The Hierarchy of Evidence, Meta-Analysis, and Systematic Review

Robyn Bluhm

Published online on: 25 Oct 2016


Accessed on: 08 Aug 2023


Please scroll down for document

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
THE HIERARCHY OF EVIDENCE, META-ANALYSIS, AND SYSTEMATIC REVIEW

Robyn Bluhm

Contemporary medical practice is strongly influenced by the idea that clinical decision-making should be based on the results of clinical trials. A number of interrelated trends, stretching back several decades, have converged on a view of what counts as good medical research. This chapter begins with a brief history of three connected movements—clinical epidemiology, evidence-based medicine (EBM), and the Cochrane Collaboration—which aim to promote the use of high-quality evidence from clinical research in patient care. “High-quality evidence” is usually understood as evidence at the top of a “hierarchy of evidence,” specifically randomized controlled trials (see Chapter 18) and systematic reviews or meta-analyses of such studies. The remainder of the chapter will discuss these ideas in detail, examining both the arguments for and the criticisms of hierarchical rankings of study methods, and of the combination of research results in a meta-analysis.

Clinical epidemiology is the oldest of the trends toward using the results of clinical trials to inform clinical practice; Jeanne Daly (2005) traces it back to the late 1960s, to the work of Suzanne Fletcher and Robert Fletcher, and, beginning around the same time, to that of Alvan Feinstein. This turn to epidemiological methods as a way of gathering evidence to support clinical decision-making was motivated primarily by the perceived shortcomings of physiological research. In an interview with Daly, Robert Fletcher says:

We grew up in a very biomedical era. We were trained at Stanford in internship and residency, where everyone was a laboratory scientist, and it seemed to me that the kind of science that they brought to bear on patient care, which was mainly logical argument from laboratory data on mechanisms of disease, just wasn’t the best possible way of answering these clinical questions.

(Daly, 2005: 21)

As will become clear below, the contrast between laboratory-based physiological research and clinical trials that use epidemiological methods continued to be prominent in EBM.

According to Daly, “[e]vidence-based medicine was the form in which clinical epidemiology promoted its findings to practicing clinicians, who was its primary target” (2005: 4). The term “evidence-based medicine” was first used in an article in the journal *ACP Journal Club* in 1991 (Guyatt, 1991), but the “manifesto” introducing EBM to a broad clinical audience appeared
in *JAMA: The Journal of the American Medical Association* the following year (Evidence-Based Medicine Working Group, 1992). The authors of this second paper were members of the Evidence-Based Medicine Working Group, which was based at McMaster University in Hamilton, Ontario, Canada. In this paper, they contrast “the way of the past” in medicine with “the way of the future” by using a case study of a resident who is trying to find information for a patient who wants to know his risk of experiencing a seizure. The way of the past is based on the knowledge of medical authority figures, as the resident gets an answer to this question by asking her supervisors what to tell the patient. By contrast, the way of the future has her formulating her patient’s question in terms that reflect the kind of hypothesis that could be examined in a clinical study, searching the databases of the medical literature for relevant research, appraising the studies she finds to ensure their quality, and then reporting their results to the patient.

Shortly after this, members of the group published a textbook, *Evidence-Based Medicine: How to Practice and Teach EBM* (Sackett et al., 1997), and they began to present a series of articles in *JAMA* on various aspects of EBM; these articles were later published as *The Users’ Guides to the Medical Literature* (Guyatt and Rennie, 2002). These publications laid out the main tenets of, and skills required for, EBM. The central theoretical idea underlying EBM is that evidence from clinical research can be ranked hierarchically, with the study methods that are most likely to give good evidence at the top of the hierarchy (see the section below on the hierarchy of evidence). In addition, readers were taught how to critically appraise published research in order to determine its quality. For example, the *Users’ Guides* present the techniques of critical appraisal in terms of a series of questions about the methodological features of a study and the details of its results. EBM therefore aimed to teach clinicians how to obtain and assess epidemiological studies that addressed clinical questions relevant to their practice.

