Valuing Evidence: Bias and the Evidence Hierarchy of Evidence-Based Medicine

Kirstin Borgerson

Perspectives in Biology and Medicine, Volume 52, Number 2, Spring 2009, pp. 218-233 (Article)

Published by Johns Hopkins University Press
DOI: 10.1353/pbm.0.0086

VALUING EVIDENCE

bias and the evidence hierarchy of evidence-based medicine

KIRSTIN BORGERSON

ABSTRACT Proponents of evidence-based medicine (EBM) suggest that a hierarchy of evidence is needed to guide medical research and practice. Given a variety of possible evidence hierarchies, however, the particular version offered by EBM needs to be justified. This article argues that two familiar justifications offered for the EBM hierarchy of evidence—that the hierarchy provides special access to causes, and that evidence derived from research methods ranked higher on the hierarchy is less biased than evidence ranked lower—both fail, and that this indicates that we are not epistemically justified in using the EBM hierarchy of evidence as a guide to medical research and practice. Following this critique, the article considers the extent to which biases influence medical research and whether meta-analyses might rescue research from the influence of bias. The article concludes with a discussion of the nature and role of biases in medical research and suggests that medical researchers should pay closer attention to social mechanisms for managing pervasive biases.

THE IDEA OF HIERARCHICALLY ranking research methods is not at all intuitive, nor is such ranking widely practiced by scientists. Biologists, astronomers, and chemists would likely be intrigued to learn that certain research methods in medicine are thought to be categorically better than others. Upon

Department of Philosophy, Dalhousie University, Halifax, NS, B3H 4P9, Canada.
E-mail: kirstin.borgerson@dal.ca.

The author is grateful to the Killam Trusts for fellowship support during the preparation of this manuscript and to colleagues at the University of Toronto and Dalhousie University for their ongoing support.

Perspectives in Biology and Medicine, volume 52, number 2 (spring 2009):218–33

learning of the evidence hierarchy of evidence-based medicine (EBM), these fellow scientists might ask how it is that medical scientists ranked different research methods against each other. Perhaps they would be interested to know how they might go about recreating such a hierarchy in their own field. The surprising answer to this query is that there are very few explicit justifications offered for the EBM hierarchy of evidence. This is true despite the widespread influence of EBM in health-care settings worldwide and the vast number of articles and books on the subject.

In what follows, I discuss two implicit justifications offered for the evidence hierarchy of EBM. One of these justifications—that the hierarchy ranks research methods according to their ability to identify causal relationships between treatments and effects—has been soundly critiqued in the medical and philosophical literature. The other justification for the evidence hierarchy is that it ranks research methods according to their ability to secure less biased results. Even the most vocal critics of the hierarchy concede that certain research methods may be ranked categorically above others according to their ability to minimize bias. However, my analysis reveals that this second justification is as flawed as the first, and thus that there are no epistemic justifications for the hierarchical ranking of research methods advanced by EBM.

EVIDENCE-BASED MEDICINE (EBM)

EBM requires that physicians integrate the best available clinical research evidence into decisions made in the clinical care of individual patients (Sackett et al. 1996). At the core of the EBM movement is the evidence hierarchy, which was designed to reflect the methodological strength of scientific studies. It is assumed that higher-ranked evidence on this scale is better than lower-ranked evidence, and that such evidence provides greater justification for clinical action. The Oxford Centre for Evidence-Based Medicine (2001) offers the most well-established version of the hierarchy for medical therapy. It places systematic reviews of randomized controlled trials (RCTs) and individual RCTs above cohort studies, which are in turn ranked above case-control studies and case series, and all of these methods are positioned above expert opinion and bench research.

It is important to be clear about the nature of this hierarchical ranking. EBM advocates are not just claiming that it is helpful to be able to distinguish, for instance, good from bad RCTs or better from worse cohort studies. They have made an assumption about the necessity of ranking these methods against one another so that a critical review of the literature will produce one, hopefully decisive, answer. The desire for a decisive answer is understandable in the medical context, where decisions are morally weighty (quite often matters of life and death) and there is an overload of conflicting information arising from medical research. Even if we acknowledge the difficult nature of this situation, however,

this does not mean that a hierarchy of research methods—or this hierarchy, in particular—is the right solution.

Proponents of EBM assume that a hierarchy of evidence is needed to guide medical research and practice. However, the particular evidence hierarchy advanced by EBM is only one of many possible hierarchies. If, for example, complexity of methods and individuality or specificity of results were thought most indicative of high-quality evidence in medicine, the evidence hierarchy might have been inverted. No one has seriously advocated for an inverted hierarchy, but the point is that no one has seriously (that is, explicitly and methodically) argued for *any* particular hierarchy. Hierarchies are more often asserted than argued for. In fact, recent developments in medicine have led to a proliferation of different evidence hierarchies, though they tend to follow the same basic principles of organization as the original hierarchy adopted by the EBM Working Group (Upshur 2003).

