Euro Jnl Phil Sci (2011) 1:451–466
DOI 10.1007/s13194-011-0034-6

ORIGINAL PAPER IN PHILOSOPHY OF MEDICINE

Just a paradigm: evidence-based medicine in epistemological context

Miriam Solomon

Received: 8 October 2010 / Accepted: 28 July 2011 / Published online: 6 September 2011

Abstract Evidence-Based Medicine (EBM) developed from the work of clinical epidemiologists at McMaster University and Oxford University in the 1970s and 1980s and self-consciously presented itself as a "new paradigm" called "evidence-based medicine" in the early 1990s. The techniques of the randomized controlled trial, systematic review and meta-analysis have produced an extensive and powerful body of research. They have also generated a critical literature that raises general concerns about its methods. This paper is a systematic review of the critical literature. It finds the description of EBM as a Kuhnian paradigm helpful and worth taking further. Three kinds of criticism are evaluated in detail: criticisms of procedural aspects of EBM (especially from Cartwright, Worrall and Howick), data showing the greater than expected fallibility of EBM (Ioaanidis and others), and concerns that EBM is incomplete as a philosophy of science (Ashcroft and others). The paper recommends a more instrumental or pragmatic approach to EBM, in which any ranking of evidence is done by reference to the actual, rather than the theoretically expected, reliability of results. Emphasis on EBM has eclipsed other necessary research methods in medicine. With the recent emphasis on translational medicine, we are seeing a restoration of the recognition that clinical research requires an engagement with basic theory (e.g. physiological, genetic, biochemical) and a range of empirical techniques such as bedside observation, laboratory and animal studies. EBM works best when used in this context.

Keywords Evidence-based medicine · Philosophy of medicine · Kuhnian paradigm · Nancy Cartwright · Randomized controlled trial · Translational medicine · Evidence hierachy · Meta-analysis · John Ioannidis · John Worrall

I am grateful to two anonymous reviewers for the European Journal for Philosophy of Science for helpful corrections and comments on this paper.

M. Solomon (✉)
Department of Philosophy, Temple University, Philadelphia, PA 19122, USA
e-mail: msolomon@temple.edu

1 Introduction

Evidence-Based Medicine (EBM)[1] is the application of methods of clinical epidemiology to the practice of medicine more generally. It was inspired by the post-World War II work of Archibald Cochrane, developed from the work of clinical epidemiologists at McMaster University and Oxford University in the 1970s and 1980s, and self-consciously presented as a "new paradigm" called "evidence-based medicine" in the early 1990s (Evidence-Based Medicine Working Group 1992). EBM was embraced in Canada and the UK in the 1990s, received with some ambivalence in the United States, and adopted in many other countries, both developed and developing (Daly 2005). Its techniques of population based studies and systematic review have produced an extensive and powerful body of knowledge about medical diagnosis and treatment. A canonical and helpful definition of EBM[2] is that of Davidoff et al. (Davidoff et al. 1995) in an editorial in the *British Medical Journal*:

> "In essence, evidence based medicine is rooted in five linked ideas: firstly, clinical decisions should be based on the best available scientific evidence; secondly, the clinical problem - rather than habits or protocols - should determine the type of evidence to be sought; thirdly, identifying the best evidence means using epidemiological and biostatistical ways of thinking; fourthly, conclusions derived from identifying and critically appraising evidence are useful only if put into action in managing patients or making health care decisions; and, finally, performance should be constantly evaluated."

EBM regards its own epistemic techniques as superior to other more traditional methods such as clinical experience, expert opinion, and physiological reasoning. This is because the more traditional techniques are viewed as more fallible. There is not one new technique, but several. The following are typically regarded as part of EBM:

1. Rigorous design of clinical trials, especially the randomized controlled trial (RCT). The RCT is to be used wherever physically and ethically feasible. The trial should be double-masked (traditionally, "double-blinded"[3]) wherever possible.
2. Systematic evidence review and meta-analysis, including grading of the evidence in "evidence hierarchies."
3. Outcome measures (leading to suggestions for improvement)

[1] The term "evidence-based practice" may be replacing EBM, acknowledging the fact that the practice of medicine requires not only physicians but other health care professionals.

[2] Some canonical definitions are unhelpful for a general understanding of EBM, for example, that given in (Sackett et al. 1996): "Evidence-based medicine is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients." This particular definition is probably used widely because it is brief, and because it has a rhetorical purpose—to address the frequent criticism of EBM that it applies to populations, not individuals.

