REVIEW ARTICLE

Meta-analysis: Methods, strengths, weaknesses, and political uses

JOHN H. NOBLE, JR.

WASHINGTON, D.C.

The general methodology, strengths and weaknesses, and political uses of meta-analysis are examined. As a systematic study of all studies that have been conducted to answer a specific question or hypothesis, meta-analysis is strong in revealing structural flaws and sources of bias in primary research and in posing promising research questions for future study. It cannot exceed, however, the limits of what is reported by primary researchers. Meta-analysis is particularly challenged to quantify the size of a common effect of treatment across reported trials because of (1) the clinical diversity of the trials and (2) the myriad of potential differences among patients with varying characteristics within the trials. Without access to the original data of reported trials, meta-analysis cannot overcome the bias of underpowered trials toward overstatement of the size of main treatment effects, nor the tendency for such trials to falsely conclude there were no statistically significant adverse events. Although severely compromised by ghost-written or honorary-authored reports of primary research, meta-analysis can make use of its methods to focus on the conflicts of interest and likely sources of bias of such research and make known what precautions should be taken by would-be consumers. Examples show how meta-analysis has clarified thinking about the off-label use of selective serotonin reuptake inhibitors for treating child and adolescent depression, use of low-tidal volume respirator assistance for acute lung injury and acute respiratory distress syndrome patients, and the long-term use of COX-2 inhibitors for relieving arthritic pain. Recommendations are made for Congressional action. (J Lab Clin Med 2006;147:7–20)

Abbreviations: ALI = acute lung injury; ANOVA = analysis of variance; ARDS = acute respiratory distress syndrome; FDA = U.S. Food and Drug Administration; IQ = intelligence quotient; NDA = new drug application; NSAID = nonsteroidal anti-inflammatory drug; RCT = randomized controlled trial; SSRI = selective serotonin reuptake inhibitor

Meta-analysis refers to the statistical analysis of a collection of individual studies to summarize what is known from empirical research in answer to a specific question or hypothesis. "Meta" from the Greek for "after" refers to the analysis of at least two primary datasets after each one has been initially analyzed and published or otherwise communicated. As an analysis of analyses, meta-analysis is to be distinguished from reanalysis of primary data either to confirm original findings or to answer new questions. Interest in synthesizing research results dates back to the work in the 1930s of L. H.C. Tippett,[1] R. A. Fisher,[2] Karl Pearson,[3] and W.G. Cochran,[4] all of whom sought to combine the results of different agricultural studies.[5] Meta-analysis was reinvented in 1974 by the psychologist, Gene Glass,[6] to refute the conclusion of an eminent colleague, H. J. Eysenck, that psychotherapy was essentially ineffective. Independently, the physician

From the National Catholic School for Social Service, The Catholic University of America, Washington, D.C.

Submitted for publication March 25, 2005; revision submitted August 1, 2005; accepted for publication August 2, 2005.

Reprint requests: John H. Noble, Jr., PhD, 508 Rio Grande Loop, Georgetown, TX 78628; e-mail: jhnoble@verizon.net.

0022-2143/$ – see front matter

doi:10.1016/j.lab.2005.08.006

and biomedical researcher, Thomas Chalmers,[7] soon after devised a strategy for combining clinical trials to summarize findings and published the first meta-analysis in medicine, although he did not recognize the term when informed of winning the 1982 annual research award of the Evaluation Research Society for his accomplishment.[6]

With the exception of psychiatry, the more sharply focused questions of medicine have tended to provoke less controversy than the fuzzy, more complex questions of the social sciences as, for example, what factors influence school achievement? There is reason for caution when assessing the published claims of meta-analysis in both the biomedical and the social sciences. The strengths and weaknesses of meta-analysis in biomedical and social science research are becoming better known through the systematic efforts of the Cochrane and the Campbell collaborations to standardize review methodology and to disseminate the results of best practice.[8,9] There are many sources of guidelines for conducting a meta-analysis,[10–16] as well as a growing literature of critical commentary[17–22]—especially in the *British Medical Journal*.[23]

This article presents the general methodology of meta-analysis, notes salient differences in approach and use between the biomedical research and the social sciences, assesses strengths and weaknesses, discusses the political uses of the genre as a guide to improving health-care policy and clinical practice, and provides a portal into the essential literature of meta-analysis. The recent controversy about use of COX-2 inhibitors to control arthritic pain illustrates how powerful a role the meta-analysis of only two primary studies can play in helping to resolve contentious interpretations of available research.[24]

GENERAL METHODOLOGY OF META-ANALYSIS

Meta-analysis can be conceived as a systematic study of all studies that have been conducted to answer a specific question or hypothesis. It investigates not only the reported results of the studies but all aspects of research designs that produced them, including theoretical constructs, operational definitions of the independent or manipulated variable, moderating and mediating variables, and dependent variables, population samples, data collection procedures, statistical analysis, and especially the handling of possible confounding variables that would provide an alternative explanation for the reported results. Like primary research, meta-analysis proceeds by framing a research question to be addressed by sampling a defined population of completed primary studies to be surveyed, coded for relevant methodological characteristics, and analyzed to test hypotheses derived from the research question.

In the context of research about the outcomes of interventions to preserve or enhance physical, psychological, or social functioning, meta-analysis addresses two principal questions: (1) Is there support in the sampled population of studies for the causal inference that the intervention made a statistically significant difference in the outcome(s)? And if so, (2) how large an effect or difference did the intervention make? The substantive importance or significance of the size of the intervention effect is a separate matter that invokes value judgments based on criteria that are outside the scope of meta-analysis per se; for example, "Is the statistically significant difference also substantively large enough to merit adoption generally as a clinical application in medicine or as public policy applicable to all members of society within some defined cost constraint?"

Just as there is a sequenced set of activities to guide the conduct of primary research, the meta-analytic research review proceeds through successive stages, namely, (1) problem formulation; (2) data collection, ie, selection of the relevant studies; (3) evaluation of the collected data; (4) analysis and interpretation; and (5) presentation of results. Cooper[25] conceptualizes meta-analysis as a research project wherein at each stage there is a research question addressed, a primary function served by the review process, and awareness of how procedural differences at each stage, such as the criteria for including and excluding primary studies, can create variation as well as potential invalidity in review conclusions. The sources of invalidity in the conduct of a meta-analysis highlight the potential pitfalls for the would-be meta-analyst and explain why well-conducted meta-analysis is so difficult and expensive (and, indeed, why the key to the validity of meta-analysis lies in the choice of method).[25,26]

Examples of how some of the more important sources of invalidity in Cooper's stages of meta-analysis can lead to biased or misleading conclusions follow. At the problem formulation stage of meta-analysis involving specification of causally related variables, abstracting the content of primary studies that provide too little detail about how latent constructs are operationalized can mask the influence of interacting moderating and/or mediating variables that serve to explain cause–effect relationships. At the data collection stage, accessing primary studies that contain population samples that are so different because of exclusions from the intended target population that generalized application of the results to the target population is invalid. The definition of the target population can be so vague that it makes sample selection arbitrary, or the sample selected even with a clear definition does not represent the defined population.

At the data evaluation stage, missing relevant information in the primary study reports, such as the incidence of adverse events, can lead to unreliable and invalid conclusions. At the analysis and interpretation stage of meta-analysis, differences in the rules for inference in the primary studies can lead the meta-analyst to infer causality where none exists, eg, weighting equally experimental, quasi-experimental, and case studies or weighting equally studies with different size samples or potentially biasing differential attrition rates across subgroups of the study population. Last, at the publication phase of meta-analysis, omission of important review details, such as the criteria for including and excluding studies, any ad hoc decisions, or procedures adopted while conducting the analysis, can prevent replication of results or lead to erroneous conclusions about the true scope of the review.