By contrast, the third movement that advocated using epidemiological study designs to inform clinical practice focuses on producing summaries of the literature. The Cochrane Collaboration, formed in the UK in 1993, describes itself as “a global independent network of researchers, professionals, patients, carers, and people interested in health” that aims to “gather and summarize the best evidence from research” (www.cochrane.org). One of the problems faced by clinicians who want to practice EBM is the sheer volume of clinical literature. The Cochrane Collaboration aims to solve that problem by gathering, appraising, and systematically reviewing studies relevant to a specific clinical question, and by updating the reviews periodically as new evidence becomes available. From its beginnings, the Cochrane Collaboration was closely allied with the McMaster University group’s EBM, as David Sackett (often described as the father of EBM) was also instrumental in the development of the Collaboration.

In summary, looking at clinical epidemiology, EBM, and the Cochrane Collaboration, we can see a convergence on the idea that clinical decision-making should be based on the results of research that examines patient outcomes in large groups, rather than on physiological research or on the advice of medical authorities.

### What Kind of Research?: The Hierarchy of Evidence

The central idea underlying EBM is that research can be ranked hierarchically based on study design: studies that use methods ranked higher on the hierarchy are more likely to be high quality and therefore to provide good evidence. Although there are different hierarchies for different kinds of research question (e.g., questions about the efficacy or effectiveness of a treatment, about the harms associated with an intervention, about patient prognosis, or about the accuracy of diagnostic tests), much of the discussion in the EBM literature has focused
THE HIERARCHY OF EVIDENCE

on research on therapies, and this kind of research will therefore be the primary focus of this chapter.

In addition, there are different versions of the hierarchy of evidence, although all of these have the same basic structure. Here, the hierarchy of evidence presented in the Users’ Guides to the Medical Literature will be used, as it is among the first and most influential of the proposed hierarchies, and is also therefore representative of the basic structure of hierarchies for treatment evidence. [More recent versions of the hierarchy, notably the one produced by the GRADE Working Group, also incorporate assessments of study quality, although the general ranking of study methods is unchanged (see www.gradeworkinggroup.org).] The hierarchy is as follows:

- Systematic reviews/meta-analyses of RCTs
- Single RCT
- Systematic reviews/meta-analyses of observational studies (cohort, case-control)
- Single observational study
- Physiologic studies
- Unsystematic clinical observations (Guyatt & Rennie, 2002: 7)

There are several important points to note regarding the hierarchy of evidence. First, all of the top levels of evidence use methods adopted from epidemiology; they compare outcomes in large numbers of people, one group of which receives the intervention being tested while the other group does not. Ideally, the groups will be similar (on average), except for differences in their exposure to the intervention being studied. This approach to research is contrasted with studies that aim to understand physiological mechanisms that underlie disease and that are supposed to be affected by therapeutic interventions. This kind of physiological research occupies the lowest level of the hierarchy, even though for most of the 20th century, it was the main kind of medical research. David Sackett, who was strongly influenced by Alvin Feinstein’s version of clinical epidemiology, argues for the importance of studies using epidemiological methods, saying that “in sharp contrast to most bench research, the results of clinical-practice research are immediately applicable” (Sackett, 2000: 380). In general, there is more variability toward the bottom of the hierarchy than at the top; some versions include case studies or case series reports, or expert consensus, albeit as relatively poor-quality evidence.

This quotation also draws attention to an additional reason for placing physiological research at the bottom of the evidence hierarchy, which is that EBM emphasizes the importance of using “patient-important,” clinically relevant outcomes, such as the occurrence of death, heart attack, or stroke, in clinical research. The alternative approach is to use physiological measurements as “surrogate outcomes” that are related to and predict the clinically relevant events. In many cases, EBM cautions, these surrogate outcomes are only weakly related to the clinical outcomes of real interest, and are therefore poor predictors. This aspect of EBM reflects the concern of the earliest clinical epidemiologists that physiological research is not directly relevant to clinical practice.

