Journal of Law and Policy

Volume 12
Issue 2
SCIENCE FOR JUDGES II:
The Practice of Epidemiology and Administrative
Agency Created Science

Article 3

2004

Systematic Review of Medical Evidence

John P.A. Ioannidis

SYSTEMATIC REVIEW OF MEDICAL EVIDENCE

*John P.A. Ioannidis, M.D. & Joseph Lau, M.D.**

I. THE NEED FOR SYSTEMATIC REVIEWS OF MEDICAL EVIDENCE

Medical questions that arise in everyday clinical practice are often complex. The fascinating advances of the basic biomedical sciences of the last two decades, such as the mapping of the human genome, have created a widespread notion in the general public that medical knowledge is highly advanced, well founded, scientifically documented, and exact. Despite considerable progress, however, the clinical practice aspect of the medical science has not reached the same level of exactness as the physical sciences or even the basic biological sciences. The complexity of the human organism and the unpredictable nature of interactions between specific interventions and specific patients are difficult to reduce to the simplicity of physical laws. Knowledge about molecular and cellular mechanisms is enlightening, and animal data are very useful for medical progress, but it cannot be taken for granted that they will be readily translated to medical practice, let alone to the care of individual patients. First, basic knowledge from biological and animal systems must be verified in humans, and evidence must be corroborated from the application of various medical interventions in patients. Biological concepts must be tested in large numbers of human subjects to reduce the level of statistical uncertainty. Even when this is done, highly promising findings of basic biological research often do not turn out to hold

* John P.A. Ioannidis, M.D. is Adjunct Professor of Medicine, Tufts University. Joseph Lau, M.D. is Professor of Medicine, Institute for Clinical Research and Health Policy Studies, Tufts-New England Medical Center.

true in routine clinical settings.[1]

An outsider may expect decisions about the best medical action in a specific setting, the alternatives, and whether some actions are clearly inadequate or unacceptable, to be uncomplicated. Yet, medical decisions often involve a large number of complex clinical questions. Typically several decisions need to be made concurrently or sequentially. This is most clear in the case of hospitalized patients who have several medical problems that may affect each other. These patients may already be receiving many medications, and while each one of the medications may have been tested alone in clinical studies, there may be limited knowledge on their interactions. In addition, the clinical course of hospitalized patients may be very fragile, and rapid changes in their conditions may occur. The available background information for making rapid decisions may vary, and it is not uncommon for even experienced physicians to miss large parts of the diagnostic puzzle or even the main diagnosis itself. For patients with highly complex interrelated problems, there may be little instructive precedent in the medical literature on how to handle similar cases with the same combination of problems. For example, for a patient with hypertension (increased blood pressure), there may be extensive data available on similar cases, since one in five Americans has high blood pressure.[2] On the other hand, prior knowledge may be more limited and more difficult to apply when someone also has heart disease, diabetes mellitus, high cholesterol, osteoporosis, liver dysfunction, kidney impairment, and is already taking several medications.

Even for uncomplicated situations, medical decision-making still can be quite challenging. For a healthy young woman who

[1] *See* Despina G. Contopoulos-Ioannidis et. al., *Translation of Highly Promising Basic Science Research into Clinical Applications*, 114 AM. J. MED. 477 (2003) (reporting that of 101 findings published in major basic science journals from 1977 through 1982 where explicit promises were made for clinical applications, only 5 of them had resulted in some licensed clinical use twenty years later and only one of them had a considerable clinical impact).

[2] *See* American Heart Association, High Blood Pressure Statistics, *at* http://www.americanheart.org/presenter.jhtml?identifier=4621 (last visited March 2, 2004).

comes to a physician for a routine check-up, there are many decisions to make. One of the many decisions may be whether to offer mammography as a screen test for breast cancer. Screening mammography has been recommended by several professional organizations and government agencies and is widely deemed as an effective means to reduce breast cancer mortality; yet, this apparently straightforward recommendation is controversial, and the evidence supporting this recommendation has been challenged.[3] As all medical tests suffer from unavoidable errors, finding an abnormality on a mammogram does not necessarily mean the woman has breast cancer. It is possible that the abnormality may be a case of "false positive" results. False positive results are inherent uncertainties in the ability of a test to differentiate disease from non-disease conditions; even the best tests are susceptible to false positive results, even when performed carefully with state-of-the-art equipment. A false positive test can lead down a path of undesirable outcomes as a consequence of further testing and medical management. For example, a woman who has a false positive mammogram suggesting abnormalities is a candidate for a breast biopsy. While the risk of serious complications resulting from a breast biopsy is low, it still carries small risks of bleeding and infection, and it also creates anxiety in the patient. For other potentially more dangerous tests like liver biopsy, serious bleeding may lead to hospitalization. Hospital admission may lead to other complications, such as acquiring an infection while in the hospital. In all, every action in medicine has to be carefully balanced for its potential benefits and risks,[4] and it is often difficult to take into account all the possible interactions and developments in patients.

The complexity and uncertainty in medical decision-making has shown that it is very important to utilize the best available evidence in each case and to scrutinize the quality of the evidence

[3] *See, e.g.*, Peter C. Gotzsche & Ole Olsen, *Is Screening for Breast Cancer with Mammography Justifiable?*, 255 LANCET 129 (2000).

[4] For a discussion of approaches towards balancing risks against benefits, see Paul P. Glasziou & Les Irwig, *An Evidence Based Approach to Individualising Treatment*, 311 BRIT. MED. J. 1356 (1995).

accumulated from past medical research. The questions are, where should this evidence come from, and how should it be appraised and synthesized to arrive at a most meaningful and seasoned decision?

II. OVERVIEW OF THE HIERARCHY OF MEDICAL EVIDENCE

A. Expert Opinion

Medical practice and policy-making often relied on expert opinions. Expert opinions may be offered as patient care consultations, informal discussions among colleagues, lectures, or in written forms as book chapters, review articles, or editorials. In the last few decades, there has been an increasing recognition that relying on expert opinion to make medical decision has significant shortcomings. Analyses of medical review articles have shown that coverage of the potentially relevant data by the authors is often sketchy, anecdotal, and highly subjective or even biased.[5] The proposed recommendations may not represent the findings of the actual data from medical studies, and sometimes the recommendations may be in favor of totally ineffective therapies, or against therapies that have been shown to be beneficial.[6]

The unbalanced reviews offered by experts should not be surprising. In the current era of the information revolution, the amount of medical information that is being generated is staggering. Over a million biomedical articles are published in international journals every year. Even an expert in a highly

[5] The first major assessment of the deficiencies of the traditional review article in the medical literature was published in 1987. *See* Cynthia D. Mulrow, *The Medical Review Article: State of the Science*, 106 ANNALS INTERNAL MED. 485 (1987).