In light of the variety of possible and actual evidence hierarchies, the particular version offered by EBM needs to be justified. Advocates of EBM have not been forthcoming on this issue.[1] Because of this, I have attempted to reconstruct the most plausible justifications for the hierarchy. In order to do so, I have drawn upon a number of classic papers on EBM as well as more recent articles and guidebooks (EBMWG 1992; Guyatt and Rennie 2001; Sackett et al. 1996; Straus et al. 2005).

JUSTIFICATIONS FOR THE EVIDENCE HIERARCHY

According to my analysis, the evidence hierarchy ranks research methods according to two interrelated criteria. Evidence produced by methods at the top is thought to isolate causal relationships and to minimize bias. While I have attempted to distinguish these two arguments, they do share some common assumptions. Confounding factors, for instance, are part of the problem for those

[1]The justifications offered for the hierarchy are either unsupported or vague assertions. For instance, an authoritative statement on the hierarchy offers this attempt at justification: "mightn't a high-quality cohort study be as good as, or even better than, an RCT for determining treatment benefit? Some methodologists have vigorously adopted this view. I disagree with them, for two reasons. First there are abundant examples of the harm done when clinicians treat patients on the basis of cohort studies. . . . My second justification is an unprovable act of faith. It professes that the gold standard for determining the effectiveness of any health intervention is a high-quality systematic review of all relevant, high-quality RCTs" (Haynes et al. 2006, p. 177). The first of these justifications seems to radically misunderstand the nature of research (surely any research results can turn out to be wrong, regardless of method), and the second is actually an attempt to evade justification. The authors of the *Users' Guide to the Medical Literature* do slightly better, suggesting the hierarchy organizes research methods according to those that are more "systematic" and "unbiased" (Guyatt and Rennie 2001). "Systematic" is left undefined and could mean anything at all. The claim about bias is largely unexplained, but I attempt to come to terms with it as a possible justification.

who want to isolate cause-effect relationships and for those who identify confounding factors as biases. I shall draw attention to these points of overlap as they arise; for reasons that will become clear, however, it is important that the two claims are kept as distinct as possible. I shall offer a brief overview of the first justification, and the arguments made against it in the medical and philosophical literature, before moving to a detailed analysis of the second.

X Causes Y: Isolating Causal Relationships

One of the principal divisions in the evidence hierarchy is that between randomized and nonrandomized trials. If a trial is randomized, it is ranked near the top of the hierarchy; if not, it is ranked lower. There are plenty of strong statements on the epistemic powers of randomization in the EBM literature. For instance: "If the study wasn't randomized, we'd suggest that you stop reading it and go on to the next article in your search. (Note: We can begin to rapidly critically appraise articles by scanning the abstract to determine if the study is randomized; if it isn't, we can bin it.) Only if you can't find any randomized trials should you go back to it" (Straus et al. 2005, p. 118).[2] This chatty advice appears in the 2005 edition of an official EBM handbook. This recent statement exposes the dependence of EBM on evidence hierarchies and challenges the popular view that EBM has evolved beyond its early tendencies to discredit nonrandomized trials. EBM does tend to privilege RCTs, and advocates do tell physicians to ignore other sources of evidence when RCTs are available. In addition, a careful examination of the guidelines used in the evaluation of research evidence indicates a persistent tendency to set aside all studies that are not RCTs, despite claims to the contrary (Grossman and MacKenzie 2005). The question for epistemologists and epidemiologists alike is: does randomization confer the epistemic benefits claimed?

Extensive critiques of the overblown claims made on behalf of RCTs in the medical and statistical literature have done away with circular arguments regarding the overestimation of effects in nonrandomized trials (since a difference in effect—were it to be present—might just as easily imply an underestimation of effects in RCTs) and have revealed the confused reasoning beneath claims that nonrandomized trials are "misleading" (as if randomized trials could never be misleading!; Grossman, and MacKenzie 2005; Worrall 2002). The claims about causation, however, have been more persistent. Only randomized trials are thought to be capable of establishing genuinely causal relationships between treatments and effects; studies lower on the hierarchy get at "mere correlation."

[2]Consider also: "we owe it to our patients to minimize our application of useless and harmful therapy by basing our treatments, wherever possible, on the results of proper randomized controlled trials" (Sackett et al. 1991, p. 195), and "To ensure that, at least on your first pass, you identify only the highest quality studies, you include the methodological term 'randomized controlled trial (PT)' (PT stands for publication type)" (Guyatt, Sackett, and Cook 1994, p. 59).