Ideally, raw data showing separate comparisons provide the optimal basis for combining studies, and according to Cooper,[10] "the level of analysis to which the reviewer should aspire . . . before other less adequate means for combining results are undertaken." Unfortunately, the ideal is seldom attained because primary studies seldom, if ever, report raw data, and efforts to retrieve raw data from the authors usually end in failure,[10] as is also the case with most primary biomedical research reports.[27] Consequently, meta-analysis typically proceeds by converting either (1) the probability values of statistical significance for each relevant study comparison or (2) the relevant effect sizes into a common metric, both of which permit direct comparison and summation of the independent studies.

Of the several available methods for converting and combining the probability values of statistical significance for independent studies, the easiest and most frequently used is Stouffer's Adding Zs formula in either unweighted or weighted form.[28] Weighted, the formula is:

$$Z_w = \frac{\sum_{i=1}^{N} WiZi}{\sqrt{\sum_{i=1}^{N} W_i^2}}$$

where Z_w = the standard normal deviate, or Z-score, for the weighted combination of study comparisons; Z_i = the standard normal deviate for the ith comparison; W_i = the weighting factor for each study; and N = the total number of comparisons in the series.

Primary studies may be weighted to reflect sample size, quality of research design (eg, experimental vs quasi-experimental) or other factors (eg, attrition) that may influence their reliability and validity.

The second principal question to which meta-analysis is addressed relates to how large a difference the intervention made, ie, the size of its effect? In Cohen's[29] terms, "effect size" refers to "the degree to which the phenomenon is present in the population," or "the degree to which the null hypothesis is false." Two scale-free, common-metric measures of size effect used in the social sciences are (1) the d-index, ie, the distance between two group means expressed in terms of their common standard deviation; and (2) the r-index or Pearson product-moment correlation coefficient. Biomedical researchers more commonly use the odds-ratio statistic, ie, the odds of the outcome in the treated or exposed group divided by the odds of the outcome in the control group, to express size effect. For example, the odds ratio of an outcome showing at least 50% pain relief from Ibuprofen 400 mg in a hypothetical randomized trial of 80 patients would be 5.7, where 22 in 40 treated with Ibuprofen (0.55 experimental event rate) experienced relief compared with 18 who did not, and where 7 in 40 in a placebo control group (0.18 control event rate) experienced similar relief compared with 33 who did not. This translates into experimental event odds of 1.2 (22/18) versus control event odds of 0.21 (7/33) and a corresponding odds ratio of 5.7 (1.2/0.21). Biomedical researchers sometimes employ the relative-risk ratio, ie, the experimental event rate divided by the control event rate, in place of the odds ratio to measure size effect. For the experimental and control event rates in the hypothetical randomized trial, the relative risk would be 3.1 (0.55/0.18).[30] Meta-analysts use a variety of formulas to convert the reported statistical findings of primary studies that are based on different statistical models into a common size effect measure with confidence intervals.[31]

Statistically significant nonzero size effects are the basis for determining the influence of the independent variable by itself and, data permitting, in interaction with a host of possible moderating variables, such as gender, age, socioeconomic status, education, health history, and a variety of mediating variables, such as components of a complex intervention that correlate with the intervention and occur after the intervention begins.[32] Although successful randomization in the typical clinical trial may control for the effects of moderating variables in estimating the overall effect of the experimental variable, knowledge of interaction effects is important when deciding how individual patients should be treated or which patient groups will most benefit. The interaction of moderating variables with the treatment defines the patient groups and identifies their differential response. With respect to the interaction of mediating variables with treatment, the usual assumption of a common effect across trials is implausible because the mediating variables are highly unlikely to be implemented identically from trial to trial, especially in behavioral interventions. Thus, the

clinical diversity of the trials themselves (mediators) and the myriad of potential differences among patients with varying characteristics within the trials (moderators) pose stiff challenges for the would-be meta-analyst.[27,32]

Multiple operationalism, ie, acceptance and use of different "manifest" or observable measures of the same theoretical or latent construct that captures different facets of a particular "latent" or unobservable construct, characterizes much of meta-analysis in the social and behavioral sciences. This entails, for example, combining and summarizing the results of studies of IQ (latent construct) using several of its different manifest measures, such as the Stanford-Binet Intelligence Scale, the Wechsler Adult Intelligence Scale, the California Test of Mental Maturity, and the Leiter International Performance Scale. Whether multiple operationalism should be considered a methodological strength or weakness has been a contentious matter from the beginning. Indeed, Eysenck reacted harshly to the original Glass and Smith meta-analysis that called into question his own published conclusions about the effects of psychotherapy by calling it "an exercise in mega-silliness" and "garbage in, garbage out" when studies of higher and lesser quality are combined.[6]

The expanded use of meta-analysis in biomedical and social research attests to the embrace of the Glass and Smith position. Indeed, some hold that meta-analysis reduces the uncertainty associated with specific primary studies that employ different manifest measures of the same latent construct, arguing that "If a proposition can survive the onslaught of a series of imperfect measures, with all their irrelevant error, confidence should be placed in it."[33] However, the critically important proviso in this regard is that the measures encompassed within the meta-analysis capture similar facets of a common latent construct and be of at least satisfactory reliability and validity. Cooper[10] argues that ". . .if the majority of operations bear no correspondence to the underlying concept or the operations share a different concept to a greater degree than they share the intended one, the conclusion of the review will be invalid regardless of how many items or operations are involved."

Concept-to-measurement correspondence may be less of a problem in biomedical research, wherein measured outcomes are "harder" (eg, body weight, blood pressure) and do not involve as much idiosyncratic judgment as social science measures (eg, attitudes, self-appraisal). Meta-analysis in the social sciences in contrast to biomedical research typically combines a much larger number of primary studies containing diverse operational definitions of latent constructs that are susceptible to greater measurement error by virtue of the wider possible range of interpretation or judgment among both study subjects and the investigators. The original Glass and Smith meta-analysis reviewed 375 published and unpublished studies that survived initial screening of the 1000 found in the psychotherapy outcome literature at that time.[6] After screening for quality, the typical social science encompasses 50 to 500 primary studies, eg, 93 in the Yu and Cooper[34] review of research design effects on questionnaire response rates, 44 in the Ambady and Rosenthal[35] meta-analysis of thin slices of expressive behavior as predictors of interpersonal consequences, 345 in the Rosenthal and Rubin[36] review of interpersonal expectancy effects, and 443 in the Lipsey[37] meta-analysis of the effectiveness of treatment for juvenile delinquents. One meta-analysis[38] of published meta-analyses relating to the efficacy of psychological, educational, and behavioral treatment reviewed 156 such studies that provided treatment effect estimates based on control or comparison group designs.

Meta-analyses in biomedical research typically encompass smaller numbers of primary studies but contain ostensibly more reliably measured dependent variables, eg, death rates during specified time periods, as well as a better articulated statement of the nature of the independent or experimental variable. Halvorsen et al[39] reviewed the frequency of cited meta-analyses and secondary data analyses in the first 10 issues of four weekly general medical journals in 1982, including the *New England Journal of Medicine*, the *Journal of the American Medical Association*, the *British Medical Journal*, and the *Lancet*. The 20 cited meta-analyses, published from 1960 to 1983, combined and reviewed from 5 to 473 studies each; 26.7% reviewing less than 10 studies; another 26.7% from 10 to 30 studies; 10% from 31 to 103 studies, and one that reviewed 473 studies. In contrast, the Lipsey and Wilson[38] meta-analysis of 302 meta-analyses relating to the efficacy of psychological, educational, and behavioral treatment reveals that the typical social science meta-analysis is based on a much larger number of primary studies; 4.6% reviewing less than 10 studies; 31.5% from 10 to 30; 21.2% from 31 to 50; 24.8% from 51 to 150; 8.9% from 151 to 200; 2.6% from 201 to 250; and 1% reviewing more than 250 primary studies. Chalmers and Lau[40] cite 495 meta-analyses of almost all randomized clinical trials published as of June 12, 1992 but with no mention of the number of studies that each one combined.