[6] In 1992, it was shown that the best written documents based on expert opinion often failed to follow what the true data suggested in the treatment of acute myocardial infarction, and sometimes they even contradicted the data. *See* Eliott M. Antman et. al., *A Comparison of Results of Meta-analyses of Randomized Control Trials and Recommendations of Clinical Experts. Treatments for Myocardial Infarction*, 268 J. AM. MED. ASS'N 240 (1992).

specialized field would have to read many thousands of articles a year to be certain that she is not missing essential information. The variable quality of this information further adds to the problem, since one needs to have considerable training in research methodology and to spend a lot of time evaluating the strengths and limitations of each study. Experts who are influential often are excellent in their specialty, but they may lack formal training in research methodologies and in critically reviewing the literature. In addition, subjectivity is unavoidable. Finally, it is probably quite common for experts to have some conflicts of interest, financial or otherwise, or topics in which they have invested their careers or reputations.[7] The relative contribution of these reasons towards the inadequacy of using experts alone for judging medical evidence is unknown, and probably it varies from case to case. Given these limitations, however, it is currently generally accepted that expert opinion in the absence of data comprises the lowest level of evidence in the hierarchy of medical evidence.[8]

[7] The potential conflicts of experts have recently drawn considerable attention. *See, e.g.*, J. Abraham, *The Science and Politics of Medicines Control*, 26 DRUG SAFETY 135 (2003); Sheldon Krimsky et. al., *Scientific Journals and Their Authors' Financial Interests: a Pilot Study*, 67 PSYCHOTHERAPY & PSYCHOSOMATICS 194 (1998); George N. Papanikolaou et. al., *Reporting of Conflicts of Interest in Guidelines of Preventive and Therapeutic Interventions*, 1 B.M.C. MED. RES. METHODOLOGY 3 (2001).

[8] There are several different scales for rating evidence. Simple three-tier systems such as those used by the United States Preventive Services Task Force are commonly adopted or modified. According to this scheme, level I evidence corresponds to evidence obtained from at least one properly randomized controlled trial; level II corresponds to evidence obtained from well-designed controlled trials without randomization (II-1), well-designed cohort or case-control analytic studies, preferably from more than one center or research group (II-2), or multiple time series with or without the intervention or uncontrolled experiments with dramatic results (II-3); and level III corresponds to opinions of respected authorities, based on clinical experience; descriptive studies and case reports; or reports of expert committees. For a comprehensive review of available systems to rate the medical evidence, *see* Suzanne West et. al., *Systems to Rate the Strength of Scientific Evidence*, Agency for Healthcare Research and Quality, Evidence report/Technology Assessment Number 47, Apr. 2002.

B. Non-randomized Studies

There is wide consensus that the strength of the evidence in medicine should depend on the amount, quality, and consistency of the available data that has been generated from biomedical research. The amount of data is important but far from sufficient. The quality of the data is very important, since small studies that are well designed may give more accurate conclusions than large studies that have clear flaws in their design and execution. Finally, for each medical question of interest many different studies may have been conducted. It is thus important to be able to acquire a synthesis of the available evidence and to examine the whole picture emerging from the totality of the data. When all the different studies on the same question agree, this consistency is reassuring, while disagreements need to be probed and explained, if possible. Exceptions to the rule are common in medicine.

1. Observational and Cross-Sectional Studies

The quality of medical studies is not straightforward to assess. Based on theoretical considerations, however, some types of studies are less susceptible to potential errors and biases compared with others. Case-reports of single observations and series of cases can provide useful information, but they lack a control comparison and thus one can only see what has happened to one or several patients without being able to tell what might have happened if a different course of action had been employed. Therefore, such observational studies without a control population are typically considered to be superior to plain expert opinion in the absence of data, but inferior to other types of study designs.[9] The same largely holds true of cross-sectional designs, where a population is studied in terms of outcomes and candidate factors of interest and

[9] There is an effort to improve the quality of case reports in the medical literature and to upgrade their status based on the principles of evidence-based medicine. *See, e.g.*, MILOS JENICEK, CLINICAL CASE REPORTING IN EVIDENCE-BASED MEDICINE, (2d ed. 2001); Jan P. Vandenbroucke, *In Defense of Case Reports and Case Series*, 134 ANNALS INTERNAL MED. 330 (2001).

associations are evaluated. In such studies, outcomes and candidate factors associated with them are measured at the same time, so it is not possible to tell where there is a causative relationship in the absence of a temporal sequence.

2. Case-Control and Retrospective Cohort Studies

Case-control studies and retrospective cohort studies have the advantage that both patients with an adverse outcome and those without an outcome of interest are available and that these groups can be compared in terms of various factors that may or may not be associated with the outcomes. A temporal sequence of events is available. Again, however, the detection of significant associations on the basis of statistical tests does not guarantee clinical or biological causality. Furthermore, there can be bias in the choice of the control groups (subjects without the outcome of interest). Finally, these studies are retrospective, in that they are based on collecting information from the past, and such information may be subject to major errors or problems from missing data since the data has not been collected with a specific prospective plan in mind.