As causation is a complex concept, it is important to be clear about what is meant by a "cause" in this context. Two types of causes are common to discussions in medicine: mechanistic causes and probabilistic causes. Mechanistic causes are provided by bench research in biochemistry, genetics, physiology, and other basic sciences, and are thought to be especially stable because they hold in all cases (not just selected subpopulations, however carefully or randomly selected). Probabilistic causes establish strength of association between dependent and independent variables in a given population, ideally in repeated studies (Russo and Williamson 2007). These causes are often identified through epidemiological research.

While mechanistic and probabilistic causes might intuitively be thought of as complementary ways of understanding the empirical world, the evidence hierarchy identifies probabilistic causes as epistemically superior. Claims about the special ability of RCTs to isolate causes refer to probabilistic causes and downplay the possibility that mechanistic causes could be just as well established, just as epistemically strong, and just as useful in medical practice. Consider Bradford-Hill's (1965) nine criteria for causation: strength of association, temporality, consistency, theoretical plausibility, coherence, specificity, dose-response relationship, experimental evidence, and analogy. Of these criteria, several explicitly relate to mechanisms: temporality, theoretical plausibility, coherence, and experimental evidence all rely on a characterization of a cause as a mechanism of some sort (Russo and Williamson 2007). Many of the remaining criteria relate to probabilistic causes. It is unclear why some of these criteria (those that are probabilistic) have been elevated within EBM while others (those that are mechanistic) have not. In addition to the neglect of other types of causes, the assumption that RCTs uniquely isolate probabilistic causes runs into its own problems.

Philosopher of science John Worrall has examined the most prevalent argument in favor of randomization: only randomized trials can balance treatment and control groups on all known and unknown confounding factors, and thus only randomized trials can isolate cause-effect relationships. Randomized trials are said to eliminate possible alternative hypotheses, permitting reasoning by eliminative induction. This claim goes back to Fisher (1947), who writes that the significance test can be "guaranteed against corruption" by the use of randomization (p. 19). But as Worrall (2002) points out, this is far too strong a claim, and Fisher and other statisticians who have made similar claims must have been aware of this. The treatment and control groups can at best be balanced for all factors only "in some probabilistic sense"; thus, the defenders of randomized trials temper their claims with statements like "as balanced as possible," and they refer to the "tendency" for balance rather than any guarantee (p. S322). More specifically, randomizers argue that it is improbable that the two groups are imbalanced with respect to *any one* particular unknown confounder. However, as Worrall points out: "Even if there is only a small probability that an individual factor is unbalanced, given that there are indefinitely many possible confounding factors, then

it would seem to follow that the probability that there is some factor on which the two groups are unbalanced . . . might for all anyone knows be high" (p. S324).

In order to begin to address this problem of confounding factors, the randomization would have to be repeated an indefinite number of times. But in RCTs, randomization is usually done only once. Thus, defenders of the special causal ability of RCTs make claims about the epistemic powers of *actual* RCTs based on what would happen in *ideal* RCTs. (The presence of the phrase "in the long run" betrays the slide to theoretical claims.) If we were to randomize forever, the limiting-average effect of the treatment would yield information of the sort desired by RCT enthusiasts. However, even on the infrequent occasion when an RCT is repeated, it is done on different subjects, in a different context—it is not, strictly speaking, replicated. And, unfortunately, "there is no reason to think that any actual randomized trial gives the same results as would be got from the 'limiting-average'" (Worrall 2007, p. 465). Because of the number of variables at play, it is more likely that, were a trial to be run many times, each set of results would be slightly different. So while we might be justified in making claims about the causal powers of randomization in the long run, in the short run (which is all we have) those powers are greatly diminished. It is not just that it is logically possible for RCTs to fail to establish causation (we already knew that based on the number of conflicting RCTs), it is that we never know how close they have come to doing so. This is not significantly different from the sorts of claims that can be made about the results of, for instance, well-conducted historical trials. There is no special access to causes gained only through the use of RCTs.

As a result, then, randomization does not create the conditions for justified reasoning by eliminative induction: "The premise that the experimental groups were *probably* balanced does not imply that the differences that arise in the clinical trial were *probably* due to the experimental treatment" (Howson and Urbach 2006, p. 197). If the two groups are only probably balanced, it is no longer possible to claim that we are reasoning by eliminative induction, because we have not *eliminated* the possible options, but only made them less likely. This does not mean that randomization is entirely ineffective—as I just noted, it still makes it *less likely* that confounding factors are at play, and this has some epistemic value. But this value is much more limited than generally recognized, and it does not provide a basis for ranking randomized methods categorically above carefully matched or historically controlled trials, since there is no special guarantee that one has isolated causes simply because of randomization.