Published meta-analyses vary in scope and critical acumen. The Cochran Collaboration[8] reviews provide uniform information about meta-analysis background, objectives, search strategy, study selection criteria, data collection and analysis procedures, main results, and

Table I. Claimed strengths of meta-analysis

Strength	Source
1. Can summarize from available studies the effects of interventions across many patients.	Thompson and Higgins[27]
2. Can reveal research designs as moderators of study results.	Cooper[25]
3. Can determine if the effect of the intervention is sufficiently large in practical as well as statistical terms.	Lipsey and Wilson[38]
4. Can, through multi-operationalism, reduce uncertainty of interpretation—if the measures encompassed are sufficiently valid.	Webb. Campbell, Schwartz, Sechrest, and Grove[33]
5. Can allow more objective assessment of evidence and thereby reduce disagreement.	Egger and Smith[20]
6. Can reduce false negative results and thereby hasten introduction of effective treatments into clinical practice.	Egger and Smith[20]
7. Can allow testing *a priori* hypotheses about treatment effects in patient subgroups.	Egger and Smith[20]
8. Can clarify heterogeneity between study results.	Egger and Smith[20]
9. Can suggest promising research questions for future study.	Egger and Smith[20]
10. Can assist accurate calculation of sample size needed in future studies.	Egger and Smith[20]
11. Can increase precision of literature reviews.	Cooper and Hedges[51]
12. Can, through consistent coding of primary study characteristics and use of multiple judges, reduce bias in judgements about the "quality" of individual studies.	Cooper and Hedges[51]
13. Can, through various statistical formulae, provide confidence interval calculation of effect size estimates.	Cooper and Hedges[51]
14. Can, using different assumptions or alternative statistical models, clarify and interpret the range of possible conclusions about the "quality" of segments of the literature review.	Cooper and Hedges[51]
15. Can overcome problems of traditional literature reviews involving (a) selective inclusion of studies, (b) subjective weighting of studies and their interpretation, (c) failure to examine study characteristics as source of disparate or consistent results across studies, and (d) failure to address influence of moderating variables in the relationship being examined.	Wolf[5]
16. Can, through systematic use of threats-to-inference framework, reveal structural flaws and sources of bias in research procedures.	Campbell and Stanley[52]; Noble[53]

reviewer conclusions. Compared with meta-analyses published in paper-based journals, Cochran reviews exhibit several superior features, including description of trial inclusion and exclusion criteria, assessments of research quality, no restriction on the language of publication, periodic updating, and most importantly, less likelihood of bias.[11,41,42] Sampling of available Cochran reviews on recent controversial topics substantiates reliance by biomedical meta-analysts on mostly small numbers of combined studies, eg, 5 RCTs for ventilation with lower versus traditional tidal volumes for ALI and ARDS,[43] 98 RCTs for treatment of depression with SSRIs versus other antidepressants,[44] 5 RCTs for treatment of rheumatoid arthritis with celecoxib,[45] 4 crossover studies testing treatment of rheumatoid arthritis with paracetamol versus nonsteroidal anti-inflammatory drugs,[46] and 2 RCTs for treatment of rheumatoid arthritis with rofecoxib.[47]

Other recently published meta-analyses bearing on these controversial topics combine similarly small numbers of primary studies. One by Greenberg et al[48] reviewed all 13 computer-search locatable double-blind, placebo-controlled fluoxetine trials with sample sizes ranging from 42 to 540. Another by Eichacker et al.[49] reviewed five randomized, prospective trials of the use of low tidal volumes (5–7 mL/kg measured body weight) as a protective lung strategy in the mechanical ventilation of ALI and ARDS patients with sample sizes ranging from 52 to 861 patients. The meta-analysis of the cardiovascular risk of the COX-2 inhibitor, valdecoxib (Bextra; Pfizer, Inc., New York, NY) by Furberg et al.[24] combined two trials, one published and the other unpublished, with unspecified sample sizes.

In short, differences in the reliability and validity of measurement, ie, the "signal to noise" ratio, in addition to the relative cost of conducting the primary research, may well explain the disparate number of studies that are typically encompassed in meta-analysis of social science as compared with biomedical research. Reliability of measurement is a necessary but not sufficient condition for validity. Power analysis can reveal what size sample is needed to detect an effect of a given size between experimental and control subjects, but whether it can be detected is largely determined by the reliability and validity of measurement.[50]

STRENGTHS AND WEAKNESSES

There is reasonable consensus about the methodological strengths and weaknesses of meta-analysis. Table I[51–53] and Table II list the major claimed strengths and admitted weaknesses. Meta-analysis, if carefully constructed and implemented, can assist biomedical and

Table II. Admitted weakness of meta-analysis

Weakness	Source
1. Can pass along inflated estimates of size effects based on (a) research design characteristics, (b) published vs. unpublished study composition, and (c) small sample sizes.	Lipsey and Wilson[38]
2. Can be limited by unspecified "black box" treatment and control conditions as well as lack of attention to potential interactions with subject characteristics, range of outcomes, and temporal factors.	Lipsey and Wilson[38]
3. Can be compromised by inclusion of non-peer-reviewed data—especially if derived from biased sources, eg, the pharmaceutical industry.	Smith and Egger[21]
4. Application of the results of meta-analysis to individual patients remains a difficult judgment call because "uncertainty with respect to a particular patient will always be greater than with respect to the overall patient group."	Smith and Egger[21]
5. Interpretation of different attrition rates as indicative of therapeutic success or failure is problematic because of judgment subjectivity.	Smith and Egger[21]
6. Cannot overcome subjectivity in choice of outcomes and their weighting in analysis.	Smith and Egger[21]
7. The combined statistically significant estimate of size effect may still prove inconclusive because of ignorance or uncertainty about other relevant matters.	Smith and Egger[21]
8. The correlational nature of review-generated evidence precludes conclusive inference about the strength of possible confounds or rival explanations for the reported effects of primary studies.	Cooper and Hedges[51]
9. The post hoc nature of research synthesis prevents use of review-generated evidence to develop and test theory simultaneously.	Cooper and Hedges[51]
10. Partially complete information in some studies compromises coding certainty and reliability as well as the extent to which studies with complete information can represent the universe of all relevant studies.	Cooper and Hedges[51]
11. Reader judgments about the "quality" of a specific research synthesis are dependent on such subjective criteria as (a) self-assessed ability to interpret the review, (b) organization, (c) writing style, (d) clarity of focus, (e) use of citations, (f) attention to variable definitions, (g) attention to details of methodology, and (h) manuscript layout.	Cooper and Hedges[51]
12. Cannot eliminate without registration of all trials publication bias wherein positive findings get published and negative ones do not.	Moncrief[12]
13. Synthesis may mask relevant heterogeneity with respect to distinguishing characteristics of participants, circumstances of interventions, and the conduct of the primary research.	Moncrief[12]
14. Arithmetic nature of meta-analysis can produce false impression of certainty in an inherently uncertain process with many subjective elements.	Moncrief[12]
15. Statistics for calculating "heterogeneity" in primary study size effects are weak and lack power to distinguish true differences among trials from chance effects; hence, statistical nonsignificance does not mean there is sufficient homogeneity to justify their combination.	Moncrief[12]; Thompson and Higgins[27]
16. Meta-regression, in which treatment benefit is related, where possible, to some average characteristic of patients in each trial, eg, mean age or proportion of women, is fraught with interpretive difficulty and is generally misleading.	Thompson and Higgins[27]
17. Meta-regression describes observed relationships across trials that—even if all randomized—are still subject to confounding by other variables that may vary between trials and thus invalidate inferences about relationship cause and effect.	Thompson and Higgins[27]

social scientists to determine the extent to which accumulated evidence tends to provisionally confirm or conclusively refute a specified theory about a set of phenomena under investigation. In this view, pursuant to Popper's falsifiability principle, based on the quality of the primary studies that enter into a meta-analysis as well as the quality of the meta-analysis, we can only eliminate demonstrably false theories about putative cause–effect relationships and choose among any remaining unfalsified theories as the basis for belief and action.[54]