3. Prospective Cohort Studies

Prospective cohort studies have a higher level in the hierarchy of evidence. In this case, groups of people are followed into the future and the potential association between various factors and outcomes can be discerned over time. These studies by definition may take a long time to conduct, and they are more expensive than retrospective studies, but there is room for better data collection according to pre-specified rules. Still, these studies suffer from the limitation that people are not assigned randomly to having or not having a specific factor, so there is considerable room for bias. For example, one can compare patients who had a new aggressive type of surgery against those who chose to have the old type of operation. It is possible that patients and physicians who choose the new aggressive type of intervention are those who have the worst or more advanced disease and are desperate to try something new. If they fare worse than those who got the old type of operation, this may be due simply to the fact they were at a worse condition even before the operation. Conversely, the new operation may be reserved only for patients who are in the best possible condition if patients and physicians feel the risk of adverse events may be too high for patients having any background problems. In this case, the new operation may have superb outcomes only because the most favorable patients were selected upfront. This problem is known as confounding by indication. Confounding may involve also a number of other known or unknown patient characteristics beyond the severity of illness. Thus, for all semi-experimental studies (both prospective and retrospective cohort studies and case-control studies), to arrive at a net effect of the factor of interest, it is important to adjust the compared groups for all possible imbalances that may exist between them. Adjustment is rarely straightforward, however, and it is unlikely that all potential imbalances between the compared groups can be identified a priori and taken into account.[10]

4. Randomized Trials

The problem of confounding is avoided when randomized trials are conducted. In randomized trials, subjects are divided into two or more groups in a random fashion that guarantees that imbalances between the compared groups are unlikely, and if present, they are purely due to chance and likely to cancel themselves out if a sufficiently large number of subjects can be randomized. Randomized trials have been used for over fifty years in medicine, and they are accepted as the reference standard for assessing the efficacy of medical interventions with the least bias.[11]

Even within randomized trials, not all studies are the same. Besides sample size, studies may differ on various aspects of study

[10] For more details on semi-experimental studies, see KENNETH J. ROTHMAN & SANDER GREENLAND, MODERN EPIDEMIOLOGY (1998).

[11] Randomized trials are the perceived gold standard for evidence despite their acknowledged limitations. *See, e.g.*, Alejandro R. Jadad & Drummond Rennie, *The Randomized Controlled Trial Gets a Middle-Aged Checkup*, 279 J. AM. MED. ASS'N 319 (1998).

design. These aspects include whether blinding is used or not; how well randomization has been performed; whether the allocation sequence has been adequately concealed or not; and whether follow up and patient flow has been carefully examined, as far as patients withdrawing from the assigned treatment are concerned. There is considerable debate on whether these or other study parameters may be affecting the results of the trials. Some evidence suggests that poor quality studies may tend to inflate the estimated efficacy of medical interventions.[12] However, quality is often hard to define and/or may be difficult to discern from published medical reports because what is reported may not be an accurate reflection of what actually happened during the trial.[13] Thus in some cases, studies with poor quality characteristics may actually show a smaller effect for an experimental intervention than studies with good quality ratings.[14] In the presence of quality defects, the direction of bias in the results is often difficult to determine.

Given this uncertainty, there is increasing understanding that the quality of medical research needs to be improved and held at high standards. For some medical fields, the overall feeling is that the quality of research methods used has been lagging behind acceptable standards. Several years ago, the editor of the *Lancet* wrote an editorial where in the title he questioned whether surgical trial research exists or is all just "comic opera."[15] Although this

[12] *See, e.g.*, Kenneth F. Schulz et. al., *Empirical Evidence of Bias. Dimensions of Methodological Quality Associated with Estimates of Treatment Effects in Controlled Trials*, 273 J. AM. MED. ASS'N 408 (1995) (finding that lack of allocation concealment inflates the observed treatment effect by 40% and lack of double-blinding inflates the treatment effect by about 20%). *See also* Peter Juni et. al., *Systematic Reviews in Health Care: Assessing the Quality of Controlled Clinical Trials*, 323 BRIT. MED. J. 42-46 (2001) (reviewing such studies).

[13] John P.A. Ioannidis & Joseph Lau, *Can Quality of Clinical Trials and Meta-analyses be Quantified?*, 352 LANCET 590 (1998).

[14] This has been demonstrated in studies of infectious disease-related interventions by Ethan M. Balk et. al., *Correlation of Quality Measures with Estimates of Treatment Effect in Meta-analyses of Randomized Controlled Trials*, 287 J. AM. MED. ASS'N 2973 (2002).

[15] Richard Horton, *Surgical Research or Comic Opera: Questions, but Few*

may be an exaggeration, it reflects the difficulties in improving the quality of medical research. One vein of effort has focused attention on standardizing the reporting of the results of medical studies in the peer-reviewed literature. Comprehensive checklists that ensure all the important information is conveyed have been developed and accepted for randomized trials.[16] Efforts are underway to develop and agree on similar standards for the reporting of other types of studies.

III. DIVERSITY OF MEDICAL EVIDENCE

The hierarchy of evidence discussed above is not cut in stone. For some types of important medical questions, some or all of these types of study designs may not be equally applicable. For example, in studying the effects of potentially harmful factors (e.g., the harmful effects of radiation or smoking), it is unethical to use randomized trials. Randomized trials may also be infeasible when it is difficult to commit individuals to specific interventions, e.g., making some behavioral or nutritional changes.[17] To evaluate the accuracy of diagnostic tests, a different type of study is necessary, where people with and without the disease are subjected to the diagnostic test of interest and a comparator reference standard test to see whether the test under study comes close to the reference standard.[18] Medical prognosis also cannot be addressed with randomized studies, but usually semi-experimental studies are

Answers, 347 LANCET 984 (1996).

[16] The most widely adopted is the *Consolidated Standards of Reporting Trials* (CONSORT). *See* Doug G. Altman et. al., *The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration*, 134 ANNALS INTERNAL MED. 663 (2001).

[17] This rationale is exemplified by positions held by Meir Stampfer. *See* Meir Stampfer, *Observational Epidemiology Is the Preferred Means of Evaluating Effects of Behavioral and Lifestyle Modification*, 18 CONTROLLED CLINICAL TRIALS 494 (1997).

[18] Discussion of diagnostic test studies is beyond the focus of this article. For more details, see XIAO-HUA ZHOU, NANCY A. OBUCHOWSKI & DONNA K. MCCLISH, STATISTICAL METHODS IN DIAGNOSTIC MEDICINE (2002).

used.[19] Moreover, even when several different designs can be equally applied to a specific question, there is no guarantee that studies using designs at a higher level in the hierarchy of evidence are necessarily superior to studies using designs at a theoretically lower level. Randomized trials sometimes disagree with the results of semi-experimental and observational studies on the same question,[20] but this does not mean that in all such cases of disagreements the randomized trials are correct and the observational evidence is wrong. Since bias may interfere in all kinds of medical studies, sometimes poorly-done randomized studies may be more unreliable than well-done cohort studies. Furthermore, random error may affect the results of any human study; by chance the results of randomized trials may occasionally be further away from the truth than the results of other studies.