[3] "Double-masked" has largely replaced "double-blinded," in order to avoid inappropriate use of terms relating to disability.

The RCT has often been described as the "gold standard" of evidence for effectiveness of medical interventions. It is a powerful technique, originally developed by the geneticist R.A. Fisher and applied for the first time in a medical context by A. Bradford Hill's 1948 evaluation of streptomycin for tuberculosis (Doll et al. 1999). The double-masked RCT is designed to control for the placebo effect, for selection and other confounding biases, and for confirmation biases.

EBM also includes systematic and formal techniques for combining the results of different clinical trials. A systematic review does a thorough search of the literature and an evaluation and grading of clinical trials. An evidence hierarchy is typically used to structure the judgments of quality and strength of evidence. Meta-analysis integrates the actual data from different but similar high-quality trials to give an overall single statistical result.

Often, EBM is supplemented with formal techniques from Medical Decision Making (MDM) such as risk/benefit calculations. The risk/benefit calculations can be made for individual patients, making use of patient judgments of utility, or they can be made in the context of health care economics, for populations. MDM seeks to avoid common errors of judgment, such as availability and salience biases, in medical decision making.[4]

The overall project is to use the techniques of EBM (and sometimes also MDM) to construct practice guidelines and to take care of individual patients. Each technique—the RCT, other high quality clinical trials, meta-analysis and systematic review–is based on its own core technical successes. The techniques fit together, and share a reliance on statistics, probability theory and utility theory. Journals, centers, clearinghouses, collaborations, educational programs, textbooks, committees and governments all produce and disseminate EBM.

EBM rose to dominance right *after* the prominence of consensus conferences for assessment of complex and sometimes conflicting evidence and may have been partly responsible for the decline of traditional consensus conferences. As late as 1990, an Institute of Medicine report evaluating the international uses of medical consensus conferences said "Group judgment methods are perhaps the most widely used means of assessment of medical technologies in many countries" (Baratz et al. 1990). Just a few years later, expert consensus is viewed in the same medical circles as the lowest level of evidence, when it is included in the evidence at all. For example, the Canadian Task force on Preventive Health Care which began in 1979 as a consensus conference program now explicitly declares "Evidence takes precedence over consensus" (Canadian Task Force on Preventive Health Care).

EBM has not completely replaced group judgment, however. Consensus conferences (or something similar) are often still used for producing evidence-based guidelines or policy, that is, for translating a systematic review into a practical recommendation. And group judgment may be needed to set the standards to be used in systematic review. I will comment on this continued reliance on consensus methods later in the paper.

There is some indication that EBM is now past its peak, and being overshadowed in part by a new approach, that of "translational medicine" (Woolf 2008). Considerable resources from the NIH, from the European Commission and from

[4] A useful essay discussing the differences and interactions between EBM and MDM is (Elstein 2004)

the National Institute for Health Research in the UK have been redirected to "bench to bedside and back" research, which is typically the research that takes place *before* the clinical trials that are core to EBM. Donald Berwick, the founder of the Institute for Healthcare Improvement (which is the leading organization for quality improvement in healthcare), now claims that "we have overshot the mark" with EBM and created an "intellectual hegemony" that excludes important research methods from recognition (Berwick 2005). Berwick calls the overlooked methods "pragmatic science" and sees them as crucial for scientific discovery. He mentions the same sorts of approaches (use of local knowledge, exploration of hypotheses) that "translational medicine" advocates describe. After the discussions in this paper, some reasons for the recent turn to translational medicine will become clearer.

There is a vast literature on evidence-based medicine, most consisting of systematic evidence reviews for particular health care questions. A substantial portion of the literature, however, is a critical engagement with EBM as a whole, pointing out both difficulties and limitations. These discussions come from outsiders as well as insiders to the field of EBM. My goal in this paper is to do something like a systematic review of this literature, discerning the kinds of criticisms that seem cogent and presenting them in a structured manner. EBM, like all methodologies in medicine, has both core strengths and limitations. I will begin with an overview of some general social and philosophical characteristics of EBM, and then turn to the criticisms.