Claims that RCTs isolate causes, while other methods identify merely correlations, have resulted in undefined and undefended accounts of causation that unfairly denigrate mechanistic causes, depend on problematic arguments about the ability of randomization to balance groups on known and unknown factors, and rely on characterizations of ideal RCTs (such as the indefinite repetition of the trial) that are never attainable in practice. All research methods that make use of probabilistic methods of analysis have some ability to get at probabilistic

causes. It may be that, in cases where an RCT is the best method for a particular question, it is especially good at narrowing down the possible causes, but this does not mean that RCTs have a unique capacity to identify causal relationships. And were we to have good reason, perhaps based on bench research, to believe we had a proper account of the mechanisms for a particular treatment, there is no reason to think that the lowly case study (ranked at the bottom of the hierarchy) couldn't do just as good a job at establishing causation on Hill's criteria. The hierarchy is not justifiably ranked according to the special causal abilities of particular research methods.

Objective Results: The Ability to Minimize Bias

I shall now turn to a justification for the hierarchy that has received less attention in the critical literature: the claim that it ranks research methods according to their ability to produce less biased results. The EBM Working Group (1992) writes about the systematic attempts to record observations in an "unbiased" fashion as one of the key features distinguishing clinical research from clinical practice. According to the Canadian Task Force on Preventive Health Care (2008), which produced the first formalized version of the hierarchy, the evidence hierarchy is designed to "place greatest weight on the features of study design and analysis that tend to eliminate or minimize biased results." In Richard Ashcroft's (2004) words, the evidence hierarchy rests on the notion "that it is possible to rank methods of inquiry by their susceptibility to bias" (p. 131). Of all the available methods that deal in direct empirical evidence, the RCT is thought to be least subject to bias. Against this popular position, I argue that research methods ranked highest in the hierarchy provide no greater guarantee that biases have been minimized than those below.

In statistical terminology, *bias* is "a systematic distortion of an expected statistical result due to a factor not allowed for in its derivation; also, a tendency to produce such distortion" (*OED*). One of the tasks of research methods is to minimize bias. The value placed on RCTs is most evident in the sharp line drawn between the RCT and the lower-ranked cohort study. All versions of the hierarchy maintain a categorical placement of RCTs above cohort studies. An examination of this ranking offers a clue to the problems in the hierarchy at all levels. Given that cohort studies are also controlled trials (they have treatment and control groups), can be double-blinded (though this depends on the type of intervention, as it does for RCTs), can be analyzed under the intention-to-treat protocol, and have an identical causal inferential structure (eliminative induction), the only feature distinctive of RCTs is the random allocation of participants to the two groups. Yet RCTs are thought to be less biased than other research methods. The superiority of RCTs is usually illustrated with reference to two forms of bias: selection bias and ascertainment bias (Jadad and Enkin 2007). Only RCTs, the claim goes, can control for these kinds of bias. As a result, RCTs produce less biased results than other methods.

To begin, consider selection bias, which the authoritative CONSORT (Consolidation of the Standards of Reporting Trials; 2008) statement defines as: "systematic error in creating intervention groups, such that they differ with respect to prognosis. That is, the groups differ in measured or unmeasured baseline characteristics because of the way participants were selected or assigned." This form of bias can occur when selecting participants for a trial from the general public. In the early days of clinical research, before randomization was popularized, medical researchers attempted to achieve balanced treatment and control groups by alternating the allocation of patients to the two groups as they were enrolled into the trial. The problem with this was that physicians modified their behavior depending on whether the next patient was to be enrolled into one group or the other. Physicians would, on occasion, refrain from inviting patients into a trial when they knew the next participant would receive placebo, or would purposely enroll patients who were more likely to do well on the treatment into one or the other group depending on what they hoped to establish with the results of the trial.

To deal with selection bias (at least in these types of trials), researchers must institute some form of *allocation concealment*. Allocation concealment, as it turns out, is secured independently of randomization. In fact, a study can be randomized and yet be without allocation concealment; this was of great concern to the proponents of EBM, who pushed for explicit statements about allocation concealment in published studies. And a nonrandomized cohort study can have concealed allocation; it is just a matter of keeping the allocation criteria—whatever they may be—from the physicians doing the intake. So, for instance, the allocation may be according to the patient's day of birth (odds in one group, evens in the other). As long as researchers do not know that this is the allocation criterion, selection bias can be managed. Furthermore, selection bias does not plague all research endeavors. Other research designs, such as case studies and qualitative research (in-depth interviews, for instance), do not face concerns about selection bias because they do not divide patients into two groups. The ability to manage selection bias, even if it were to be a characteristic of only some research methods, would not be the end of the discussion about relative bias.