Within the same study design, sample size is important to consider, and larger studies are likely to be more definitive than smaller ones. There is no certainty that they are always likely to be more correct, though. Empirical evaluations have shown that small studies disagree with larger studies about a quarter of the times beyond what would be anticipated by chance alone. In these cases, it is not straightforward to tell which one is right or whether both small and larger studies provide different sides of the truth and complementary evidence. Interestingly, it has also been found that large trials (defined as studies with at least 1,000 patients) disagree among themselves about as frequently as the discrepancy between

[19] However, it is possible to evaluate whether the application of a prognostic system improves patient outcomes in the long-term or the efficiency and cost-benefit of the health system. The same holds true for the application of diagnostic tests. Such applications and their effects may then be studied with randomized trials.

[20] *See, e.g.*, John P.A. Ioannidis et. al., *Comparison of Evidence of Treatment Effects in Randomized and Nonrandomized Studies*, 286 J. AM. MED. ASS'N 821 (2001) (comparing large versus smaller trials); *but see, e.g.*, John Concato et. al., *Randomized, Controlled Trials, Observational Studies and Randomized, Controlled Trials*, 342 NEW. ENG. J. MED. 1878 (2000) (offering a dissenting view). For discrepancies between very large trials (also called "megatrials"), *see* T.A. Furukawa et. al., *Discrepancies among Megatrials*, 53 J. CLINICAL EPIDEMIOLOGY 1193 (2000).

large trials and meta-analyses of small trials.[21]

In all, it is unavoidable that sometimes studies seemingly addressing the same question may reach different results. There could be many reasons for this variability. Studies are conducted with different designs, at different settings, in different populations, with different background management and treatment, in different countries, at different time periods.[22] All of these factors, in addition to the play of chance, could contribute to variability. The whole emerging picture may be quite confusing when an effort is made to reach a final conclusion even by the most well informed and experienced experts. The quantity of data, although useful in assessing consistency, can be overwhelming. Systematic reviews and meta-analyses can provide a transparent framework in such situations to make sense of the disparate data.

IV. Systematic Reviews and Meta-analyses

A systematic review is a comprehensive assessment of the medical literature on a topic of interest using a priori specified rules for the search, identification, and eligibility of the pertinent studies and for the abstraction of relevant data.[23] The systematic nature of the process according to clear-cut rules differentiates a

[21] Several empirical evaluations have been published on the rate and reasons of disagreements between small and larger studies. *See* John P.A. Ioannidis et. al., *Issues in Comparisons between Meta-analyses and Large Trials*, 279 J. Am. Med. Ass'n 1089 (1998); Joseph C. Cappelleri et. al., *Large Trials vs Meta-analysis of Smaller Trials: How Do Their Results Compare?*, 276 J. Am. Med. Ass'n 1332 (1996); Jose Villar et. al., *Predictive Ability of Meta-analyses of Randomized Controlled Trials*, 345 Lancet 772 (1995); Jacques LeLorier et. al., *Discrepancies between Meta-analyses and Subsequent Large Randomized, Controlled Trials*, 337 N. Eng. J. Med. 536 (1997); T. A. Furukawa et al., *Discrepancies among Megatrials,* 53 J. Clin. Epidemiology 1193 (2000).

[22] Joseph Lau et. al., *Summing up Evidence: One Answer Is Not Always Enough*, 351 Lancet 123 (1998).

[23] Deborah J. Cook et. al., *Systematic Reviews: Synthesis of Best Evidence for Clinical Decisions*, 126 Annals Internal Med. 376 (1997); Cynthia D. Mulrow et. al., *Systematic Reviews: Critical Links in the Great Chain of Evidence*, 126 Annals Internal Med. 389 (1997).

systematic review from a traditional review authored by experts without specific rules. Systematic reviews currently have assumed a key place in the generation of recommendations for medical practice and for clinical decision-making.[24] An international initiative such as the Cochrane Collaboration is aiming to conduct systematic reviews to cover all major aspects of health care.[25] As of the fall of 2003, the Cochrane Library includes 1,754 completed Cochrane reviews and 1,304 protocols for ongoing reviews. In addition, there are 4,123 abstracts of completed systematic reviews in medical journals, theses, and reports that have been included in the Database of Abstracts of Reports of Effects (DARE), also a part of the Cochrane Library.[26] In the United States, the Agency for Healthcare Research and Quality has designated 13 evidence-based practice centers to produce comprehensive evidence reports and technology assessments on a variety of health care topics.[27] Over 100 of these reports have been produced over the past six years. These reports are based on systematic reviews of the medical literature and analyses of relevant databases, and are used by various government agencies and professional and health care organizations.

Systematic reviews target upfront a precisely defined clinical question or set of questions. Depending on the question and the

[24] Lisa A. Bero & Alejandro R. Jadad, *How Consumers and Policymakers Can Use Systematic Reviews for Decision Making*, 127 ANNALS INTERNAL MED. 37 (1997); Deborah J. Cook et. al., *The Relation between Systematic Reviews and Practice Guidelines*, 127 ANNALS INTERNAL MED. 210 (1997).

[25] *See* THE COCHRANE COLLABORATION, *at* http://www.cochrane.org (last visited March 2, 2004); Lisa Bero & Drummond Reenie, *The Cochrane Collaboration: Preparing, Maintaining, and Disseminating Systematic Reviews of the Effects of Health Care*, 274 J. AM. MED. ASS'N 1935 (1995); Mike Clarke & Peter Langhorne, *Revisiting the Cochrane Collaboration: Meeting the Challenge of Archie Cochrane—and Facing up to Some New Ones*, 323 BRIT. MED. J. 821 (2000).

[26] *See* THE COCHRANE COLLABORATION, *at* http://www.cochrane.org (last visited March 2, 2004) (access to the Cochrane Library is limited to those with a subscription).