Another important thing to notice about the hierarchy of evidence is that it places randomized studies (either a single RCT, a systematic review, or meta-analysis of RCTs) above nonrandomized, “observational” studies. In some places, this ranking is treated as absolute, for example: “[i]f the study wasn’t randomized, we’d suggest that you stop reading it and go on to the next article in your search. . . . Only if you can’t find any randomized trials should you go back to it” (Straus et al., 2005: 118). In other places, for example, the GRADE Working Group’s system, it is acknowledged that a well-designed nonrandomized study can provide better evidence than a less well-designed randomized trial. Furthermore, in some cases, if the
evidence (from nonrandomized trials) supporting an intervention is particularly strong, EBM acknowledges that it may not be necessary to conduct RCTs at all. In general, though, the consensus is that randomization is such an important feature of study design that RCTs almost always provide better evidence than studies that do not randomly assign patients to the treatment or the control group.

What does randomization do that makes it so important? There are two main arguments for randomization. First, it is considered to be the best way to ensure that the treatment and the control groups really are similar. Similarity is important because if, for example, one group is significantly older (on average), or contains a greater proportion of women, than the other does, then it is not clear whether any observed outcome differences between the groups are due to the presence or absence of the intervention being studied or to the other (age, sex) differences between the groups. These other factors are said to confound the results of the study. In some cases, specific clinical or demographic factors may be known or suspected to affect the clinical outcomes of interest, but different unknown factors may also have such an influence. For example, an unmeasured genetic or physiological characteristic may be present in one study group more often than in the other, and that may influence either the progression of disease or the action of the treatment. Proponents of randomization say that it is the best way to balance both known and unknown confounders between study groups.

The second major argument for randomization is that if study researchers are deciding who gets assigned to the treatment or the control group, they may (whether consciously or unconsciously) bias the assignment. For example, a clinician may assign sicker patients to receive an experimental drug rather than a placebo because she wants to ensure that they receive an active treatment. Alternatively, a clinician who stands to benefit financially or professionally if a clinical trial shows that a new treatment is efficacious may put patients who are relatively healthy in the experimental group. (Again, this need not be a conscious, deliberate decision.) In both of these cases, the study groups end up being unbalanced with regard to disease status, which is likely to affect the study outcome. This kind of bias is known as allocation bias (because the allocation of patients to study groups biases the trial’s results).

Philosophical Criticisms of the Hierarchy of Evidence

It is important to recognize that the hierarchy of evidence is based on and defended by epistemological claims. It is therefore a philosophical theory of medical knowledge. Philosophers have examined and challenged both the implicit assumptions underlying the hierarchy and the explicit defenses supporting it. Most of this philosophical work has addressed the claims made regarding randomization and, in particular, have tended to conclude that the placement of randomized trials above nonrandomized studies is not entirely justified. More recently, there has also been discussion in the philosophical literature of the relative importance of epidemiological and physiological research.

With regard to the first issue, John Worrall’s 2002 paper “What Evidence in Evidence-Based Medicine” provides a clear analysis of the claims made in favor of placing randomized studies at the top of the hierarchy and concludes that this placement is not justified. With regard to the balancing of known confounders, he points out that it is just as feasible to deliberately balance the groups, for example, by assigning (nearly) equal numbers of women and of men to each of the treatment and the control groups. With regard to unknown confounders, which by definition cannot be deliberately balanced, Worrall says that randomization is not (despite some claims to the contrary) guaranteed to do so; what is actually correct is the much weaker claim that randomization controls for these factors “in some probabilistic sense” (2002: S322).
Worrall also addresses the argument that randomization prevents bias in the allocation of patients to the treatment versus the control groups by ensuring that study personnel do not deliberately assign patients to groups in such a way that confounding factors are not balanced across groups. He concedes the importance of guarding against this bias, but notes that random assignment of patients to study groups is merely a means of achieving this goal: “It is blinding (of the clinician) that does the real methodological work—randomization is simply a means of achieving this end” (S325).