The claimed strengths of meta-analysis are all contingent on the important proviso that the measures that represent and share the same theoretical concepts within a meta-analysis be of at least satisfactory validity and thus permit "triangulation" of evidence. Use of original raw data for combining the results of primary studies is the ideal but seldom achieved approach to meta-analysis. It minimizes the considerable uncertainty that surrounds the extraction and coding of relevant details about the primary studies, including basic research design, operational definitions, population samples, hypothesis testing, overall and subgroup attrition, and the like. The claimed strength (#7) that meta-analysis can allow testing *a priori* hypotheses about treatment effects in patient subgroups may be overstated in view of the admitted weaknesses of meta-regression (#16 and #17) in isolating and interpreting such treatment effects.

Departure from the ideal of using the raw data of combined primary studies largely contributes to the weaknesses of meta-analysis. The first of these admit-

ted weaknesses, combining separate estimates of size effects, is particularly perilous in face of the confounding influences of underlying (1) research design characteristics (including the measurement problems of the ALI and ARDS trials discussed below), (2) published versus unpublished composition of the primary studies, and (3) sample size. Weakness #2, lack of clarity about the nature of the treatment and control conditions, including their erroneous specification, is particularly worrisome in social science research but can be a threat to inference in biomedical research. The confounding effects of the operational definition of the independent variable in one experiment[55] testing the effects of high versus low tidal volumes respirator setting for ALI and ARDS patients might have explained the inconsistency found in the Eichacker et al.[49] meta-analysis of five such trials.

As indicated by Noble,[56] using data contained in Eichacker et al.,[49] the distribution of abrupt changes in tidal volume that resulted from random assignment to the high- and low-volume experimental conditions of the ARDS Network trial[55] from the original tidal volume prescribed by the original primary physician is a rival explanation for the resulting reported differences in mortality. Further suggested as perhaps more appropriate to testing the null hypothesis was the use of a random- or mixed-effect factorial experimental design instead of the fixed-effect model that was actually employed to account for time-dependencies of measures. In any event, the Eichacker et al.[49] meta-analysis demonstrates the strengths of the genre in posing promising research questions for future study (#9) and in revealing structural flaws and sources of bias in research procedures by use of the threats-to-inference analytic framework (#16). Reanalysis of the original data eventually supported the hypothesis that abrupt change to static experimental ventilator tidal volumes higher or lower than those prescribed by the patient's primary physician accounted for higher mortality.[57]

Noteworthy among the listed weaknesses of meta-analysis is that it is incapable of correcting the limitations imposed by underpowered primary studies that introduce small sample bias with attendant overstatement of estimated effect size as well as insensitivity to the influence of relevant moderators and mediators on trial outcomes, including the detection of clinically important adverse events that can occur as a function of moderators (patient characteristics) and mediators (study procedures or conditions). The stakes in this regard are vastly different in the social sciences versus biomedical research. Whereas in the social sciences the emphasis is typically the relationship between two variables and not their interactions with a third, there is most always in biomedical research with clinical applications a tradeoff between benefits and risks. The tradeoff involves judgments about measured benefits, the potency of any concomitant adverse events, and their interactions with subgroups in the target clinical population.

Thus, although interaction effects between the main effect and any third variables are important, receiving even "inferential priority" in Cooper's[10] view, in meta-analysis of social science primary studies where life-safety is not at issue, so much more so do they deserve priority in biomedical research. The fact that interaction effects are so seldom (or incompletely) reported in social science primary research is an annoyance and drag on the advancement of knowledge. In biomedical research, long-standing ignorance about the interaction effects between the main effect and the clinically relevant third variables, eg, history of high blood pressure or heart disease in patients receiving a COX-2 inhibitor for arthritic pain, can well spell the difference between life and death for many thousands of patients and, when revealed, cause resentful backlash and reactions from worried patients, angry physicians, and aroused patient advocates, as well as the defensive posturing of financially threatened drug companies and exposed government regulators.[58–64]

Underpowered primary research, that is, research that employs samples of insufficient size and insufficient variability (restriction in range) relative to the anticipated results of measurement, simultaneously constrains detection of small statistically significant differences between experimental and control subjects in clinical trials and increases the likelihood of accepting the null hypothesis as true when it is false (type II error). Small samples increase sampling error in measuring both main and interaction effects and increase the difficulty for meta-analysis to detect the sources of variance in the results of several primary studies that might be explained by differences in (1) how the studies were conducted or (2) who participated in them.

This matter is the subject of an ongoing debate about the tradeoffs and costs involved. Chalmers and Lau[40] argue the case for encouraging numerous small clinical trials that through cumulative meta-analysis will speed the discovery of more effective treatments for disease and illness on the assumption that a series of small trials will be sufficiently similar in their design and procedures to produce a corpus of studies that can be aggregated to test a common hypothesis. Halpern et al.[65] oppose and argue that conducting an underpowered clinical trial that cannot answer the research question is unethical in all but two exceptional situations wherein there is provision for disclosing to potential subjects the limitations of the research in which they are asked to enroll, namely, (1) small trials for rare diseases with an

Table III. *A priori* analysis* of trade-off between sample size and power for ANOVA F-test on means with Type I error equal to 0.05 and Type II error (I-Power of the Test) equal to 0.05

No. cells	Sample size		Power of test when sample size is for a medium effect and the effect is actually small
	Medium effect	Small effect	
$3 \times 1 = 3$	252	1548	<0.45
$3 \times 2 = 6$	324	1986	<0.45
$3 \times 2 \times 2 = 12$	420	2532	<0.45

*Erdfelder, Faul, Buchner[66]

explicit plan for combining the results of several similar trials in a prospective meta-analysis and (2) early phase trials for drugs or devices wherein the results may only contribute indirectly to future improved health care.

To illustrate the dilemma of biomedical researchers and the meta-analysts who depend on them, consider the tradeoff between sample size and the power of the test to detect small- and medium-size differences in a three-arm noninferiority and/or superiority clinical trial comparing any new drug both to placebo and to some one relevant standard treatment. As shown by Table 3,[66] the sample size needed to balance type I and type II error in a one-way, fixed-effect ANOVA to test the null hypothesis that there are no statistically significant small-size differences among the three groups is six times larger than for testing the null hypothesis for medium-size differences. Responding to the FDA initiative[67] to encourage measurement of gender effects in clinical trials, the sample size required to detect a medium-size difference in a two-way, fixed-effect ANOVA would be roughly 25% larger than that of the initial three-arm trial. The same six-to-one ratio holds for increasing sample size to detect small-size as compared with medium-size differences in the two-way, fixed-effect ANOVA. Adding one more moderator would require an increase in sample size of two thirds over that of the initial three-arm trial to detect medium-size differences in a three-way, fixed-effect ANOVA and, again, a six-fold increase in sample size to detect small-size effects.