[27] *See* AGENCY FOR HEALTHCARE RESEARCH & QUALITY, AHRQ PUB. NO. 03-P006, EVIDENCE-BASED PRACTICE CENTERS (2003), *at* http://www.ahrq.gov/clinic/epc/.

eligibility criteria that are considered appropriate for selecting studies on the question(s), some systematic reviews may include a large number of eligible studies, while others may systematically scrutinize hundreds and thousands of references from the medical literature, only to conclude that no eligible study exists that directly addresses the question of interest.[28] Systematic reviews that identify such lack of evidence are still useful since they clearly show the direction for designing new studies in a field where evidence-based inferences are urgently needed. Even in the presence of data, systematic reviews may often conclude that the available evidence is insufficient, controversial, or inconclusive and that further studies are needed. In other cases, systematic reviews conclude that the available data from several studies is consistent and conclusive.

Even in the presence of systematic methods for locating and appraising evidence, a systematic review can hardly be conclusive unless the collected data can be appropriately synthesized in a quantitative fashion. A meta-analysis is a quantitative synthesis of data from various sources addressing the same question; a systematic review is a prerequisite to a good meta-analysis. Most meta-analyses use statistical methods to combine data from published studies, and the information is available at the level of groups, e.g., the outcomes in patients in each compared group. Such meta-analyses of group data or meta-analyses of the published literature are basically retrospective designs and have several limitations including using only data that are available. Meta-analyses may also be designed prospectively to collect necessary data as well as to improve the consistency of collected data, i.e., with a plan to combine the results from several studies that are to be conducted.[29] Furthermore, meta-analyses may be

[28] For example, a systematic search of studies that might yield information on how to tell whether a very common condition, acute conjunctivitis, is due to viruses or bacteria, found no good study that had assessed the utility of clinical signs and symptoms for making this diagnosis. *See* Remco P. Rietveld et. al., *Diagnostic Impact of Signs and Symptoms in Acute Infectious Conjunctivitis: Systematic Literature Search*, 327 BRIT. MED. J. 789 (2003).

[29] For prospective meta-analysis may start with the construct of a study registry that tries to capture all studies performed on a specific topic right from

extended to use not only group data, but data from individual patients. Such individual-level data meta-analyses collate information from diverse pertinent studies with data on each patient on the exposures and outcomes of interest as well as a set of other parameters that may be considered to be of interest.[30] Clearly, meta-analyses of individual-level data are more difficult to conduct, and their performance must be justified by considering whether they are likely to provide additional or far more exact results compared to meta-analyses of group data.

A. Place of Systematic Reviews and Meta-analyses in Medical Decision-Making

Most schemes of the hierarchy of evidence have accepted that meta-analyses, especially those of randomized trials, are the highest level of medical evidence.[31] This is due to the fact that meta-analyses may combine data from a large number of studies and, given their systematic background, they have an objective opportunity for quantifying effects and associations, finding out whether the data are consistent, and in some cases, also quantifying the extent of inconsistency and probing into potential reasons for the existence of discrepancies between studies.

By definition, systematic reviews and meta-analyses are highly focused research designs. They address a specific question or a few questions, and it is unlikely that they can cover all the tangible and intangible issues that are involved in decision making. There is increasing appreciation that systematic reviews and meta-analyses should try to cover aspects of both efficacy and safety when medical interventions are involved. This is often very difficult to do because the quality and quantity of available data to assess the

their beginning (at the time they are designed or launched).

[30] For meta-analysis of individual-level data, *see* Leslie A. Stewart & M.K. Parmar, *Meta-analysis of the Literature or of Individual Patient Data: Is There a Difference?*, 341 LANCET 418 (1993); Leslie A. Stewart & Mike J. Clarke, *Practical Methodology of Meta-analyses (Overviews) Using Updated Individual Patient-Data—Cochrane Working Group*, 14 STATISTICS IN MED. 2067 (1995).

[31] *See, e.g.*, Robin Harbour & Juliet Miller, *A New System for Grading Recommendations in Evidence Based Guidelines*, 323 BRIT. MED. J. 334 (2001).

benefits and harms of medical interventions varies a great deal.[32] Even if several aspects of a clinical problem can be covered by a meta-analysis, it is unlikely that a final decision can be made directly on the basis of such evidence. Decision-making may involve other important parameters such as cost issues; the availability of resources; the availability or lack thereof of alternative interventions; utility ranking when several, diverse outcomes are involved; priority setting and overall strategic design in institutions or health care systems; and the subjective preferences of the patient.

Thus systematic reviews and meta-analyses provide some highly filtered and carefully analyzed information that needs to be placed in a broader context. Although physicians are increasingly educated in understanding these research tools, often the subtleties of meta-analyses may go beyond the appreciation of many practitioners and health care workers. However, the results and inferences of meta-analyses may be used efficiently for generating more easily interpretable clinical directives. Clinical practice guidelines, for example, are documents that aim to distill recommendations for important medical practice decision-making. In the past, most guidelines suffered from the same problems as expert reviews, since they were generated typically by one or several experts without any particular attention to the scientific basis of collecting, appraising, and synthesizing the available evidence. Major deficiencies in the quality of guidelines, even those published by top quality medical journals and reputable specialist societies have recently been scrutinized.[33] A robust

[32] Luis Gabriel Cuervo & Mike Clarke, *Balancing Benefits and Harms in Health Care: We Need to Get Better Evidence about Harms*, 327 BRIT. MED. J. 65 (2003).

[33] Defects on guidelines have been identified on their development, reporting, evidence base, and transparency. *See, e.g.*, Roberto Grilli et. al., *Practice Guidelines Developed by Specialty Societies: The Need for a Critical Appraisal*, 355 LANCET 103 (2000); T.M. Shaneyfelt et. al., *Are Guidelines Following Guidelines? The Methodological Quality of Clinical Practice Guidelines in the Peer-Reviewed Medical Literature*, 281 J. AM. MED. ASS'N 1900 (1999); Ioannis A. Giannakakis et. al., *Citation of Randomized Evidence in Support of Guidelines of Therapeutic and Preventive Interventions*, 55 J.

process must be set for producing high-quality guidelines. This process includes the systematic appraisal of the available data and a transparent procedure for reaching consensus among experts involved in guideline development, clarity in the presentation of alternative options, a balance of risks and benefits, editorial independence, and rigorous standards for prompt updating to include newly released research data that, in some cases, may strengthen, modify, or invalidate prior beliefs.[34]

B. Cumulative Meta-analysis

In this regard, meta-analysis can be seen as an exercise of updating the totality of available information over time. In cumulative meta-analysis, studies on a given medical question are ordered chronologically.[35] The results of each study are added one at a time in the order in which they appear. Cumulative meta-analysis provides a picture of evolving trends as medical evidence accumulates. It is possible to examine whether evidence has remained steady over time or major fluctuations have occurred over time. For example, occasionally, initial studies suggested an intervention may be highly effective, while subsequent studies may show that the same intervention is totally ineffective. Such big changes are infrequent when a large amount of evidence has accumulated from large and/or several studies. In some cases, however, changes were seen even when many early studies, including large ones, had accumulated over time.