A number of other authors have addressed EBM’s claims about the importance of randomization. For example, Grossman and Mackenzie (2005) survey some of the limitations of RCTs and discuss their implications for research. In particular, they worry that the emphasis on randomization will prevent people from conducting, and from appreciating the contributions that can be made by, research that uses other methods. Borgerson (2009) argues that the claim that RCTs are the only type of study that can avoid (or at least the study that best avoids) certain kinds of bias obscures the fact that RCTs are still subject to many other kinds of bias.

More recently, there has been philosophical discussion about the second key feature of the hierarchy, which is the relationship between the epidemiological studies that occupy the top few layers of the hierarchy and laboratory-based research on physiological mechanisms, which is located near the bottom of the hierarchy. One key question in this discussion is whether evidence (or, more generally, reasoning) about mechanisms can ever substitute, as a basis for treatment decisions, for epidemiological research. The EBM position appears to be that, as a last resort, this research can be used: while the Users’ Guides say that the hierarchy of evidence “implies a clear course of action for physicians addressing patient problems: they should look for the highest available evidence from the hierarchy” (Guyatt and Rennie, 2002: 8), this implies that if no randomized (or nonrandomized) controlled trials are available, then evidence from physiological studies should be used.

A deeper question is whether this kind of research can ever provide strong enough evidence to establish the efficacy or effectiveness of a treatment, so that it is not necessary to conduct epidemiological studies. The idea here is that the physiological mechanism by which the treatment works is so well-understood that no further research is needed. The suggestion that RCTs are always necessary to establish whether a treatment works has been spoofed by a study that purported to review the RCT evidence that parachutes reduce mortality associated with “gravitational challenge.” The authors noted that they were unable to find high-quality RCT evidence in favor of parachute use (Smith and Pell, 2003). More seriously, people who claim that evidence from physiological mechanisms may be sufficient have observed that nobody claims that it is necessary to do a placebo-controlled RCT to determine whether people with dehydration should be treated with water; our understanding of the relevant physiology is enough to justify using this therapy.

Jeremy Howick (2011) has argued that “mechanistic reasoning” can provide sufficient evidence for the use of a treatment when the following conditions are met: knowledge of the relevant mechanisms must be complete (so that there are no “gaps” between the intervention and the outcome), and both the probabilistic nature and the complexity of mechanisms must be recognized (including the way that their functioning might be influenced by other mechanisms, which may result in “undesirable side-effects or even paradoxical effects” (2011: 144). Howick acknowledges, particularly in his more recent work (Howick et al., 2013), that cases in which knowledge of physiological mechanisms is sufficient to justify the use of a treatment is rare, but he does not want to rule out the possibility that these situations may sometimes occur.

Holly Andersen (2012) has argued that even Howick’s guarded optimism about the sufficiency of mechanistic evidence is not justified. She points out that the physiological mechanisms, and
the functional relationships among them, are so complex that it is unlikely that we can ever be justified in claiming that Howick’s conditions are met. Yet she does (in agreement with the Users’ Guides) say that there is a second—and legitimate—role that knowledge of mechanisms can play in clinical reasoning. This is the use of knowledge of mechanisms to bridge the gaps between evidence from clinical trials and the circumstances and characteristics of particular patients. Bluhm (2013) has argued, however, that situations in which this kind of reasoning about specific cases is justified are probably almost as rare as ones in which Howick’s criteria for reasoning about general claims are met. Similarly, Howick et al. (2013) argue that knowledge of mechanisms cannot justify inferring from clinical trials how a treatment will work in a population that is different from the one in which the study results were obtained, ultimately because of the complexity of physiological mechanisms.

