JAMA. 2005. *JAMA instructions for authors—current*. http://jama.ama-assn.org/ifora_current.dtl

52. Johnson WD, Hedges LV, Ramirez G, Semaan S, Norman LR, et al. 2002. HIV prevention research for men who have sex with men: a systematic review and meta-analysis. *J. Acquir. Immune Defic. Syndr.* 30:S118–29

53. Johnson WD, Semaan S, Hedges L, Ramirez G, Mullen PD, Sogolow E. 2002. A protocol for the analytical aspects of a systematic review of HIV prevention research. *J. Acquir. Immune Defic. Syndr.* 30:S62–72

54. Juni P, Witschi J, Bloch R, Egger M. 1999. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 282:1054–60

55. Kaplan RM, Christensen AJ. 2004. *Annals of Behavioral Medicine* endorses transparent reporting of evaluations with nonrandomized designs. *Ann. Behav. Med.* 27:147

56. Kaplan RM, Trudeau KJ, Davidson KW. 2004. New policy on reports of randomized clinical trials (editorial). *Ann. Behav. Med.* 81
57. Kelsey KS, Kirkley BG, DeVellis RF, Earp JA, Ammerman AS, et al. 1996. Social support as a predictor of dietary change in a low-income population. *Health Educ. Res.* 11:383–95
58. Khan K, Daya S, Jadad AR. 1996. The importance of quality of primary studies in producing unbiased systematic reviews. *Arch. Intern. Med.* 156:661–66
59. Light RJ, Pillemer DB. 1984. *Summing Up: The Science of Reviewing Research*. Cambridge, MA: Harvard Univ. Press
60. Lin V. 2003. Competing rationalities: evidence-based health policy? In *Evidence-based Health Policy: Problems and Possibilities*, ed. V Lin, B Gibson, pp. 3–17. New York: Oxford Univ. Press
61. Linde K, Ramirez G, Mulrow CD, Pauls A, Weidenhammer W, Melchart D. 1996. St John's Wort for depression—an overview and meta-analysis of randomised clinical trials. *BMJ* 313:253–58
62. Lipsey MW. 2005. The challenges of interpreting research for use by practitioners: comments on the latest products from the Task Force on Community Preventive Services. *Am. J. Prev. Med.* 28:1–3
63. Lipsey MW, Wilson DB. 2001. *Practical Meta-Analysis*. Thousand Oaks, CA: Sage
64. Lohr KN. 2004. Rating the strength of scientific evidence: relevance for quality improvement programs. *Int. J. Qual. Health Care* 16:9–18
65. Moher D, Pham B, Klassen TP, Schulz KF, Berlin JA, et al. 2000. What contributions do languages other than English make on the results of meta-analyses? *J. Clin. Epidemiol.* 53:964–72
66. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. 1999. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of reporting of meta-analyses. *Lancet* 354:1896–900
67. Moher D, Fortin P, Jadad AR, Juni P, Klassen T, et al. 1996. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. *Lancet* 347:363–66
68. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. 1995. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control. Clin. Trials* 16:62–73
69. Deleted in proof
70. Moher D, Olkin I. 1995. Meta-analysis of randomized controlled trials: a concern for standards [comment]. *JAMA* 274:1962–64
71. Moher D, Schulz KF, Altman D. 2001. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 285:1987–91
71a. Mullen PD, Green LW, Persinger GS. 1985. Clinical trials of patient education for chronic conditions: a comparative meta-analysis of intervention types. *Prev. Med.* 14:753–81
72. Mullen PD, Ramirez G. 1987. Information synthesis and meta-analysis. In *Advances in Health Education and Promotion*, ed. W Ward, MH Becker, PD Mullen, S Simonds, pp. 201–39. Greenwich, CT: JAI
73. Mullen PD, Ramirez G, Strouse D, Hedges LV, Sogolow E. 2002. Meta-analysis of the effects of behavioral HIV prevention interventions on the sexual risk behavior of sexually experienced adolescents in controlled studies in the United States. *J. Acquir. Immune. Defic. Syndr.* 30:S94–105
74. Mulrow CD. 1987. The medical review article—state of the science. *Ann. Intern. Med.* 106:485–88
75. Mulrow CD. 1994. Systematic reviews:

rationale for systematic reviews. *BMJ* 309:597–99

76. Mulrow CD, Thacker SB, Pugh JA. 1988. A proposal for more informative abstracts for review articles. *Ann. Intern. Med.* 108:613–15
77. Murray DM, Varnell SP, Blitstein JL. 2004. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am. J. Public Health* 94:423–32
78. Natl. Libr. Sci. 2005. *ClinicalTrials.gov by the National Library of Science and Federal Drug Administration*. http://www.clinicaltrials.gov/
79. Neumann MS, Johnson WD, Semaan S, Flores SA, Peersman G, et al. 2002. Review and meta-analysis of HIV prevention intervention research for heterosexual adult populations in the United States. *J. Acquir. Immune Defic. Syndr.* 30:S106–17
80. Northridge ME, Levin B, Feinleib M, Susser MW. 1997. Statistics in the journal—significance, confidence, and all that. *Am. J. Public Health* 87:1092–95
81. Oxman AD, Guyatt GH. 1988. Guidelines for reading literature reviews. *Can. Med. Assoc. J.* 138:697–703
82. Rosenthal MC. 1994. The fugitive literature. See Ref. 24, pp. 85–94
83. Rosenthal R. 1979. The file drawer problem and tolerance for null results. *Psychol. Bull.* 86:638–41
84. Rosenthal R. 1994. Parametric measures of effect size. See Ref. 24, pp. 231–44
85. Rosenthal R, Rosnow RL, Rubin DB. 2000. *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*. Cambridge, UK/New York: Cambridge Univ. Press
86. Rosnow RL, Rosenthal R. 2003. Effect sizes for experimenting psychologists. *Can. J. Exp. Psychol.* 57:221–37

86a. Rychetnik L, Frommer M, Hawe P, Shiell A. 2002. Criteria for evaluating evidence on public health interventions. *J. Epidemiol. Commun. Health* 56:119–27

87. Semaan S, Des Jarlais DC, Sogolow E, Johnson WD, Hedges LV, et al. 2002. A meta-analysis of the effect of HIV prevention interventions on the sex behaviors of drug users in the United States. *J. Acquir. Immune Defic. Syndr.* 30:S73–93
88. Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin
89. Shadish WR, Haddock CK. 1994. Combining estimates of effect size. See Ref. 24, pp. 261–81
90. Smith ML, Glass GV. 1977. Meta-analysis of psychotherapy outcome studies. *Am. Psychol.* 32:752–60
91. Sogolow E, Kay LS, Doll LS, Neumann M, Mezoff JS, et al. 2000. Strengthening HIV prevention: application of a research-to-practice framework. *AIDS Educ. Prev.* 12:21–32
92. Sogolow E, Peersman G, Semaan S, Strouse D, Lyles CM, The HIV/AID PRS Project Team. 2002. The HIV/AIDS Prevention Research Synthesis Project: scope, methods and study classification results. *J. Acquir. Immune Defic. Syndr.* 30:S15–29
93. Steinberg EP, Luce BR. 2005. Evidence based? Caveat emptor! *Health Aff.* 24:80–92
94. Stone GA, Hutchinson AB, Corso PS, Teutsch SB, Fielding JE, Carande-Kulis VG. 2005. Understanding and using the economic evidence. See Ref. 106, pp. 449–63
95. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, et al. 2000. Meta-analysis of observational studies in epidemiology: a proposal for reporting. *JAMA* 283:2008–12
96. The Cochrane Collab. 2005. *What is publication bias?* http://www.cochrane-net.org/openlearning/html/mod15-2.htm

97. The CONSORT Group. 2005. *CONSORT: strength in science, sound ethics*. http://www.consort-statement.org/
97a. Timmermans S, Mauck A. 2005. The promises and pitfalls of evidence-based medicine. *Health Aff.* (*Millwood*) 24:18–28
98. Tobacco Use Depend. Clin. Pract. Guidel. Panel Staff Consort. Represent. 2000. A clinical practice guideline for treating tobacco use and dependence: a U.S. Public Health Service report. *JAMA* 283:3244–54
99. Trusty J, Thompson B, Petrocelli JV. 2004. Practical guide for reporting effect size in quantitative research in the *Journal of Counseling & Development*. *J. Couns. Dev.* 82:107–10
100. Vacha-Haase T, Thompson B. 2004. How to estimate and interpret various effect sizes. *J. Couns. Psychol.* 51:473–81
101. West S, King V, Carey TS, Lohr KN, McKoy N, et al. 2002. Systems to rate the strength of scientific evidence. *Evid. Rep. Technol. Assess.* 1–11
102. Whiting P, Bagnall AM, Sowden AJ, Cornell JE, Mulrow CD, Ramirez G. 2001. Interventions for the treatment and management of chronic fatigue syndrome: a systematic review. *JAMA* 286:1360–68
102a. Whitlock EP, Orleans CT, Pender N, Allan J. 2002. Evaluating primary care behavioral counseling interventions: an evidence-based approach. *Am. J. Prev. Med.* 22:267–84
103. Wilson DB, Lipsey MW. 2001. The role of method in treatment effectiveness research: evidence from meta-analysis. *Psychol. Methods* 6:413–29
104. Woolf SH. 1994. An organized analytic framework for practice guideline development: using the analytic logic as a guide for reviewing evidence, developing recommendations, and explaining the rationale. In *Clinical Practice Guideline Development, Methodology Perspectives*, ed. KA McCormick, SR Moore, RA Siegel, pp. 105–13. Washington, DC: AHCPR Publ. No. 95–0009
105. Wortman PM. 1994. Judging study quality. See Ref. 24, pp. 97–109
106. Zaza S, Briss PA, Harris KW. 2005. *The Guide to Community Preventive Services: What Works to Promote Health?* (*Task Force on Community Preventive Services*). New York: Oxford Univ. Press
107. Zaza S, Wright-De Aguero LK, Briss PA, Truman BI, Hopkins DP, et al. 2000. Data collection instrument and procedure for systematic reviews in *The Guide to Community Preventive Services*. *Am. J. Prev. Med.* 18:44–74