The conduct of large clinical trials is both financially and logistically expensive. This and the argument that use of larger size samples exposes more patients than absolutely necessary to possible harm provide a rationale for conducting small-size clinical trials. The consequences in terms of type II error (1—power of the test) are spelled out in the last column of Table 3. Small clinical trials that can detect statistically significant medium-size effects of a new drug or device will conclude falsely more than 55% of the time that there were no statistically significant moderator effects or between-group differences in adverse events if their true effect size is small. The Chalmers and Lau[40] belief that meta-analysis of numerous small clinical trials will come to the rescue seems ill-founded even if the primary researchers publish the small-sample results of statistical tests of possible moderator effects or between-group differences in adverse events—unless they are large enough to permit detection as statistically significant pursuant to the decision rule for type I error (falsely rejecting the null hypothesis when it is true). Only combining and analyzing the raw data from numerous small trials capable of detecting medium-size effects offers the possibility of accumulating over time sufficient evidence to detect small-size effects within perhaps as much as a six-fold increased total-size sample. The practice may also serve to identify and refute much earlier the claims of smaller randomized trials whose findings tend to fail replication by subsequent studies.[68]

Last, the naiveté or discomforted unwillingness of meta-analysts, as the case may be, to identify and confront the bias introduced by unpublished, unreported, and/or misrepresented findings in primary research probably accounts for why this problem is not raised in any meta-analytic studies reviewed here. The problem is real as indicated by escalating attention to corporate-sponsored clinical trials of investigational drugs and allegations that published and unpublished reports of primary research have been manipulated or suppressed by interested parties to advance the sale of unsafe products.[69–74] The New York Attorney General recently sued GlaxoSmithKline for alleged "repeated and persistent fraud by misrepresentation, concealing and otherwise failing to disclose to physicians information in its control concerning the safety and effectiveness of its antidepressant medication paroxetin (Paxil)" in treating child and adolescent depression.[75] Suffice it to say, it is a worrisome public health concern when unsuspecting physicians are dosed with flawed or even fraudulent research to influence their judgment about the benefits and risks of specific drugs or devices in treating individual patients.

POLITICAL USES

The claims of meta-analysis and its interpreters in the political context of deciding health-care policy or medical practice rules need careful examination. That primary research can be summarized as the basis of evidence-based medical practice is the central assumption, indeed, doctrine, that motivates Cochrane Collaborative contributors.[8] After all, meta-analysis, over and above any threats to inference contained in the methods of the underlying primary studies, must confront such threats at each stage of its own process.[25] Despite numerous obstacles, most notably those that stem from fraudulent research reports, meta-analysis has an important role to play in the search for internally and externally valid knowledge to guide development of scientifically sound public policy. Indeed, meta-analysis has a complementary role to play in countering some perverse incentives that the Congress inadvertently created in the effort to speed up the review and approval process of the FDA.[76]

In this regard, it is encouraging that published meta-analytic reports have been partially responsible for the questioning of claims about the effectiveness and safety of the COX-2 inhibitors,[24] low-tidal-volume assistive respirator treatment of ALI and ARDS patients,[49] and off-label use of SSRIs for treating child and adolescent depression[77–80] as well as depression in adults.[81] Meta-analysis can often overcome the inferential biases of a single study conducted and published by investigators with potential conflicts of interest by seeking out the unpublished studies that very often find no statistically significant differences between the intervention and the comparison conditions. Reactive questioning of the methods and conclusions of the meta-analytic reports contributes to better understanding of the issues and whatever additional data may be needed to further reduce the uncertainty of knowledge in controversial matters.[82]

Succeeding paragraphs identify how the Lau et al[83] ideal of cumulative meta-analysis and software[22] to accomplish it are undermined by identifiable perverse incentives and behaviors of the current FDA review and approval system. These commerce-engendered perversions are regarded by Lemmens[71] as analogous to Kafka's leopards in the temple that have become part of the ceremony of science while actually making a mockery of scientific integrity wherein "research data are shared and scrutinized, and uncontrolled self-interested behavior should be banned." Clues are also provided on how those with a taste for blood sport can use the methods of meta-analysis as weapons to hunt the invading leopards.

Several authors[84–86] have revealed how ghost-written and honorary-authored reports are fairly commonplace in some leading medical journals (as many as 30% by one estimate).[86] Also common are the subtle and not-so-subtle influences of financial conflicts of interest on reported research findings.[87–89] Clearly, meta-analysis that does not check the listed authors for authenticity and financial conflicts of interest overlooks those sources as a plausible rival explanation for what is reported. The meta-analyst is responsible for checking and controlling for research quality and sources of bias. The "threats-to-inference" approach[25] to classifying primary research is particularly useful in this regard, especially when combined with the conduct of sensitivity analysis to explore the extent to which specific sources of invalidity affect reported research findings. The exercise sometimes can reveal a flaw that undercuts the thrust of an entire body of apparently positive research findings, eg, the efficacy of assertive community treatment for people with serious mental illness.[90]

The Cochrane Collaborative systematic reviews for use in evidence-based medical practice might well incorporate the requirement that reviewers check and control for author authenticity and financial conflicts, perhaps going so far as to suggest that such reports when combined with others in meta-analysis be either thrown out or discounted by means of a suitably weighted sensitivity analysis. Indeed, the most conservative approach for combining studies is to employ the laboratory or researcher as the smallest unit of analysis.[25] Contemporary information technology permits tagging of ethically compromised authors and research institutions for use in evaluating subsequent publications. In this regard, the meta-analyst can use available reports of compromised International Committee of Medical Journal Editors (ICMJE), accountability standards, access to data, and control of publication.[91]

Most troubling are relaxed FDA standards and pressures to approve NDAs under the Prescription Drug User Fee Act[92] and FDA Modernization Act.[93] The previous "gold standard" of two adequate and well-controlled clinical trials showing effectiveness and safety is now discretionary with one such study deemed adequate if accompanied with confirmatory evidence.[76] McCabe[76] cites a survey of FDA reviewers revealing that pressure had been brought on them by their superiors to approve drugs that should never have been approved or had been approved too quickly, with one reviewer protesting that the burden of proof had shifted from proving safety to disproving dangerousness. Putting aside small sample bias with attendant overstatement of estimated effect size, insensitivity to detection of clinically important adverse events, the ease with which statistical findings can be manipulated in any

single clinical trial, and the FDA bias to approve NDA applications, the meta-analyst cannot penetrate the "trade secret" status of information that was provided by industry to obtain approval, including the results of all trials before the one that showed statistical significance, and thus it cannot properly combine all results to determine their overall statistical significance.

As Taveggia[94] points out, " . . . in and off themselves, the findings of any single research are meaningless—they may have occurred by chance." All data points, not some, are needed to evaluate the truth of claimed findings. There is no way for the meta-analyst to work around the obstacle of partial disclosure or concealment in published and unpublished studies. Meta-analyses that purport to summarize all relevant primary research on a given topic, including those of the Cochrane Collaboration, are potentially contaminated. Hammerschmidt and Franklin[95] amply describe the tensions that surround the publication decision of a medical journal when confronted by so-called evidentiary asymmetry or bias in a manuscript that trims content to protect a trade secret. Faced with the ever-present possibility of partial disclosure or concealment, the meta-analyst has no recourse but to rely on occasional whistle-blowers and the sleuthing of investigative reporters for cautionary tidbits to add to what can be gleaned from the usual channels of information.[96,97]