Before discussing ascertainment bias, it is worth noting the potential for controversy on this last point. There is a certain amount of confusion in the medical and epidemiological literature on the sources of selection bias. While the most common definitions (such as the CONSORT definition offered above) focus on the bias introduced by researchers selecting patients for a trial, in some cases the definition is apparently meant to be more expansive: the term is used to cover cases in which patients self-select into one group through their personal behavioral choices. So, for instance, in a trial investigating the difference between smokers and nonsmokers with respect to some particular health outcome, it is the patients who have, in effect, chosen their trial group (by choosing to smoke or not to smoke). When selection bias is used in this very broad sense to include

not only physician-introduced selection bias but also patient-introduced selection bias, it is fair to say that *some* cohort studies will be less able to control for this type of selection bias. These will be the "observational" cohort studies in which patients select, rather than are assigned to, treatment or control groups. These studies can be contrasted with "interventional" (or "experimental") cohort studies in which patients are put into groups by researchers.

The evidence hierarchy does not distinguish between different types of cohort studies (interventional vs. observational), and so it is unlikely that this expansive definition of selection bias has been used in its construction. If, however, we imagine that it has been used in this way, we see pretty quickly why this will not save the hierarchy from the arguments of this section. Patient-introduced selection bias is controlled for either by designing a trial to be interventional and instituting allocation concealment or by carefully matching the two groups and establishing that there is no reason to suspect confounders. Noninterventional research methods can still control for patient-introduced selection bias in some cases. In trials on neonates, for instance, researchers have no reason to suspect the different "lifestyle choices" of the neonates will confound the trial, so they may be just as confident about the match between two groups of neonates in a retrospective observational study as they would be an a prospective interventional study (Worrall 2007). Even with a broader definition of selection bias, then, the priority given to interventional over observational cohort studies is a bit hasty.

One final point: the more expansive definition of selection bias seems to me to be particularly unhelpful, since it lumps together sources of bias that can and should be distinguished and makes it more difficult for researchers to recognize the value—and also the limitations—of allocation concealment. It also invites a slide from bias arguments to causal arguments. The concern about patient-introduced selection bias is that it injects a possible confounder. Confounders interfere with our ability to isolate cause-effect relationships. This takes us back to the causal argument outlined (and critiqued) above. But concerns about bias are not just concerns about confounding factors, or we would not be able to make sense of, say, research design bias or publication bias. Thus, this confusion over selection bias is instructive, in that it reminds us of the level of confusion within the medical research community generally about the nature and sources of bias in research. I shall say more about these general confusions below.

Returning to the possible reasons why the RCT might be less biased, let us now consider ascertainment bias. Ascertainment bias is defined by the CONSORT statement as the "systematic distortion of the results of a randomized trial as a result of knowledge of the group assignment by the person assessing [the] outcome, whether an investigator or the participant themselves." Ascertainment bias arises in the patient reports and analyses of the trial as it nears completion. If either the patient or physician is aware of the group the patient ended up in, this may lead to the reporting of more positive, or more negative, results. For

instance, a patient may overstate his or her improvement in order to gain praise from the physician, or the physician may ask fewer questions or adopt a more detached attitude in order to get more subdued reports from patients in the placebo group. As with all biases, these may be conscious or unconscious. The mechanism for addressing such bias is blinding: keeping study participants, and those charged with their care, unaware of their assigned group. Note that it is blinding, not randomization, that is important here. And blinding is not unique to RCTs, nor even always possible. It is possible to have blinded cohort studies; conversely, interventions that cannot be blinded (such as many lifestyle interventions) may be evaluated in unblinded RCTs. Ascertainment bias is not uniquely controlled for in RCTs, and it does not justify the categorical placement of RCTs above other study designs in the evidence hierarchy.

To argue against my position, advocates of the evidence hierarchy first would have to find a type of bias that has the potential to affect all clinical research trials ranked in the hierarchy. Then they would have to demonstrate that this type of bias is either uniquely controlled for in RCTs, or that the magnitude of this form of bias is consistently smaller for RCTs than for other research methods. The first condition is crucial, since even if RCTs did manage to control for one or two biases that no other trial could address, if that bias was not faced by other research methods then the achievement would not necessarily be grounds for preferential ranking. While there may be unique forms of bias faced only by case studies, which only case studies can address, this does not necessarily mean that case studies are more objective than all other research methods. It is not meaningful to suggest that the results of such trials are less biased simply because the trials have conquered or greatly diminished the possibility of one or two particular biases.