Annual Review of Public Health
Volume 27, 2006

CONTENTS

EPIDEMIOLOGY AND BIOSTATISTICS

Effective Recruitment and Retention of Minority Research Participants, *Antronette K. Yancey, Alexander N. Ortega, and Shiriki K. Kumanyika* 1

Measuring Population Health: A Review of Indicators, *Vera Etches, John Frank, Erica Di Ruggiero, and Doug Manuel* 29

On Time Series Analysis of Public Health and Biomedical Data, *Scott L. Zeger, Rafael Irizarry, and Roger D. Peng* 57

The Promise and Pitfalls of Systematic Reviews, *Patricia Dolan Mullen and Gilbert Ramírez* 81

Hypertension: Trends in Prevalence, Incidence, and Control, *Ihab Hajjar, Jane Morley Kotchen, and Theodore A. Kotchen* 465

ENVIRONMENTAL AND OCCUPATIONAL HEALTH

Environmental Justice: Human Health and Environmental Inequalities, *Robert J. Brulle and David N. Pellow* 103

Speed, Road Injury, and Public Health, *Elihu D. Richter, Tamar Berman, Lee Friedman, and Gerald Ben-David* 125

The Big Bang? An Eventful Year in Workers' Compensation, *Tee L. Guidotti* 153

Shaping the Context of Health: A Review of Environmental and Policy Approaches in the Prevention of Chronic Diseases, *Ross C. Brownson, Debra Haire-Joshu, and Douglas A. Luke* 341

PUBLIC HEALTH PRACTICE

Health Disparities and Health Equity: Concepts and Measurement, *Paula Braveman* 167

The Politics of Public Health Policy, *Thomas R. Oliver* 195

Vaccine Shortages: History, Impact, and Prospects for the Future, *Alan R. Hinman, Walter A. Orenstein, Jeanne M. Santoli, Lance E. Rodewald, and Stephen L. Cochi* 235

What Works, and What Remains to Be Done, in HIV Prevention in the United States, *David R. Holtgrave and James W. Curran* 261

SOCIAL ENVIRONMENT AND BEHAVIOR

A Public Health Success: Understanding Policy Changes Related to Teen Sexual Activity and Pregnancy, *Claire D. Brindis* 277

An Ecological Approach to Creating Active Living Communities, *James F. Sallis, Robert B. Cervero, William Ascher, Karla A. Henderson, M. Katherine Kraft, and Jacqueline Kerr* 297

Process Evaluation for Community Participation, *Frances Dunn Butterfoss* 323

Shaping the Context of Health: A Review of Environmental and Policy Approaches in the Prevention of Chronic Diseases, *Ross C. Brownson, Debra Haire-Joshu, and Douglas A. Luke* 341

Stress, Fatigue, Health, and Risk of Road Traffic Accidents Among Professional Drivers: The Contribution of Physical Inactivity, *Adrian H. Taylor and Lisa Dorn* 371

The Role of Media Violence in Violent Behavior, *L. Rowell Huesmann and Laramie D. Taylor* 393

HEALTH SERVICES

Aid to People with Disabilities: Medicaid's Growing Role, *Alicia L. Carbaugh, Risa Elias, and Diane Rowland* 417

For-Profit Conversion of Blue Cross Plans: Public Benefit or Public Harm? *Mark A. Hall and Christopher J. Conover* 443

Hypertension: Trends in Prevalence, Incidence, and Control, *Ihab Hajjar, Jane Morley Kotchen, and Theodore A. Kotchen* 465

Preventive Care for Children in the United States: Quality and Barriers, *Paul J. Chung, Tim C. Lee, Janina L. Morrison, and Mark A. Schuster* 491

Public Reporting of Provider Performance: Can Its Impact Be Made Greater? *David L. Robinowitz and R. Adams Dudley* 517

Health Disparities and Health Equity: Concepts and Measurement, *Paula Braveman* 167

INDEXES

Subject Index 537

Cumulative Index of Contributing Authors, Volumes 18–27 565

Cumulative Index of Chapter Titles, Volumes 18–27 570

ERRATA

An online log of corrections to *Annual Review of Public Health* chapters may be found at http://publhealth.annualreviews.org/