The extent of the foregoing problems are revealed by recent disclosures about the increased cardiovascular risks of COX-2 inhibitors that have been prescribed for millions of Americans to avert real or anticipated gastrointestinal symptoms of NSAID use to control arthritic pain.[58–64] The increased cardiovascular risks were not identified by earlier Cochran Collaborative systematic reviews[45–47] of the effectiveness and safety of the COX-2 inhibitors on the basis of available primary research reports. It was the results of three randomized, placebo-controlled trials of COX-2 inhibitors for different conditions, two to prevent colorectal cancer[58,59] and one to control postoperative pain after cardiac surgery[60], that gave indication of the increased cardiovascular risks. Soon after these reports came the results of a nested case-control study explicitly designed to test whether the COX-2 inhibitors compared with NSAID use increased cardiovascular risks.[98]

Psaty and Furberg[61] describe how rofecoxib was approved by the FDA in 1999 despite signals of safety problems in the available small, short-term trials; how a later larger trial inadequately reported a five-times greater rate myocardial infarction for rofecoxib compared with naproxen; and what is still not known about "the exact levels of risk for each drug, the time course of the risk during therapy, and the populations of patients, if any, in whom the benefits might exceed the known risks." The trials that led to approval were defective because they were too small, made inadequate provision for measuring cardiovascular events, and excluded the very high-risk patients for whom the drugs were later prescribed. Looking back on what went wrong, Psaty and Furberg[61] recommend that drugs applicable to long-time use by millions of people should be vetted by large, long-term clinical trials from the outset of the approval process. Brief trial periods will not reveal lagged adverse effects that only emerge months to years after starting treatment of chronically ill patients with a single or combination of drugs.

Had they been available, could meta-analysis of the data from the original small, short-term trials that led to the approval of refecoxib have revealed the increased risk for cardiovascular events? The answer is "no" because meta-analysis depends on what is reported and cannot overcome what is not reported in published studies or withheld altogether by suppressing publication. On the other hand, had there been interest and available resources, direct analysis of the FDA postmarketing adverse drug reaction database might have revealed elevated cardiovascular events among COX-2 inhibitor versus NSAID users even though only 3% to 10% of the actual number of drug reactions are reported annually.[99] Furthermore, had the FDA and the biomedical research community been proactive in taking a threats-to-inference approach in meta-analyzing the cumulative results of all clinical trials involving the use of COX-2 inhibitors, the structural flaws now identified by Psaty and Furberg[61] just might have influenced drug-labeling in terms of what is not known about the benefit–risk ratios for each drug in terms of the time course of therapy.

At a minimum, such meta-analysis might have used whatever flawed data were available on adverse reactions to evaluate the likelihood of type II error (erroneously accepting the null hypothesis) versus type I error (erroneously rejecting the null hypothesis) in preliminary findings of the lack of statistically significant differences between experimental and control subjects. Such a meta-analysis might well have picked up on the fact that the primary research had excluded the high-risk patients for whom the drugs were later prescribed, raised questions about the practice, and pointed the FDA and the biomedical research community in the direction of reviewing the FDA postmarketing adverse drug reaction database for what might be revealed through that source despite its severe limitations.

Why were none of these steps taken? The growing consensus points to a compromised FDA regulatory function that has become unbalanced in its dual role of protecting public health and industry wealth and that has permitted too many industry leopards to roam un-

checked in the temple of science.[71,76] The media have picked up on this and are in the process of educating the public with the facts and the consequences even as the FDA is taking steps to stifle criticisms of its failures in this regard.[99,100] It remains to be seen whether and how Congress responds to the growing unease about the adequacy of the American drug-safety system. Will it push the FDA toward the conduct of RCTs to determine the true risk associated with treatments, as called for by Drazen,[62] or perhaps require that all RCTs be registered and their results be disseminated as quickly as possible, as called for by Psaty and Furberg?[61]

For meta-analysis to make its contribution, however, the Congress will have to pass legislation that either bans or severely limits "trade secret" status to the results of clinical trials in support of new drug or device applications for FDA approval. Such intervention would be consistent with a recent report of the U.K. House of Commons Health Committee[101] that indicts "traditional secrecy in the drug regulatory process" as the underpinning of "publication bias and other unacceptable practices," which in conjunction with the "closeness" between regulators and pharmaceutical companies, "has deprived the industry of rigorous quality control and audit." Clearly, full access to the raw data needs to be part of any clinical trial registration scheme to facilitate fully efficient meta-analysis or systematic reviews of the primary research on which the evidence-based practice of medicine depends.

Although Congress cannot mandate study design, it can in its oversight capacity require and fund systematic meta-analytic reviews of clinical trials to determine the threats to inference they contain as well as how well they have been designed to (1) approximate the population to whom treatment will be delivered, (2) run long enough to match trial results with the course of the illness and treatment regimen, (3) use comparable measures of outcome and consistent procedures for adjudication of treatment response, (4) handle the so-called placebo effect by use of a prebaseline time-series measurement to neutralize the effects of regression-to-the-mean, (5) sample populations to represent variability in illness severity, and (6) avoid underpowered analysis of the external validity or generalizability of investigational agents across the entire range of severity.

The emerging facts argue for strong corrective action in the United States and in all countries in which biomedical research is conducted subject to government regulation. Fixing the problems in the United States only will not suffice, although doing so would likely have a very large impact throughout the world. Sightings of Kafka's leopards in the temple of science have been reported in the United Kingdom,[101] Canada,[102,103] Australia,[104] Eastern Europe,[105] and India.[106] Close linkage of U.S. PhRMA with the Office of the U.S. Trade Representative is considered a threat to access to affordable medicines and equitable health care throughout the world by inserting preferential property provisions in bilateral and multilateral trade agreements.[104] Protection of frivolous patent claims through these agreements promotes the practice of "evergreening" of brand-name drugs to the detriment of generic-drug development and serves to distort biomedical research priorities in the interest of monopolistic market share[76] while encouraging biomedical researchers to behave badly.[70]

I wish to thank H. Stephen Leff, PhD and staff of the Evaluation Center@HSRI Human Services Research Institute for help in identifying the essential meta-analysis literature as well as Mark Wilson, PhD, and Paul B. Gold, PhD, for their suggested revisions of the original manuscript.

REFERENCES

1. Tippett LHC. The methods of statistics. London: Willams and Norgate; 1931.
2. Fisher RA. Statistical methods for research workers. 4th ed. London: Oliver and Boyd; 1932.
3. Pearson K. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. Biometrika 1933;25:379–410.
4. Cochran WG. Problems arising in the analysis of a series of similar experiments. J Royal Stat Soc 1937;100(Suppl 4):102–18.
5. Wolf FM. Meta-analysis: quantitative methods for research synthesis. Beverley Hills, Calif.: Sage; 1986:13.
6. Hunt M. How science takes stock: the story of meta-analysis. New York: Russell Sage Foundation; 1997.
7. Chalmers TC, Matta RJ, Smith H, Kunzler AM. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. N Engl J Med 1977;297:1091–6.
8. Available at: http://www.cochrane.org/index0.htm. Accessed February 25, 2005.
9. Available at: http://www.campbellcollaboration.org/guidelines.htm. Accessed February 25, 2005.
10. Cooper H. Scientific guidelines for conducting integrative research reviews. Rev Educ Res 1982;52:291–302.
11. Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M, et al. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. JAMA 1998;280:278–80.
12. Moncrieff J. Research synthesis: systematic reviews and meta-analysis. Int Rev Psychiatr 1998;10:304–11.
13. Robey RR, Dalebout SD. A tutorial on conducting meta-analyses of clinical outcome research. JSLHR 1998;41:1227–41.
14. Rosenthal R. Writing meta-analytic reviews. Psychol Bull 1995;118:183–92.
15. Rothstein HR, McDaniel MA. Guideline for conducting and reporting meta-analyses. Psychol Rep 1989;65:759–70.
16. Duriak JA, Lipsey MW. A practitioner's guide to meta-analysis. Am J Commun Psychol 1991;19:291–352.