2 EBM as a "Kuhnian paradigm"

When the Evidence-Based Medicine Working Group described themselves as having a "new paradigm" of medical knowledge (Evidence-Based Medicine Working Group 1992), they particularly had in mind Kuhn's (1962) characterization of a paradigm as setting the standards for what is to count as admissible evidence.[5] EBM assessments make use of an "evidence hierarchy" (often called "levels of evidence") in which higher levels of evidence are regarded as of higher quality than lower levels of evidence. A typical evidence hierarchy[6] puts double-masked (or "double-blinded") RCTs at the top, or perhaps right after meta-analyses or systematic reviews of RCTs. Unmasked RCTs come next, followed by well designed case controlled or cohort studies and then observational studies and case reports. Expert opinion, expert consensus, clinical experience and physiological rationale are at the bottom. The rationale for the evidence hierarchy is that higher levels of evidence are thought to avoid biases that are present in the lower levels of evidence. Specifically, randomization avoids selection and other confounding biases (but see (Worrall 2007b)) and masking helps to distinguish real from placebo effects (but see (Howick 2008; Howick 2011)) and avoid confirmation bias. Powering the trial with sufficient numbers of participants and using statistical tools avoids the salience and availability

[5] They were not, however, claiming along with many Kuhnians that there is subjectivity or relativity involved in what gets to count as evidence.

[6] There are many such hierarchies in use, but all put double-masked RCTs at the top, or right after meta-analyses of RCTs, and clinical experience, expert consensus and physiological rationale at the bottom.

biases that can skew informal assessments and unsystematic clinical experience. Ultimately, trial results are graded for both quality and strength of evidence.

The language of Kuhnian paradigms has been overused and become somewhat clichéd, meaning something like "transformational new theory" in the typical quote from the EBM Working Group cited in the previous paragraph. In fact, EBM has many characteristics of a *traditional* Kuhnian paradigm,[7] having all three of the characteristics of a Kuhnian paradigm discerned by Margaret Masterman (Lakatos and Musgrave 1970) and agreed to by Kuhn (Kuhn et al. 2000). These characteristics[8] are helpful for understanding the import of EBM. First, EBM is a social movement with associated institutions such as Evidence-Based Practice Centers, official collaborations, textbooks, courses and journals. It is also, secondly, a general philosophy of medicine, defining both the questions of interest and the appropriate evidence. It is seen as the central methodology of medicine by its practitioners and as an unwelcome politically dominant movement by its detractors (e.g. (Charlton and Miles 1998; Denny 1999)). And third (sometimes overlooked by those who use Kuhn's term "paradigm"), it is characterized by a core of technical results and successful exemplars that have been extended over time. Kuhn referred to such exemplars as "concrete puzzle solutions…employed as models or examples" (Kuhn 1970) and later as a "disciplinary matrix" including "symbolic generalization, models and exemplars" (Kuhn 1977). He regards this third meaning as the original and fundamental meaning of the term "paradigm" (Kuhn 1977).

Contrary to appearances and self-presentation, this core of technical results is not produced by a general algorithm or set of precise methodological rules. One of the things that Kuhn emphasized about paradigms is that they are driven primarily by exemplars, and not by rules. He writes (Kuhn 1970) that exemplars are "one sort of element…the concrete puzzle-solutions employed as exemplars which can replace explicit rules as a basis for the solution of the remaining puzzles of normal science. Kuhn argued that this is significant because the rules are not the basis for the development of the science. Rather, Kuhn argues, less precise judgments about similarity of examples are used (Kuhn 1977).

The medical RCT traces its beginning to A. Bradford Hill's 1948 evaluation of streptomycin for tuberculosis (Doll et al. 1999). It was initially resisted by many physicians used to treating each patient individually, therapeutically and with confidence in treatment choice (Marks 1997). Nevertheless, important trials such as the polio vaccine field trial of 1954 and the 1955 evaluation of treatments for rheumatic fever helped bring the RCT into routine use (Meldrum 1998; Meldrum 2000). In 1970 the RCT achieved official status in the USA with inclusion in the new FDA requirements for pharmaceutical testing (Meldrum 2000). One of the most well-known early successful uses of RCTs was the 1980s international study of aspirin and streptokinase for the prevention of myocardial infarction. However, not

[7] Interestingly, Sehon and Stanley (2003) argue that EBM should not be thought of as a Kuhnian paradigm, but, instead, as part of a Quinean holistic network of beliefs. My focus here is on the historic, rather than the popular meaning of "paradigm" and on what EBM is, not on what it should be.