RCTs are widely thought to be less biased than other trial designs. But the (causal) inferential structure of the RCT is almost identical to the cohort study, even though cohort studies are consistently ranked below RCTs in various versions of the EBM hierarchy. Furthermore, the one or two biases that RCTs allegedly eliminate are either equally well managed by other methods (because they are not necessarily connected to randomization), or they are not necessarily encountered by other methods. As such, the claim that RCTs, by design, produce results that are necessarily less biased than other trials is false.

BIAS IN THE BIG PICTURE

A recent edition of a well-known guide to randomized controlled trials offers a contemporary catalog of the types of biases that can influence medical research (Jadad and Enkin 2007). The authors acknowledge that there are potentially limitless sources of bias, and they outline 60 or so of the most common types, at five stages of research. Table 1 gives a modified version of Jadad and Enkin's list. This provisional catalog is helpful for demonstrating, in concrete detail, the pervasive

TABLE 1 BIASES IN RANDOMIZED CONTROLLED TRIALS

Planning phase	*Duration*	*Reporting*	*Dissemination*
Choice of question (hidden agenda/vested interest, self-fulfilling prophecy, cost and convenience, funding availability, secondary gains search)	Population choice (gender, age, special circumstances, recruitment, informed consent, literary, language, severity of illness)	Withdrawal	Publication
Regulation	Intervention choice (too early, too late, learning curve, complexity)	Selective reporting (social desirability, optimism, data-dredging, interesting data)	Language (country of publication)
Wrong design	Comparison choice (measurement, time term)		Time lag
	Selection		
	Ascertainment		

Source: Adapted from Jadad and Enkin (2007).

role of values at all stages of medical research—from the planning phase right through to the dissemination of research results—even when best methods are used.

Even in the most methodologically rigorous studies, significant biases can occur. We now have plenty of empirical evidence of persistent bias in RCTs. Researchers have been quite inventive at coming up with new ways to subvert legitimate inquiry (without committing outright fraud), including suboptimal dosing of the competitor's drug in a head-to-head trial, publication of only positive results, publication of only part of the results of a trial, analysis on the basis of secondary endpoints when primary endpoints do not indicate a significant effect of the treatment, and so on (Angell 2004; Parker 2002; Sackett and Oxman 2003). Researchers have found that even when the quality of studies appeared to be the same (that is, the methodological rigor was consistent), positive outcomes were more frequently reported for privately funded drug trials (Cho and Bero 1996). Thus, as Norman (1999) suggests: "methodological rigor is an insufficient measure of freedom from bias" (p. 141). In other words, despite equally good methods in the different studies, bias still played a role in the research outcome.

Even if we were to set aside global social concerns about political and economic influences on the direction of research, and the individual biases introduced by researchers, the catalog of specific biases identified by Jadad and Enkin suggests that bias is pervasive in research. These findings have direct implications for an evidence hierarchy that claims to diminish bias through methodological rigor. Dealing with biases in research will require some creativity and a much broader outlook on the resources scientists have available to them. Textbooks in

epidemiology do sometimes recognize the problem of pervasive bias in research, but in the EBM context the tendency toward predigested evidence limits the opportunity individual clinicians have to engage with original research and identify these biases. Instead, they have to rely on the good will and critical eye of the reviewers who produce systematic reviews and synopses. Needless to say, this trust is not necessarily justified in all cases. It also is not clear what sort of action a systematic reviewer can take to incorporate concerns about bias into a review, just as statements of conflicts of interest on research publications provide information but no clear guideline on how to proceed. (Should I reject the trial from consideration? Should I "flag" a concern with the trial to the other members of the reviewing body?)

For every bias, or negative value, on Jadad and Enkin's list, there is a corresponding positive value. So, for instance, we avoid hidden agenda bias because we assign a positive value to open agendas. We avoid publication bias because we assign a positive value to equality or justice in the evaluation of publications. These positive values, in turn, are justified on the basis of epistemological assumptions about how to best arrive at knowledge in the scientific domain. Philosopher of science Helen Longino (1990) instructively writes: "the question of whether social values can play a positive role in the sciences is really the wrong question. Social and contextual values do play a role, and *whether it is positive or negative depends on our orientation to the particular values in question*" (p. 281, emphasis added). This indicates a need for greater attention to the role of values in medical research. Identifying and evaluating biases that have a negative impact on inquiry is an important project, as is the reeducation of health-care professionals regarding the positive and productive role of values in inquiry. Without an appreciation for this range of roles, the job of weeding out negative values will be superficial. In addition, there is a need for transparency about all values in research. Pervasive values and assumptions need to be critically discussed and evaluated to ensure that idiosyncratic assumptions and values are not unduly shaping research.[3] And we need to begin with the recognition that procedural mechanisms, such as research methods, are only part of any solution to the influence of biases on research, and that social mechanisms and social structures, such as those exemplified by the recent "open science" movement (which stresses transparency, diversity and publicity in research) are in need of fortification.