For example, several studies on the efficacy of intravenous magnesium salts in acute myocardial infarction, including a study

Clinical Epidemiology 545 (2002).

[34] A widely adopted checklist for appraising guidelines has been developed by the AGREE Collaboration. *See* AGREE Collaboration, *Development and Validation of an International Appraisal Instrument for Assessing the Quality of Clinical Practice Guidelines: the AGREE Project*, 12 Quality & Safety Health Care 18 (2003).

[35] Joseph Lau et. al., *Cumulative Meta-analysis of Therapeutic Trials for Myocardial Infarction*, 327 N. Eng. J. Med. 248 (1992); Joseph Lau et. al., *Cumulative Meta-analysis of Clinical Trials Builds Evidence for Exemplary Medical Care*, 48 J. Clinical Epidemiology 45 (1995).

of 2,300 patients, had clearly documented large, significant reductions in the mortality risk with this inexpensive intervention. A larger study of over 50,000 patients showed no effect at all, however, and the same lack of efficacy was documented in a subsequent study with over 6,000 patients.[36] Thus, recommendations on the use of this therapy would change dramatically over time, based on the evolution of meta-analysis results. The interpretation of this example is further complicated by the fact that the standard of care of patients in these trials has evolved considerably over time. Many of the patients in the later trials were concurrently receiving other forms of effective treatments, thus confounding the interpretation of the results. Comparison of the older trials and the newer ones may no longer be valid and the treatment being evaluated may no longer be relevant, given the current standard of care.

In another example, several very large non-randomized studies of over 10,000 subjects suggested that estrogen replacement is a highly beneficial intervention for post-menopausal women with strong protective effects against cardiovascular disease. Nevertheless, a large randomized study showed clearly that there is no cardiovascular protection and the overall balance of risks against benefits makes estrogen replacement a highly unfavorable course of action for women in this age group.[37] In the case of estrogen replacement, different study designs (observational vs. randomized data) led to different conclusions, even though both types of designs entailed very large numbers of subjects.

[36] For more details on the magnesium controversy, see the report of the latest relevant trial: Magnesium in Coronaries (MAGIC) Trial Investigators, *Early Administration of Intravenous Magnesium to High-Risk Patients with Acute Myocardial Infarction in the Magnesium in Coronaries (MAGIC) Trial: A Randomised Controlled Trial*, 360 LANCET 1189 (2002). That report also alludes to the prior evidence on this controversial topic.

[37] The Women's Health Initiative (WHI) trial was the pivotal study that reversed prior beliefs on the indications of estrogen replacement in post-menopausal women. The study results have been published in a series of manuscripts in the *Journal of the American Medical Association* in 2002 and 2003 and have been extensively commented on in the scientific and lay literature.

The cases where large changes have occurred in our appreciation of the role of specific medical interventions are probably the exception, rather than the rule. In most cases, medical evidence does not change so much over time. Provided that they are well conducted,[38] systematic reviews and meta-analyses may thus provide the scientific ground for trusting the adequacy of medical actions or lack thereof in specific settings. Still, the evolutionary, cumulative nature of medical evidence should be kept in mind as a way of understanding that uncertainty about medical actions is unlikely to disappear completely even with large-scale evidence and well-conducted studies.

C. Methods of a Systematic Review

The general approach to conducting a systematic review consists of the following steps:

- Formulate answerable research question(s)
- Define the systematic review (meta-analysis) protocol (establish inclusion and exclusion criteria)
- Perform literature search
- Screen titles and abstracts of literature search results for potentially relevant studies and retrieve full articles
- Evaluate full articles according to criteria
- Critically appraise articles that met criteria
- Extract data for analysis
- Perform meta-analyses and sensitivity analyses as appropriate
- Interpret results

One of the most important tasks in conducting a systematic

[38] The conduct of meta-analysis requires rigorous standards as that of any good research. Guidelines have been developed for the reporting of meta-analyses. *See, e.g.*, David Moher et. al., *Improving the Quality of Reports of Meta-analyses of Randomised Controlled Trials: The QUORUM Statement. Quality of Reporting Meta-analyses*, 354 LANCET 1896 (1999); Donna F. Stroup et. al., *Meta-analysis of Observational Studies in Epidemiology: A Proposal for Reporting. Meta-analysis of Observational Studies in Epidemiology (MOOSE) Group*, 283 J. AM. MED. ASS'N 2008 (2000).

review is defining the question of interest and its limits.[39] This is important, since reviews with similar target-questions may reach different conclusions if they have important differences on how exactly the questions of interest are defined. It is important to clarify the interventions or associations of interest; what types of studies are to be examined and why some specific types of studies are to be selected or excluded; what types of people are eligible for including in these studies; which clinical settings are eligible and which ones are not; and also what electronic or other databases of medical information are to be searched for identifying the relevant studies.

Some subtle decisions may make a difference sometimes. For example, should published studies only be included or should abstracts be eligible as well?[40] For a recently targeted topic of interest, much of the pertinent literature may still be in abstract form, but for long-standing medical questions this is rarely an issue. Another problem is whether both English and foreign-language sources of evidence should be used. A tower of Babel bias has been described according to which non-English speaking scientists may publish their results in English language journals when they find significant differences, while they may prefer native language journals when their results show no significant differences.[41] Hopefully, the English language covers the overwhelming majority of the current scientific literature, so in most instances this problem is not a major issue, but exceptions may occur. Moreover, a reverse tower of Babel bias has been described; where all the Chinese and Russian language articles on acupuncture published in local non-English journal have significant results.[42]

[39] Carl Counsell, *Formulating Questions and Locating Primary Studies for Inclusion in Systematic Reviews*, 127 ANNALS INTERNAL MED. 380 (1997).