[8] There are many other characteristics of Kuhnian paradigms, such as the narrative of paradigm change, the emphasis on high-level theory, and the idea that paradigms replace one another, that do not apply to this case. I am not trying to apply Kuhn exhaustively, just to use some of his concepts where they may be explanatorily useful.

all early use of the RCT was straightforward or successful: the 1960s-1970s NCI randomized controlled trial of lumpectomy versus mastectomy for early stage breast cancer was not masked,[9] yet it was highly regarded, and the attempt to conduct a Diet-Heart study in the 1960s was hampered and finally frustrated by the difficulties in implementing major changes in diet in one arm of the study (Marks 1997). This is an example of the finding that the methodology of the RCT does not readily apply to all the situations in which we might wish to use it. As Kuhn might put it, normal science is not a matter of simple repetition of the paradigm case; it requires minor or major tinkering, and sometimes ends in frustration (or what Kuhn would call an "anomaly").

There are also variations in the design and analyses of RCTs. For example, some trials do an "intention to treat" analysis, dropping no experimental subjects from the trial, even if they fail to go through the course of treatment, and some trials do a "per-protocol" analysis in which only patients who complete the trial are included in the final results. Some trails have a placebo in the control arm and some trials have an established treatment in the control arm. It is often said that design and evaluation of an RCT requires "judgment" (see for example (Rawlins 2008)); by this what is meant is that trials cannot be designed by a universal set of rules and that the design and evaluation of trials requires domain expertise, not only statistical expertise. For example, domain expertise is needed in order to design both the dosage and the intervention in the control arm, and domain expertise is needed in order to specify appropriate trial selection criteria.

The same insights apply to systematic reviews and meta-analyses. The first systematic review is often identified as the Oxford Database of Perinatal Trials 1989 study of corticosteroids for fetal lung development; this study was the basis for the development of the Cochrane Collaboration in 1993 which has since then done over 3,000 systematic reviews. Other organizations producing systematic reviews include the Agency for Health Care Research and Quality (AHRQ) and its fourteen Evidence-Based Practice Centers and the American College of Physicians (ACP) Journal Club. All systematic reviews use evidence hierarchies, but there is some variation in the hierarchies in use. The RCT is always at the top or just below meta-analyses of RCTs, but there are variations in where other kinds of studies are ranked, and in whether or not animal trials, basic science and expert opinion are included. Hierarchical rank is just one measure of the quality of a trial, which needs to be considered together with other measures of quality such as how well the trial handles withdrawals and how well it is randomized and masked. In 2002 the AHRQ reported forty systems of rating in use, six of them within its own network of evidence-based practice centers (AHRQ 2002). The GRADE Working Group, established in 2000, is attempting to reach consensus on one system of rating the quality and strength of evidence (Guyatt et al. 2008). This is an ironic development, given that EBM intends to replace group judgment methods!

Meta-analysis combines the results of several high-quality trials to get an overall measure of strength of evidence. It requires judgments about the similarity of trials for combination and the quality of evidence in each trial, as well as about the possibility of systematic bias in the evidence overall, for example due to publication

[9] Double-masking is the standard for high quality RCTs.

bias and pharmaceutical company support. Meta-analysis is a formal technique, but not an algorithmic one: judgments need to be made about trial quality (as with systematic reviews, use of an evidence hierarchy is part of the process) and similarity of trial endpoints or other aspects of studies. Different meta-analyses of the same data have produced different conclusions (Juni et al. 1999; Yank et al. 2007). Steven Goodman (2002) is concerned that the disagreement between meta-analyses, specifically in the case of mammography screening, represents a "crisis for EBM." I think it is not so much a crisis as a reminder of the limits of EBM.

The identification of EBM with a Kuhnian paradigm, useful though it is, should not be taken too scrupulously. Exemplars and judgments of similarity are important, but rules also play a role. Kuhnian claims about incommensurability between paradigms and the social constitution of objectivity are controversial here and would certainly be denied by practitioners of EBM.[10] We have moved on from Kuhn's ideas, revolutionary in the 1960s, but now built upon and transformed in more sophisticated ways.

3 Critical discussions of EBM

Critical discussions of EBM tend to focus on questioning the procedural necessity and sufficiency of the technical requirements (especially for the RCT), the reliability of EBM in practice, or on EBM's explicit or implicit claims to be a general philosophy of medicine. I'll examine these three areas in turn.

3.1 Criticisms of procedural aspects of EBM

Many of these criticisms of EBM procedures have come from British philosophers of science associated with the London School of Economics. Their main approach is to argue that the "gold standard" (the double-masked RCT) is neither necessary (Howick 2011; Worrall 2007b) nor sufficient (Cartwright 2010) for clinical research. They argue that RCTs do not always control for the biases they are intended to control, they do not produce reliably generalizable knowledge, or they can be unnecessary constraints on clinical testing. These arguments are theoretical and abstract in character, although they are sometimes illustrated by examples. I distinguish them from arguments that RCTs have difficulties *in practice*, that is, from evaluation of RCTs based on the *actual* outcomes of such studies, which will be discussed in the next subsection (3B).