17. Furkawa T. Meta-analyses and megatrials: neither is the infallible, universal standard. EBMH Notebook. Available at: www.ebmentalhealth.com. Accessed February 25, 2005.
18. Howard GS, Maxwell SE, Fleming KJ. The proof of the pudding: an illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. Psychol Meth 2000;5: 315–32.
19. Moynihan R. Evaluating health services: a reporter covers the science of research synthesis. New York: Milbank Memorial Fund; 2004.
20. Egger M, Smith GD. Meta-analysis: potentials and promise. BMJ 1997;315:1371–74.
21. Smith GD, Egger M. Meta-analysis: unresolved issues and future developments. BMJ 1998;316:221–5.
22. Egger M, Stern JAC, Smith GD. Meta-analysis software. BMJ 1998;316:221–5.
23. Available at: http://bmj.bmjjournals.com/cgi/collection/systematic_reviews%3Astatistics_descriptions. Accessed February 25, 2005.
24. Furberg CD, Psaty BM, Fitzgerald GA. Parecoxib, valdecoxib, and cardiovascular risk. Circ J Am Heart Assoc 2005;111:249.
25. Cooper H. The integrative research review. Beverley Hills, Calif.: Sage; 1997:15.
26. Dickersin K, Scherer R, LeFebvre C. Identifying relevant studies for systematic reviews. BMJ 1994;309:1286–91.
27. Thompson SG, Higgins JPT. Treating individuals 4: can meta-analysis help target interventions at individuals most likely to benefit? Lancet 2005;365:341–6.
28. Rosenthal R. Combining results of independent studies. Psychol Bull 1978;85:185–93.
29. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Erlbaum; 1988.
30. Bandolier. Describing results of trials and reviews. Available at: http://www.jr2.ox.ac.uk/bandolier/booth/glossary/outputs.html. Accessed July 29, 2005.
31. Rosenthal R. Meta-analytic procedures for social research. Newbury Park, Calif.: Sage; 1991:34–5.
32. Kraemer HC, Wilson T, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. Arch Gen Psychiatry 2002;59:877–83.
33. Webb E, Campbell D, Schwartz R, Sechrest L, Grove J. Nonreactive measures in the social sciences. Boston, Mass.: Houghton Mifflin; 1981.
34. Yu J, Cooper H. A quantitative review of research design effects on response rates to questionnaires. J Marketing Res 1983;20:36–44.
35. Ambady N, Rosenthal R. Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. Psychol Bull 1992;111:256–74.
36. Rosenthal R, Rubin DB. Interpersonal expectancy effects: the first 345 studies. Behav Brain Stud 1978;3:363–74.
37. Lipsey M. What do we learn from 400 research studies on the effectiveness of treatment with juvenile delinquents? In: McGuire J, ed. What works: reducing re-offending. New York: Wiley; 1995:63–111.
38. Lipsey MW, Wilson DB. The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. Am Psychol 1993;48:1181–1209.
39. Halvorsen KT, Burdick E, Colditz GA, Frazier HS, Mosteller F. Combining results from independent investigations: meta-analysis in clinical research. In: Bailar JC, Mosteller F, eds. Medical uses of statistics. 2nd ed. Boston, Mass.: NEJM Books; 1992: 413–26.
40. Chalmers TC, Lau J. Meta-analytic stimulus for changes in clinical trials. Stat Methods Med Res 1993;2:162–72.
41. Egger M, Smith D, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315: 629–34.
42. Clarke M, Chalmers I. Discussion sections in reports of controlled trials published in general medical journals: islands in search of continents? JAMA 1998;280:280–2.
43. Petrucci N, Iacovelli W. Ventilation with lower tidal volumes versus traditional volumes in adults for acute lung injury and acute respiratory distress syndrome. The Cochrane Database Systematic Reviews 2004;2:CD003844.pub2.DOI:10.1002/14651858.CD003844.pub2.
44. Geddes JR, Freemantle N, Mason J. Eccles MP, Boynton J. Selective serotonin reuptake inhibitors (SSRIs) versus other antidepressants for depression. The Cochran Database of Systematic Reviews 2005;4:CD001851.DOI:10.1002/14651858.CD001851.
45. Garner S, Fidan D, Frankish R, Judd M, Shea B, Towheed T, et al. Celecoxib for rheumatoid arthritis. The Cochran Database of Systematic Reviews 2002;4:CD003831.DOI:10.1002/14651858.CD003831.
46. Wienecke T, Gotzsche PC. Paracetamol versus nonsteroidal anti-inflammatory drugs for rheumatoid arthritis. The Cochran Database of Systematic Reviews 2004;1:CD003789.pub2.DOI:10.1002/14651858.CD003789.pub2.
47. Garner S, Fidan D, Frankish R, Judd M. Towheed T, Wells G, et al. Rofecoxib for rheumatoid arthritis. The Cochran Database of Systematic Reviews 2005;1:CD003685.pub2.DOI:10.1002/14651858.CD003685.pub2.
48. Greenberg RP, Bornstein RF, Zborowski MJ, Fisher, S, Greenberg MD. A meta-analysis of fluoxetine outcome in the treatment of depression. J Nerv Ment Dis 1994;182:547–51.
49. Eichacker PQ, Banks SM, Cui X, Natanson C. Meta-analysis of ALI and ARDS trials testing low tidal volumes. Am J Respir Crit Care Med 2002;166:1510–4.
50. Li H. The resolution of some paradoxes relating to reliability and validity. J Educ Behav Stat 2003;28:89–95.
51. Cooper H, Hedges LV. Potentials and limitations of research synthesis. In: Cooper H, Hedges LV, eds. The handbook of research synthesis. New York: Russell Sage; 1994.
52. Campbell D, Stanley J. Experimental and quasi-experimental designs for research. Chicago, Ill.: Rand McNally; 1963.
53. Noble JH. Peer review: quality control of applied social research. Science 1974;185:916–21.
54. Thornton S. Karl Popper. In: Zalta EN, ed. The Stanford encyclopedia of philosophy. (2002). Available at: http://plato.stanford.edu/archives/win2002/entries/popper/. Accessed February 25, 2005.
55. Acute Respiratory Distress Syndrome Network. Ventilation with lower tidal volumes compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. N Engl J Med 2000;342:1301–8.
56. Noble JH. Critique of the ARDS Network experimental design, implementation, and reporting. Testimony presented to the Office of Human Research Protection (OHRP) ARDS Network Trials Expert Review Panel, June 10, 2003. Washington, D.C. Available at: http://www.ahrp.org/testimonypresentations/ARDSNet0603/noble.html. Accessed March 14, 2005.
57. Deans KJ, Minneci PC, Cui X, Banks SM, Natanson C, Eichacker PQ, et al. Editorial—Mechanical ventilation in ARDS: one size does not fit all. Crit Care Med 2005;33:1141–3.
58. Solomon SD, McMurray JJV, Pfeffer MA, Wittes J, Fowler R, Finn P, et al. Cardiovascular risk associated with celecoxib in a