[40] Maureen O. Meade & W.S. Richardson, *Selecting and Appraising Studies for a Systematic Review*, 127 ANNALS INTERNAL MED. 531 (1997).

[41] Matthias Egger et. al., *Language Bias in Randomised Controlled Trials Published in English and German*, 350 LANCET 326 (1997).

[42] Andrew Vickers et. al., *Do Certain Countries Produce Only Positive Results? A Systematic Review of Controlled Trials*, 19 CONTROLLED CLINICAL TRIALS 159 (1998).

Other parameters that relate to the way the clinical question is circumscribed may have an even greater impact on the conclusions of a systematic review. Sometimes, a systematic review may have very loose eligibility criteria and may include people who differ in important aspects among themselves. In this case, one may question whether the overall findings of the review may be extrapolated equally to each type of people and settings. In other situations, a systematic review may use very narrow eligibility criteria. This would result in a more sharpened target-population where the results would be pertinent, but at the cost of the potential to generalize and with considerable loss of potentially useful information.[43]

D. Methods of a Meta-analysis

These considerations affect also the interpretation of a meta-analysis, since the first step towards conducting a quantitative synthesis of the data is typically the conduct of a systematic review. However, in a meta-analysis there are additional methodological issues that need to be decided and that may affect the results and their interpretation. We will avoid mathematical details in our presentation of the key issues, focusing on the meaning of the key issues and the impact that they may have.

First, there exist a large number of statistical methods for combining data across studies.[44] Even though these methods may seem to an outsider to be a source of potentially large diversity, empirical evidence suggests this is not a major concern.[45]

[43] For example, a large number of meta-analyses (at least 9 we are aware of) have been conducted trying to evaluate whether it is better to administer these drugs once a day or multiple times a day in terms of the potential to damage the kidneys and their efficacy for treating infections. The number of studies included in these meta-analyses has varied by more than 3-fold since the eligibility criteria of the different teams conducting the meta-analyses were different.

[44] For a brief, non-technical overview, see Joseph Lau et. al., *Quantitative Synthesis in Systematic Reviews*, 127 ANNALS INTERNAL MED. 820 (1997).

[45] *See, e.g.*, Jose Villar et. al., *Meta-analyses in Systematic Reviews of Randomized Controlled Trials in Perinatal Medicine: Comparison of Fixed and*

Moreover, as the field of meta-analysis has matured over time, analytical approaches have become more standardized. The first step usually is to evaluate whether there is any formal, statistically significant heterogeneity (diversity) between the results of the eligible studies included in the meta-analysis. The detection of formal heterogeneity does not mean that something is wrong with the data used in the meta-analysis, but it provides a hint that the results of the constituent studies differ between themselves. Between-study differences may exist either because of genuine diversity or because of bias, or in some cases chance may have played its role. Conversely, when the heterogeneity test is not significant, one cannot rule out completely that genuine diversity and/or bias may exist, especially when the number of studies is limited and the test has limited power to identify existing diversity.[46]

One commonly used approach in many meta-analyses is the estimation of an overall effect (average). The average is not obtained by simply summing up the data, since this may lead to paradoxical results and erroneous conclusions. Instead, the results of each study are given a weight inversely proportional to the uncertainty (variance) of the results, which is a function of the total number of study subjects and the number of events. Large studies with little uncertainty in their estimates are thus given more weight than smaller studies. In the presence of major differences in the results of the various studies, even this weighting approach may be inappropriate, since it assumes that all studies have different results due to chance alone, something that seems unlikely in this setting. Appropriate statistical models are available that also take into account the extent of diversity in the study results in generating a summary estimate.

In the presence of significant differences between studies, there

Random Effects Models, 20 STAT. MED. 3635 (2001).

[46] Empirical evidence and technical considerations on heterogeneity are covered by Julian Higgins et. al., *Statistical Heterogeneity in Systematic Reviews of Clinical Trials: A Cricial Appraisal of Guidelines and Practice*, 7 J. HEALTH SERVICES RES. & POL'Y 51 (2002); Eric Engels et. al., *Heterogeneity and Statistical Significant in Meta-analysis: An Empirical Study of 125 Meta-analyses*, 19 STATISTICS IN MED. 1707 (2000).

are several other methods that can be used in trying to explain and understand this heterogeneity. Meta-regression methods try to relate common characteristics with differences in the results across several studies. These methods are helpful when there are a large number of studies.[47] The analyses are mostly exploratory and generate results that usually would have to be validated in subsequent research. Moreover, some of these analyses are subject to the ecological fallacy (when inferences based on average characteristics of a group are extrapolated and applied to individuals who comprise the group).

Another approach is multivariate modeling, and this is feasible in meta-analyses of individual-level data. In this case, for each patient in each study, information is available on several different characteristics. One can then test whether the results are affected or modified in the presence of each of these characteristics or combinations thereof. These analyses are also exploratory, but they may help clarify why diversity exists in the results of various studies. At the end of the analysis, it may be possible to identify subgroups of people who are different in their responses to the same medical intervention or who have different magnitudes of benefit and/or risk in relationship to the same intervention.[48]

It has been debated whether it is better to have one single large study rather than a meta-analysis of several smaller studies. Large studies are useful, but for most medical questions of interest they are never performed. Therefore, the constellation of several small clinical studies may be all that is available. Even if large studies have been done, it is not certain that their results would be more reliable than the results of smaller studies. Most often, all studies, big and small, offer complementary evidence on a question of interest.

[47] For more details on meta-regressions and their relationship to other commonly used meta-analysis methods, *see* Sander Greenland, *Invited Commentary: A Critical Look at Some Popular Meta-analytic Methods*, 140 AM. J. EPIDEMIOLOGY 290 (1994).

[48] Thomas A. Trikalinos & John P.A. Ioannidis, *Predictive Modeling and Heterogeneity of Baseline Risk in Meta-analysis of Individual Patient Data*, 54 J. CLINICAL EPIDEMIOLOGY 245 (2001).