John Worrall (2002, 2007a, 2007b) argues that randomization is just one way, and an imperfect way, of controlling for confounding factors that might produce bias. The problem is that randomization can control only for *most but not for all* confounding factors. When there are indefinitely many factors, both known and unknown, which may lead to bias, chances are that any one randomization will not randomize with respect to *all* these factors. Under these circumstances, Worrall concludes, chances are that any particular clinical trial will have at least one kind of

[10] Critics of EBM have occasionally presented EBM as a political and rhetorical movement, e.g. (Charlton and Miles 1998), emphasizing the ways in which it appears to lack rationality.

bias, making the experimental group relevantly different from the control group, just by accident. The only way to avoid this is to re-randomize and do another clinical trial, which may, again by chance, eliminate the first trial's confounding bias but introduce another. Worrall concludes that the RCT does not yield reliable results unless it is repeated time and again, re-randomizing each time, and the results are aggregated and analyzed overall. This is practically speaking impossible. In context, Worrall is less worried about the reliability of RCTs than he is about the assumption that they are much *more* reliable—in a different epistemic class—than e.g. well-designed observational ("historically controlled") studies in which there is no randomization. He is arguing that the RCT should be taken off its pedestal and that *all* trials can have inadvertent bias due to differences between the control and the experimental group.

In a series of articles, Nancy Cartwright (2007a,b, 2009, 2010) points out that RCTs may have internal validity, but their external validity and hence their applicability to real world questions is dependent on the similarity of the test population and context to the population and context targeted by the intervention. For example, she cites the failure of the California class-size reduction program, which was based on the success of a RCT in Tennessee, as due to failure of external validity (Cartwright 2009). She does not give a medical example of actual failure of external validity (hence my classification of her criticisms of EBM as theoretical in character), although she gives one of possible failure: prophylactic antibiotic treatment of children with HIV in developing countries. UNAIDS and UNICEF 2005 treatment recommendations were based on the results of a 2004 RCT in Zaire. Cartwright is concerned that the Zaire results will not generalize to resource-poor settings across other countries in sub-Saharan Africa (Cartwright 2007b) Concern about external validity is reasonable and there are classic medical examples of lack of external validity. For example, some recommendations for the treatment of heart disease, developed in trials of men only, do not apply to women. There is a history of challenges to RCTs on the grounds that they have excluded certain groups from participation (e.g. women, the elderly, children) yet are used for general health recommendations. The exclusions are made on epistemic and/or ethical grounds. These days, women are less likely to be excluded because the NIH and other granting organizations require their participation in almost all clinical trials, but other exclusions, such as those based on age, remain. Cartwright expresses the concern about external validity in its most general form. In her most recent work (Cartwright 2010) she describes four conditions that need to be met for external validity.[11] These four conditions demand considerable domain knowledge i.e. knowledge of the particular causal interactions that the intervention relies upon.

Jeremy Howick (2008, 2011) argues that masking is not useful outside of contexts in which outcomes are measured subjectively and that masking is both impossible and unnecessary when dealing with large effect size. It is impossible when dealing with large effect size because the effects of the drug unmask the assignment. He also argues that masking is in practice inadequate for placebo controlled trials, since

[11] Cartwright's (2010) four conditions are knowledge of "Roman laws"(laws that are general enough), "the right support team" (all necessary conditions), "straight sturdy ladders" (for climbing up and down levels of abstraction) and "unbroken bridges" (no interfering conditions).

participants can usually tell through the presence of side effects whether or not they are receiving an active intervention.

These three sets of criticisms by Worrall, Cartwright and Howick are significant, and I evaluate them next, beginning with Worrall's papers. Randomization is unlikely to control for confounding factors only in the event that there are many *unrelated* population variables that influence outcome, because only in that complex case is one of those variables likely to be accidentally unbalanced by the randomization. Worrall considers only the abstract possibility of multiple unknown variables; he does not consider the likely relationship (correlation) of those variables with one another and he does not give us reason to think that, *in practice*, randomization generally (rather than rarely) leaves some causally relevant population variables accidentally selected for and thereby able to bias the outcome.[12] In addition, successful replication adds evidence that any confounder inadvertently introduced is not causally responsible for the outcome.