Another important discussion relevant to the relationship between epidemiological and physiological research centers on a position that has come to be known as the Russo-Williamson thesis. Federica Russo and Jon Williamson (2007) have argued that establishing causal claims in medicine (including claims about whether a treatment will have the desired outcomes) requires both population-level, epidemiological research and research on physiological mechanisms. This is because the kind of causal evidence coming from each type of research is different: epidemiological research provides statistical evidence of a correlation between an intervention and an outcome, while physiological research provides mechanistic evidence. According to Russo and Williamson, causal claims in biomedicine rely on both kinds of evidence; they develop the “epistemic theory of causality” in order to do justice to the dual nature of causal claims in medicine. They and their colleagues also note that EBM has failed to do justice to the importance of the evidence provided by mechanisms and that its reliance on correlational, epidemiological research is problematic for several reasons (Clarke et al., 2013). First, correlation on its own cannot establish causation, because of the possibility of confounding variables. Notice that this point can serve as a rejoinder to those who claim that RCTs provide the best method for eliminating the effects of confounding variables by balancing those variables between the treatment and the control group. Second, knowledge of mechanisms is, in a sense, presupposed by epidemiological studies, since they influence the design of clinical trials. Finally, information about mechanisms is necessary to allow researchers to make claims that go beyond the results of a specific RCT, including applying those results outside of the context of the trial. This last point shows that the preceding discussion of Howick’s and Andersen’s work on generalizing from mechanisms is related to the more general claims about causality in medicine and medical research that are the focus of Russo’s and Williamson’s work.

Given these problems with the main points raised by the hierarchy of evidence, it has been suggested that it might be worth getting rid of the idea of a hierarchy. Goldenberg (2009), for example, argues that because the hierarchy is problematic, and has therefore been the focus of most criticisms of EBM, attention has been diverted away from the important contributions that EBM can make to improving medical care. Bluhm (2005) suggests that a better metaphor for the relationship among different kinds of research evidence is a network; this will better allow us to focus on the relationship between epidemiological and physiological research. Most recently, Stegenga (2014) has examined more recent versions of the hierarchy, such as GRADE, and argued that they fail to overcome the criticisms raised against the earlier hierarchies.

Meta-Analysis and Systematic Reviews

The previous section surveyed issues related to the placement of different kinds of study (randomized, nonrandomized, physiological) on the hierarchy. This section examines questions related to another central feature of the hierarchy, which is the importance of systematic
reviews and meta-analyses. A systematic review is “a high-level overview of primary research on a particular research question that tries to identify, select, synthesize and appraise all high quality research evidence relevant to that question” (www.cochrane.org), whereas a meta-analysis is a form of systematic review that combines the results of the included studies in a statistical analysis. Recall that hierarchy is structured so that systematic reviews or meta-analyses of RCTs rank higher than a single RCT, and then similarly for nonrandomized studies. This ranking is based on the (very reasonable) idea that more evidence is better. Note that, even here, things are not straightforward; the hierarchy implies that randomized and nonrandomized trials cannot (or perhaps should not) be combined in a single systematic review or meta-analysis. And beyond this starting point, things rapidly become more complicated; it is not clear how best to amalgamate the available evidence and draw conclusions about what it says. To see why this is the case, we can look at some of the decisions made in the course of conducting a meta-analysis.

Perhaps the most famous example of a meta-analysis in medicine is the one represented graphically in the logo of the Cochrane Collaboration. The study reviewed the results of placebo-controlled RCTs that examined whether giving a short course of corticosteroids to women who were about to give birth early improved the chances of their babies’ survival. The first of these RCTs was published in 1972, and ten years later there was sufficient evidence to establish that the odds of infant mortality due to the complications of premature birth were reduced by 30–50%. But, the Cochrane website explains, the first meta-analysis on this topic was not published until 1989. Prior to this, because the evidence was located in a number of studies, many of which were too small to have the power to detect any effect, the studies did not affect practice and, as a result, “tens of thousands of premature babies have probably suffered and died unnecessarily (and needed more expensive treatment than was necessary)” (www.cochrane.org).