META-ANALYSES, GUIDELINES, AND BIAS

EBM supporters may argue that meta-analyses can save the day because meta-analyses average results and so wash out the biases of individual studies. A meta-

[3]More specific proposals for dealing with pervasive values have been proposed by social epistemologists. I discuss these constructive solutions further in my doctoral dissertation (Borgerson 2008).

analysis is "a statistical synthesis of the numerical results of several trials which all addressed the same question" (Greenhalgh 2006, p. 122). According to the hierarchy, meta-analyses produce the highest quality of evidence achievable in medicine. Meta-analyses are thought to be advantageous because they assimilate large amounts of information, reduce the delay in translating evidence into practice, and establish the generalizability and consistency of research results (Greenhalgh 2006). In addition, epistemic advantages, such as the ability to minimize bias, are frequently offered in support of meta-analyses. These analyses are assumed to be minimally biased because the studies they group together are already relatively unbiased: thus, meta-analyses of RCTs are thought to be unbiased because they combine the results of several (already quite unbiased) individual RCTs. In addition, meta-analyses are thought to offer practical advantages, such as the ability to assimilate and translate bodies of evidence into practical guidelines that are ready for use.

There is a trend toward the use of meta-analyses, systematic reviews (and synopses or abstracts of systematic reviews), and predigested evidence-based guidelines produced by such groups as the Cochrane Collaboration. For all the good that comes from these guidelines and meta-analyses, we cannot ignore the potential for them to mislead physicians into believing that unbiased results are represented when they are not. This is particularly worrisome when we factor in some of the powerful and influential economic forces behind the production of much medical research today and the interests they have in ensuring their research is taken up by such guidelines. A recent article by David Cundiff (2007) on the financial interests influencing members of the Cochrane Collaboration highlights the importance of critical attitudes toward even the most prestigious guidelines and meta-analyses. It may be that the abstraction from the data of original research is motivated largely by issues of expediency and practicality, but in order to be justified as a good route to knowledge, these approaches ought at least to protect the production of knowledge (if not enable it). As mentioned above, the diminished possibility of bias in meta-analyses is thought to provide at least part of this epistemic justification. But given the critique of RCTs (and all research methods) offered above, it is not clear why we would think that we are doing anything more than pooling the biases of individual studies, and—crucially—failing to acknowledge these biases in the end-product, whether meta-analysis or guideline. The detailed information on biases in trials is generally unavailable in the summaries produced by expert groups. This means that a variety of biases are removed from the view of evidence "users" who rely on these guidelines and reviews. Given the extensive range of biases known to impact clinical trials, this is dangerous. The users of evidence are further removed from the data (of all types), and thus they are less able to critically evaluate that data for biases.

CONCLUSION

In recent discussions of EBM, advocates have tended to suggest that EBM proposes nothing more revolutionary than that empirical evidence should inform medical practice. The language of EBM has changed from calls for paradigm shifts and revolutions to talk of integration and judicious, conscientious inclusion. This overly charitable characterization of a gentler, friendlier EBM, however, fails to recognize the enduring, and central, role accorded to the evidence hierarchy, and it will remain inaccurate (though sadly so) as long as the evidence hierarchy persists within the movement.

The critical analysis of the evidence hierarchy offered in this article does not indicate a lack of appreciation for the motivations behind EBM. The members of the EBM Working Group sought to bring about a more rational, more rigorous, and more humane medical practice. Although the details of their attempt to improve medicine were less than ideal, EBM has forced physicians to talk about standards of evidence, the elements of clinical decision-making, and methods of assessing clinical research. And while the movement has shifted in recent years, during the period that it emphasized critical thinking, it provided an important perspective on the value of analytic skills for medical professionals. Further, when one looks at the prominent physicians today who advocate for improvements to medicine, the early proponents of EBM are notable. For instance, it is Gordon Guyatt who drew attention to the value of the otherwise little-known "*n* of 1" method in research (Guyatt et al. 1986). It is members of the Cochrane Collaboration who have most actively lobbied for a clinical trials registry (Rennie 2004). And it is David Sackett and colleagues who have written the most comprehensive and provocative guide to the ways in which research evidence can be biased by corporate interests (Sackett and Oxman 2003).