Cartwright uses examples from education, economics and international development to show lack of external validity. In general, her examples show failures of interventions to generalize, often because of cultural differences between populations. External validity in medical trials is more explored territory. We typically already know, from some of the trial selection criteria, where the controversies about generalization lie (see also Table 2 in Rawlins (2008), which sets out the problems with generalization). This does not mean that we can figure out the domain of application of trial results in a simple or formulaic manner. Domain expertise is essential for projection, as of course Nelson Goodman argued long ago (1955). Cartwright's four conditions (2010) have a role here as non-formal criteria for assessing external validity. It should also be noted that Cartwright's discussion applies broadly to experimental and evidential reasoning and not specifically to trial methodology. It is not a specific criticism of RCTs, although because of what she calls the "vanity of rigor" of EBM (Cartwright 2007a), the criticism is especially pertinent to RCTs.

Howick is correct that masking is important only for detecting small effects with subjectively measured outcomes, but this is (unfortunately) true of many recent advances in medicine and therefore widely applicable. It is not often that we have a new intervention with the dramatic success of e.g. insulin for diabetes or surgery for acute appendicitis. Howick is right to see that the methodology of the RCT is suited for some interventions and not suited for others, but the methodology is, in fact, suited for many if not most of the health care interventions currently in development.

The approaches used by Worrall, Cartwright and Howick argue that the RCT is neither a necessary nor a sufficient method for getting knowledge from clinical trials. They argue that other methods can be equally or more effective in specific circumstances and that knowledge from trials always involves projection to untested domains. EBM enthusiasts are beginning to acknowledge this sensible moderation of their views, as the recent Harveian Oration by Sir Michael Rawlins[13] shows (Rawlins 2008). In this paper, Rawlins argues that "the notion that evidence can be

[12] Worrall might respond that he is only criticizing those methodologists who make abstract and general claims about freedom from "all possible" biases. This is fine, so long as no conclusions are drawn for RCTs in practice.

[13] Michael Rawlins is the head of NICE (National Institute of Health and Clinical Excellence) in the UK, which bases its policies and guidelines on the results of EBM.

reliably placed in hierarchies is illusory," that "striking effects can be discerned without the need for RCTs" and that the findings of RCTs should be extrapolated with caution.

Granting these qualifications puts EBM in the same category as other successful scientific methodologies. They are useful tools in the domains in which they work, but they do not work everywhere or always.

3.2 Effectiveness of EBM methods in practice

How reliable is EBM in practice? RCT and meta-analyses generate claims with stated confidence levels. Typically, RCTs give 95% confidence levels and meta-analyses much higher confidence levels. It follows that each RCT has a 5% chance of producing a false positive (and each meta-analysis much less). Yet, in practice, RCTs and meta-analyses are much more fallible.

Ioannidis (2005) did a study of 59 highly cited original research studies. Less than half (44%) were replicated; 16% were contradicted by subsequent studies and 16% found the effect to be smaller than in the original study; the rest were not repeated or challenged. Another, more well known statistic is that studies funded by pharmaceutical companies—even when properly masked and of highest quality—have an astonishingly higher chance (three or four times the probability of studies not funded by pharmaceutical companies) of showing effectiveness of an intervention than studies not funded by pharmaceutical companies (Als-Nielsen et al. 2003; Bekelman et al. 2003; Bero et al. 2007; Lexchin et al. 2003). And LeLorier et al. (LeLorier et al. 1997) found that 35% of the time, the outcome of RCTs is not predicted accurately by previous meta-analysis.

This is a large and partly unexplained failure rate. Some suggest that factors such as publication bias, time to publication bias and pharmaceutical funding bias (which subtly affects trial design and evaluation) are responsible for the worse-than-expected track record of RCTs, systematic reviews and meta-analyses. Publication bias occurs when studies with null or negative results[14] are not written up or not accepted for publication because they are wrongly thought to be of less scientific significance. Steps to address this bias have been taken in many areas of medical research by creating trial registries and making the results of all trials public. Time to publication bias is a more recently discovered phenomenon: trials with null or negative results, even when they are published, take much longer than trials with positive results (6–8 years for null or negative results compared with 4–5 years for positive results) (Hopewell et al. 2007). It is possible that steps taken to correct for publication bias will also help correct for time to publication bias.