This story illustrates powerfully the important contribution that can be made by meta-analysis; however, it is not clear that this particular meta-analysis is representative. The intervention in these studies (corticosteroid therapy) was simple and short term. The outcome measured (neonatal death) is clear and easy to measure, and was not tracked over the long term. The results (when combined) are also dramatic: the improvement in survival rates with therapy is large.

By contrast, in many other cases, the studies that have been conducted on a particular intervention may be highly variable, both in the details of their design and in their quality. RCTs may differ in a number of ways, including the details of the intervention (e.g., dosage, duration of treatment, use of concomitant therapies), the outcomes measured (and when and how they are measured), and the population included in the study (e.g., age, presence of comorbid conditions). Those who conduct a meta-analysis must decide how similar studies must be with regard to these features before they can be combined. They may even decide, if the studies are different enough, to conduct a systematic review without a meta-analysis and present information for all of the studies separately. In cases where there is a large number of studies, this approach will soon become unwieldy; moreover, the fact that meta-analyses are generally conducted if possible suggests that they are the preferred method (though the hierarchy of evidence does not explicitly rate them above systematic reviews that do not conduct a meta-analysis).

Similarly, decisions must be made about when to include studies on the basis of their quality. One option is to set a threshold, with studies falling below this quality cut-off being excluded from the meta-analysis. Another is to assign different weights to the study results being included, so that higher-quality studies make a bigger contribution to the summarized results than do lower-quality studies. Because of the variability of studies pertaining to a clinical question, part of the process of conducting a meta-analysis is deciding which studies to
include. Different groups conducting meta-analyses on the same treatment may make these decisions differently. Stegenga (2011) has pointed out that one of the advantages claimed for meta-analysis is that it is an objective way to assess evidence, but that because these different choices are possible, this claim is false.

A review by Deshauer et al. (2008), on the use of selective serotonin reuptake inhibitors (SSRIs) to treat depression, illustrates the kinds of decisions researchers must make in conducting a review. This particular paper both reports the methods and results of individual studies and conducts several analyses that combine the results of these studies for different treatment outcomes (e.g., response to treatment, overall acceptability of the treatment). Only six RCTs met the criteria for inclusion in the review. When the results of all of the studies were combined, they showed a statistically significant response to treatment, but a closer look at the individual studies suggests that combining them may obscure important information. Half of the trials included only patients who had major depression with no comorbid conditions; combining only these studies showed a statistically significant treatment response (and a larger effect size than the analysis that included all six of the studies). The other half did include patients with comorbidities, and when these studies were combined, there was no statistically significant improvement. This means, though, that clinicians making decisions about the treatment of patients with comorbidities on the basis of the meta-analysis of all six studies may overestimate the effect of SSRIs for their patients.

The presentation of the details of each of the six studies also shows clearly that they varied with regard to the treatment dose and duration, and even the drug being tested; although most of the studies examined sertraline, others tested paroxetine or citalopram. The studies also differed in how they defined “response to treatment” and in the eligibility criteria for the study. These differences among the studies do not mean that the authors of the review were wrong to combine them, but different choices—ones that might have resulted in rather different results—could also be defended. Deshauer et al. also chose to exclude a number of studies, most notably any that lasted less than 6 months and any that used a “discontinuation” study design (which begins by treating all participants with the study drug and then assigning them to either continued treatment or placebo). Again, had these studies been included, the conclusion regarding the effectiveness of SSRIs may well have been different.

**Conclusion**

The trend toward basing clinical decision-making on the results of clinical research raises interesting philosophical questions. Although much has been written by philosophers on EBM and, in particular, the hierarchy of evidence, most of the discussion of the merits of meta-analyses has taken place in the medical literature.

**References**


Cochrane collaboration website: www.cochrane.org
GRADE Working Group: www.gradeworkinggroup.org

Suggested Readings