These valuable contributions, however, do not justify the assumptions underlying the evidence hierarchy. While both critics and defenders of EBM increasingly recognize that some justifications of the hierarchy are not as robust as originally supposed, few have appreciated just how bad the situation really is. In conjunction with arguments showing that the causal justification offered for the hierarchy fails, this article identifies grounds for significant concern about the way medical research is conducted and reasons against using the EBM hierarchy as a guide to clinical practice. Because of the limited capacity of research methods to control for bias, we have good reason to insist on the transparency and publicity of medical research: insofar as meta-analyses and guidelines decrease access to information about potential biases, they do not help to address these biases and might even make the situation worse. Not only is the EBM hierarchy of evidence failing to secure knowledge, it may be used to limit access to original data. In light of pervasive biases in medical research, this can only be damaging to the pursuit of knowledge.

REFERENCES

Angell, M. 2004. *The truth about the drug companies: How they deceive us and what to do about it*. New York: Random.

Ashcroft, R. E. 2004. Current epistemological problems in evidence based medicine. *J Med Ethics* 30:131–35.

Borgerson, K. 2008. Valuing and evaluating evidence in medicine. PhD diss., Univ. of Toronto.

Bradford-Hill, A. 1965. The environment of disease: Association or causation? *Proc Roy Soc Med* 58:295–300.

Canadian Task Force on Preventive Health Care. 2008. http://www.ctfphc.org/.

Cho, M. K., and L. A. Bero. 1996. The quality of drug studies published in symposium proceedings. *Ann Intern Med* 124(5):485–89.

CONSORT (Consolidated Standards of Reporting Trials) Group. 2008. CONSORT statement. http://www.consort-statement.org/.

Cundiff, D. 2007. Evidence-based medicine and the Cochrane Collaboration on trial. *MedGenMed* 9(2):56.

Evidence Based Medicine Working Group (EBMWG). 1992. Evidence based medicine: A new approach to teaching the practice of medicine. *JAMA* 268(17):2420–25.

Fisher, R. A. 1947. *The design of experiments*, 4th ed. Edinburgh: Oliver and Boyd.

Greenhalgh, T. 2006. *How to read a paper: The basics of evidence-based medicine*, 3rd ed. London: BMJ Books.

Grossman, J., and F. J. MacKenzie. 2005. The randomized controlled trial: Gold standard, or merely standard? *Perspect Biol Med* 48(4):516–34.

Guyatt, G., and D. Rennie. 2001. *Users' guide to the medical literature*. Chicago: AMA Press.

Guyatt, G. H., D. L. Sackett, and D. J. Cook. 1994. How to use an article about therapy or prevention: What were the results and will they help me in caring for my patients? *JAMA* 271(1):59–66.

Guyatt, G. H., et al. 1986. Determining optimal therapy: Randomized trials in individual patients. *New Engl J Med* 314(14):889–92.

Haynes, R. B., et al. 2006. *Clinical epidemiology: How to do clinical practice research*, 3rd ed. Philadelphia: Lippincott, Williams and Wilkins.

Howson, C., and P. Urbach. 2006. *Scientific reasoning: The Bayesian approach*, 3rd ed. Chicago: Open Court.

Jadad, A., and M. Enkin. 2007. *Randomized controlled trials: Questions, answers and musings*, 2nd ed. Oxford: Blackwell.

Longino, H. 1990. *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton: Princeton Univ. Press.

Norman, G. R. 1999. Examining the assumptions of evidence-based medicine. *J Eval Clin Pract* 5(2):139–47.

Oxford Centre for Evidence-Based Medicine. 2008. http://www.cebm.net/levels_of_evidence.asp.

Parker, M. 2002. Whither our art? Clinical wisdom and evidence-based medicine. *Med Health Care Philos* 5:273–80.

Rennie, D. 2004. Trial registration: A great idea switches from ignored to irresistible. *JAMA* 292(11):1359–62.

Russo, F., and J. Williamson. 2007. Interpreting causality in the health sciences. *Int Stud Philos Sci* 21(2):157–70.

Sackett, D. L., and A. D. Oxman. 2003. HARLOT plc: An amalgamation of the world's two oldest professions. *BMJ* 327:1442–45.

Sackett, D. L., et al. 1991. *Clinical epidemiology: A basic science for clinical medicine*, 2nd ed. Toronto: Little, Brown.

Sackett, D. L., et al. 1996. Evidence-based medicine: What it is and what it isn't. *BMJ* 312: 71–72.

Straus, S. E., et al. 2005. *Evidence-based medicine: How to practice and teach EBM*. Toronto: Elsevier.

Upshur, R. 2003. Are all evidence-based practices alike? Problems in the ranking of evidence. *CMAJ* 169(7):672–73.

Worrall, J. 2002. What evidence in evidence-based medicine. *Philos Sci* 69(3):S316–S330.

Worrall, J. 2007. Why there's no cause to randomize. *Br J Philos Sci* 58:451–88.