The additional bias created by pharmaceutical funding is not fully understood, especially since many of these trials are properly randomized and double-masked and satisfy rigorous methodological criteria. Some suggest that pharmaceutical companies deliberately select a weak control arm, for example by selecting a low dose of the standard treatment, giving the new drug a greater chance of relative

[14] In this context, a positive trial is one in which the experimental arm of the trial is more effective, a null result is one in which both arms are equally effective, and a negative trial is one in which the control arm is more effective.

success. There can also be biases that enter into the analysis of data, particularly when endpoints are not specified in advance. It is hoped that a clinical trials registry will help correct for publication bias and ex post facto manipulation of endpoints. However, at present we are a long way from correcting for bias created by pharmaceutical funding.[15] Disclosure of funding source is helpful for evaluation, but often this information is lost in systematic review and meta-analysis.

Since the performance of RCTs is so flawed, it is worth asking the question whether other kinds of clinical trials, further down the evidence hierarchy, are even less reliable. This would be expected in the abstract, since the further down the evidence hierarchy, the more possible sources of bias. Studies by Benson and Hartz (2000) and Concato et al. (2000) find that many well-designed observational studies produce the same results as RCTs.[16] The matter is controversial, but a recent article by Ian Shrier et al. strongly argues for the inclusion of observational studies in systematic reviews and meta-analyses (Shrier et al. 2007)

This result corroborates the intervention of early AIDS activists, who argued against the imposition of RCTs for AZT on both ethical and epistemic grounds (Epstein 1996). They argued that such trials are morally objectionable in that they deprive the individuals in the placebo arm of the only hope for a cure (at that time). And they argued that such trials are epistemically unnecessary because an RCT is not the only way to discern the effectiveness of anti-retroviral drugs—a claim that, in hindsight, has proved correct as a combination of historical controlled trials and laboratory studies have provided the knowledge of dramatically effective anti-retrovirals in clinical use today. These days, of course, no-one needs to get a placebo, and RCTs can continue to detect small improvements of protocol without such strenuous moral objections.

Finally, EBM has been asked to evaluate itself using its own standards of evaluation This would involve showing not merely that a specific EBM intervention improves outcomes, but that more general use of systematic evidence reviews and so forth in clinical decision making results in improved outcomes for patients. In theory, we would of course expect improved outcomes. But what matters here is not theory but practice, and no-one has yet designed or carried out a study to test this (Charlton and Miles 1998; Cohen et al. 2004; Straus and McAlister 2000).

3.3 Criticisms of EBM as a general philosophy of medicine

Like most paradigms (new ways of knowing) the light shone on the paradigm flatters it and puts everything else into the shadows. Critics have protested, variously, that EBM overlooks the role of clinical experience, expert judgment, intuition, medical authority, patient goals and values, local health care constraints and the basic medical sciences including the structure of theory and the relations of causation. This is a long and complex list of intertwined scientific, hermeneutic, political and ethical considerations. Perhaps the most common criticism of EBM is that it deals with

[15] I recommend that in the meantime we correct for funding bias by asking for a higher level of significance from the results of trials funded by pharmaceutical companies.

[16] An editorial in the same issue of NEJM (Pocock and Elbourne 2000) strongly protests these conclusions, partly in the name of EBM orthodoxy, but partly also on the basis of some well known RCTs which contradicted the results of observational trials

statistical results, and application of those results to particular cases is said to require a different set of skills (e.g. (Cohen et al. 2004; Feinstein and Horwitz 1997; Hampton 2002; Straus and McAlister 2000; Tonelli 1998)). EBM advocates dispute this, and this is the reason for the early redefinition of the enterprise: "Evidence-based medicine is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett et al. 1996). Sorting out this complicated dispute is beyond the scope of this paper; I do so elsewhere (Solomon, in progress).

Another common complaint, which covers several specific complaints, is that EBM is a scientific approach that overlooks the "art" of medicine (e.g. (Montgomery 2006; Tonelli 1998). Elsewhere I have argued that the traditional dichotomy between the "art" and the "science" of medicine is no longer helpful (Solomon 2008). This paper focuses on what EBM leaves out, rather than on whether to count that as art or as science.