clinical trial for colorectal adenoma prevention. N Engl J Med 2005;352:1071–80.
59. Bresalier RSK, Sandler RS, Quan H, Bolognese JA, Oxenius B, Horgan K, et al. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. N Engl J Med 2005;352:1092–1102.
60. Nussmeier NA, Whelton AA, Brown MT, Langford RM, Hoeft A, Parlow JL, et al. Complications of the COX-2 inhibitors parecoxib and valdecoxib after cardiac surgery. N Engl J Med 2005;352:1081–91.
61. Psaty BM, Furberg CD. COX-2 inhibitors—lessons in drug safety. N Engl J Med 2005;352:1133–5.
62. Drazen JM. COX-2 inhibitors—a lesson in unexpected problems. N Engl J Med 2005;352:1131–2.
63. Kaufman M. FDA panel mulls whether all COX-2 drugs have same risk. Washington Post 2005;(Feb. 17):News A03.
64. Kaufman M. Vioxx alternative potentially as risky, official says. Washington Post 2005;(Feb 18):News A12.
65. Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002;288: 358–62.
66. Erdfelder E, Faul F, Buchner A. G*Power: a general power analysis program. Behav Res Meth Ins C 1996;28:1–11. Available at: http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/how_to_use_gpower.html. Accessed February 25, 2005.
67. Willis JL. Equality in clinical trials: drugs and gender. Rockville, Md.: U.S. Food and Drug Administration, Office of Special Health Issues; 1997. Available at: http://www.fda/gov/oashi/aids/equal.html. Accessed February 25, 2005.
68. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005;294:218–28.
69. Sugarman J. Editorial—Lying, cheating and stealing in clinical research. Clin Trials 2004;1:475–6.
70. Martinson BC, Anderson MS, deVries R. Scientists behaving badly. Nature 2005;435:737–8. Available at: http://www.nature.com/nature/journal/v435/n7043/full/435737a.html. Accessed June 30, 2005.
71. Lemmens T. Leopards in the temple: restoring scientific integrity to the commercialized research scene. Int Comp Health Law Ethics 2004;(Winter):641–57.
72. Wolfe SM. Letter to FDA revealing heart dangers in an unpublished clinical trial of Celebrex. Washington, D.C.: Public Citizen Health Research Group; January 31, 2005. Available at: http://www.citizen.org/publications/print_release.cfm?ID=7359. Accessed February 25, 2005.
73. Barton J, Greenwood JC. Letter to FDA seeking information on antidepressants. Washington, D.C.: U.S. House of Representatives, Committee on Energy and Commerce, Subcommittee on Oversight and Investigations; March 24, 2005. Available at: http://energycommerce.house.gov/108/letters/03242004_1242.htm. Accessed September 30, 2004.
74. Mosholder AD. Written statement summarizing Dr. Andrew Mosholder's role in the FDA review of pediatric use of antidepressant drugs, including interview by the FDA Office of Internal Affairs regarding disclosure of his findings in the February 1, 2005 edition of the San Francisco Chronicle. Washington, D.C.: U.S. House of Representatives, House Committee on Energy and Commerce, Subcommittee on Oversight and Investigations Hearing; September 23, 2004.
75. AG New York v. GlaxoSmithKline, Sec. 38, June 2, 2004.
76. McCabe AR. A precarious balancing act—the role of the FDA as protector of public health and industry wealth. Suffolk U Law Rev 2003;36:787–819.
77. Garland EJ. Facing the evidence: antidepressant treatment in children and adolescents. Can Med Assoc J 2004;170:489–92.
78. Whittington CJ, Kendall T, Fonagy P, Cottrell D, Cotgrove A, Boddington E. Selective serotonin reuptake inhibitors in childhood depression: systematic review of published versus unpublished data. Lancet 2004;363:1341–5.
79. Editorial. Depressing research. Lancet 2004;363:1335.
80. Wohlfarth T, Lekkerkerker F, van Zwieten B. Correspondence: use of selective serotonin reuptake inhibitors in childhood depression. Lancet 2004;364:659.
81. Fergusson D, Doucette S, Glass KC, Shapiro S, Healy D, Hebert P, et al. Association between suicide attempts and selective serotonin reuptake inhibitors: systematic review of randomised controlled trials. BMJ 2005;330:396.
82. BMJ rapid responses to: Fergusson, Doucette, Glass, Shapiro, Healy, Hebert, et al. BMJ 2005;330:396. Available at: http://bmj.bmjjournals.com/cgi/eletters/330/7488/396. Accessed March 13, 2005.
83. Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. J Clin Epidemiol 1995;48:45–57.
84. Healy D, Cattell D. Interface between authorship, industry and science in the domain of therapeutics. BMJ 2003;183:22–7.
85. Rennie D, Flanagin A. Authorship! Authorship! Guests, ghosts, grafters, and the two-sided coin. JAMA 1994;274:469–71.
86. Flanagin A, Carey LA, Fontanarosa PB, Phillips SG, Pace BP, Lundberg GD, et al. Prevalence of articles with honorary authors and ghost authors in peer-reviewed medical journals. JAMA 1998;280:222–4.
87. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systemic review. BMJ 2003;326:1167–70.
88. Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(i)ased medicine—selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. BMJ 2003;326:1171–3.
89. Angell M. The truth about the pharmaceutical industry: how they deceive us and what to do about it. New York: Random House; 2004.
90. Herdelin AC, Scott DL. Experimental studies of the program of assertive community treatment (PACT): a meta-analysis. J Disabil Policy Stud 1999;10(1):53–89.
91. Schulman KA, Seils DM, Timbie JW, Sugarman J, Dame LA, Weinfurt KP, et al. A national survey of provisions in clinical trial agreements between medical schools and sponsors. N Engl J Med 2002;347:1335–41.
92. P.L 102-571, Title I, Sec. 102, 105 Stat. 4491.
93. Food and Drug Administration Modernization Act of 1997, 21 U.S.C. Sec 301 (1997).
94. Taveggia T. Resolving research controversy through empirical cumulation. Sociol Method Res 1974;2:395–407.
95. Hammerschmidt DE, Franklin M. Secrecy in medical journals. Minnesota Medicine 2005;(March):34–5. [commentary]
96. Dobson R, Lenzer J. US regulator suppresses vital data on prescription drugs on sale in Britain. The Independent (UK). June 12, 2005. Available at: http://news.independent.co.uk/uk/health_medical/article225491.ece. Accessed July 29, 2005.
97. Lenzer J, Pyke N. Was Traci Johnson driven to suicide by anti-depressants? That's a trade secret, say US officials. The Independent (UK). June 19, 2005. Available at: http://news.independent.co.uk/uk/health_medical/article226432.ece. Accessed July 29, 2005.
98. Graham DJ, Campen D, Hui R, Spence M, Cheetham C, Levy G, et al. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and

non-selective non-steroidal anti-inflammatory drugs: nested case-control study. Lancet 2005;365:475–81.

99. Kotulak R. FDA cut off critic's access to drug database. Chicago Tribune 2005;(Feb 20):Health sidebar 01.
100. Graham J, James F. Flaws in drug agency put consumer at risk: critics of FDA cite conflicts of interest, lack of enforcement authority. Chicago Tribune 2005;(Feb 20):Health sidebar 01.
101. House of Commons Health Committee. The influence of the pharmaceutical industry: fourth report of session 2004-2005, volume I. London: United Kingdom Parliament. 2005. Available at: http://www.publications.parliament.uk/pa/cm200405/cmselect/cmhealth/42/42.pdf. Accessed April 7, 2005.
102. Eggerston L. Drug approval system questioned in US and Canada. CMAJ 2005;172:317–8.
103. CBC Health & Science News. Review of arthritis drugs raises questions about role of regulators. December 7, 2004. Available at: http://www.cbc.ca/story/science/national/2004/12/07/arthritis-drugs04207.html. Accessed July 30, 2005.
104. Faunce TA, Tomossy GF. The UK House of Commons report on the influence of the pharmaceutical industry: lessons for equitable access to medicines in Australia. Monash Bioethics Rev 2005;23:38–42.
105. Richards T. Conduct of drug trials in poor countries must improve. BMJ 2005;330:1466. Available at: http://bmj.bmjjournals.com/cgi/content/full/330/7506/1466-a. Accessed June 30, 2005.
106. Nundy S, Gulhati CM. A new colonialism?—conducting clinical trials in India. N Engl J Med 2005;352:16.