E. Other Considerations in Meta-analysis

1. The Relationship between Quality and Selection

As discussed earlier, the results of systematic reviews and meta-analyses may be affected by how comprehensive the criteria are for selecting the studies to be included. Sometimes criteria may be set on the basis of the perceived quality of studies; studies considered to be of poor quality may be excluded or studies of perceived poor quality may be contrasted against those of high quality. The quality and potential deficits thereof should be assessed in each of the studies considered for inclusion in a meta-analysis. However, quality may often be difficult to quantify, and the effect of poor quality on the results of a study may often be unpredictable. Some investigators have suggested that more weight should be given to the results of high-quality studies.[49] This is a problematic approach since the allocation of differential weight does not correct the quality deficits and it is unknown whether it leads the summary results closer to the truth. The detection of quality defects is important to give insight about how to improve future research in the field and to give a hint about the uncertainty that may accompany the research findings, especially when poor quality is documented. One may have to give less credibility to (or even discount) the results of poorly designed research, even if the conclusions seem to be strong and beyond dispute from a purely statistical perspective.

2. Publication Bias

Another issue to be considered is the possibility of publication bias. Publication bias reflects the fact that small studies with "negative" results may not be published because investigators, peer-reviewers, and/or editors don't find them as interesting as

[49] David Moher et. al., *Does Quality of Reports of Randomised Trials Affect Estimates of Intervention Efficacy Reported in Meta-analyses?*, 352 LANCET 609 (1998).

studies that find statistically significant results ("positive" studies).[50] The terms "positive" and "negative" are misnomers since well designed and conducted studies should be important sources of evidence regardless of what their results are. The selection of "positive" studies may nevertheless lead to spurious impressions when data are synthesized. Another possibility is time lag bias,[51] according to which studies with negative results may eventually be published, but they take longer to do so, as compared with those with "positive" results. Time lag bias will also result in more favorable estimates of effects when early data are synthesized and in diminishing effects as a more complete picture emerges over time.[52]

Several tests have been developed that try to detect publication bias. These tests basically examine whether small studies tend to give different results as compared with larger ones; or they try to investigate whether the results of a meta-analysis would change under the assumption that certain "negative" studies are imputed to have been lost and their putative results are added to the overall calculations.[53] Time lag bias may be examined by performing

[50] Publication bias has been documented both for randomized trials and for observational studies. *See, e.g.*, Phillipa J. Easterbrook et. al., *Publication Bias in Clinical Research*, 337 LANCET 867 (1991); Kay Dickerin et. al., *Factors Influencing Publication of Research Results: Follow-up of Applications Submitted on Two Institutional Review Boards*, 267 J. AM. MED. ASS'N 374 (1992).

[51] John P.A. Ioannidis, *Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials*, 279 J. AM. MED. ASS'N 281 (1998); Jerome M. Stern & R. John Simes, *Publication Bias: Evidence of Delayed Publication in a Cohort Study of Clinical Research Projects*, 315 BRIT. MED. J. 640 (1997).

[52] John P.A. Ioannidis et. al., *Recursive Cumulative Meta-analysis: A Diagnostic for the Evolution of Total Randomized Evidence from Group and Individual Patient Data*, 52 J. CLINICAL EPIDEMIOLOGY 281 (1999).

[53] Matthias Egger et. al., *Bias in Meta-analysis Detected by a Simple, Graphical Test*, 315 BRIT. MED. J. 629 (1997); Alex J. Sutton et. al., *Empirical Assessment of Effect of Publication Bias on Meta-analyses*, 320 BRIT. MED. J. 1574 (2000); Sue Duval & R. Tweedie, *Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-analysis*, 56 BIOMETRICS 455 (2000).

cumulative meta-analyses. One may notice whether the summary results tend to change in the same direction over time, in particular when there is an indication of continuously diminishing summary effects. These tests have their limitations. Genuine heterogeneity may also give the same picture as publication bias,[54] but all these tests may give signals that a simple grand mean approach may be missing important parts of the picture of the evidence.

CONCLUSIONS AND LEGAL IMPLICATIONS

Medical evidence is complex; it is heterogeneous and of variable quality. Interpreting medical evidence is difficult. Evidence must be appraised in its totality using robust systematic approaches. Quantitative methods are required to achieve a synthesis of the data across an increasing number of studies of the same topic. Evidence is cumulative and may change over time, as more data accumulate. Moreover, evidence, even when derived from well-conducted studies, is subject to biases that may stem from factors that are unrelated to the excellent performance of isolated single investigations. Meta-analysis offers a useful tool to summarize evidence across many studies, identify heterogeneity, search for possible explanations for the presence of this diversity, and offer hints about the possibility of existing bias. Even the best evidence and the best meta-analyses thereof are not sufficient for medical decision-making. Most medical decisions are complex and require a frame of interpretation. Physicians should try to use the best tools to justify their actions in everyday health care. Systematic reviews and meta-analyses as well as evidence-based recommendations that stem from them are one means for enhancing the certainty and confidence of physicians about their actions and can be used to justify medical practice. However, they also provide a measure of the uncertainty that lies behind these

[54] *See, e.g.*, Jonathan A. Sterne et. al., *Publication and Related Bias in Meta-analysis: Power of Statistical Tests and Prevalence in the Literature*, 53 J. CLINICAL EPIDEMIOLOGY 1119 (2000). Some of the most commonly used tests may sometimes be difficult to interpret. *See, e.g.*, J.L. Tang & J.L. Liu, *Misleading Funnel Plot for Detection of Bias in Meta-analysis*, 53 J. CLINICAL EPIDEMIOLOGY 477 (2000).

actions.

From a legal perspective, it is important that judges and lawyers become more familiar with the advent of these evidence-based tools. While judicial decisions may continue to seek expert testimonies, such testimonies should be increasingly based on solid scientific evidence rather than expert opinion. Typically, a judge or a lawyer may find it difficult to probe behind an expert's opinion to evaluate or test the level of credibility based on concepts of best available evidence. Experts in court are typically challenged and accepted on the basis of their qualifications, conflicts of interest, or biases, and not on the value of their opinions. However, it would be useful to understand that the medical expert is only one part of the long chain of the medical evidence, and that there often can be considerable uncertainty behind expert statements.