In this subsection I will focus on the persistent criticism that EBM ignores the basic sciences that guide both research and clinical practice (Ashcroft 2004; Bluhm 2005; Charlton and Miles 1998; Cohen et al. 2004; Harari 2001; Tonelli 1998). Basic sciences guide research in suggesting hypotheses about disease processes and mechanisms for action of interventions. Basic sciences guide clinical practice in helping physicians tailor the results of epidemiological studies to the needs of particular patients, who may have unique physiological and pathological conditions. EBM is scientifically superficial: it measures correlations. EBM does not model or theorize about the complete organism, still less the complete organism in its social and environmental context. In terms of scientific theory, it is thin; what some have called "empiricistic" (Harari 2001).[17] Charlton and Miles (1998) claim that it is "statistical rather than scientific." Ashcroft writes that EBM is "autonomous of the basic sciences" and "blind to mechanisms of explanation and causation" (2004, p. 134). Ashcroft regards this as an advantage, rather than a disadvantage, because it means that EBM does not have to worry that our basic theories may be incorrect. Ashcroft allies himself with Nancy Cartwright's realism about phenomenal laws and antirealism about deeper laws[18] at this point. Others, however, see the eschewing of scientific theorizing in favor of discovery of robust statistical correlations as problematic (Harari 2001; Tonelli 2006). They consider theorizing as important to medicine as it is to the pure sciences.

Whatever one's views about scientific realism, EBM typically depends upon a background of basic science research that develops the interventions and suggests the appropriate protocols. It is rare for an intervention without physiological rationale to be tested (although this does happen, especially in the areas of complementary and alternative medicine, but in these cases there is typically an alternative rationale, perhaps in the frameworks of Asian metaphysics). Moreover, as discussed above, scientific judgment enters into the design of appropriate randomized controlled trials (choice of control, test population etc.) and into the interpretation of the applicability of results (external validity). Of course, many interventions with excellent physiological rationales and good in vitro and in vivo performance fail when tested in human beings or fail when tested for external

[17] In fact, EBM is the successor to ancient "empiric" approaches to medicine.
[18] Cartwright is, of course, a realist about underlying causal processes.

validity. That does not mean that there is anything wrong with physiological reasoning or that we can use a more reliable method. The basic science work is fallible, but it is not dispensable. Even Nancy Cartwright (1989) would agree that we cannot replace physical theory with phenomenal laws alone.

In the past 5 years a new approach to medical research has risen to prominence internationally: what is called "translational medicine," to be achieved by creating research centers, as well as journals, conferences, training programs and so forth. The NIH has made it a priority in its "Roadmap" in 2004 and started offering Clinical and Translational Science Awards in 2006. 55 Institutes have been created (as of May 2011), mostly in universities and medical centers, and the NIH hopes to fund 60, at a total cost of $500 million annually. The European Commission plans to use most of its billion Euro a year budget for the next few years for translational research. In the UK, the National Institute for Health Research has established 11 centers at a total cost of about 100 million pounds annually. The idea behind translational medicine is to facilitate greater interaction between basic science research and research in clinical medicine.[19] The buzzwords are "synergize," "catalyze" and "interdisciplinary." The idea is to bring the different researchers and their laboratories into greater physical proximity. This is an interesting retro-intervention in these days of global electronic communication and global travel.

From the perspective of the discussion in this section, the development of translational medicine or something like it was only a matter of time. EBM has such high claims to scientific objectivity that it attracted much talent and effort from clinical researchers. Perhaps the increased focus on formal epidemiological work eventually made apparent what was left out, namely engagement with substantial physiological and biomolecular theories. The model of basic science doing the research and clinical researchers testing the products is now perceived as limited; actually, it leaves all the fun and the creativity to the basic researchers, and deprives them of the input of clinical knowledge and observations from the clinical researchers.

4 Conclusions

EBM gives a set of formal techniques for evaluating the effectiveness of clinical interventions. The techniques are powerful, especially when evaluating interventions that offer incremental advantages to current standards of care, and especially when the determination of success has subjective elements. EBM techniques do not deliver the reliability that is theoretically and statistically expected from them. Results are compromised by publication bias, time to publication bias, interests of funding organizations and other unknown factors. Maintaining a strict evidence hierarchy makes little sense when the actual reliability of "gold standard" evidence is so much less than the expected reliability. I recommend a more instrumental or pragmatic approach to EBM, in which any ranking of evidence is done by reference to the

[19] Technically, "translational research" includes both the bench-to-bedside-and-back (T1) and the clinical research to everyday practice (T2) "translational blocks." See (Woolf 2008). But most of the resources and rhetoric favor the former.

actual, rather than the theoretically expected, reliability of results. So, for example, RCTs and observational trials might be at the same level in the hierarchy (based on their comparable reliability in practice) and trials designed and funded by pharmaceutical companies a level below independent trials (irrespective of apparent trial rigor, based on their track record of biased outcomes).

Emphasis on EBM has eclipsed other necessary research methods in medicine, even those methods necessary for its own development and application. With the recent emphasis on translational medicine, we are seeing a restoration of the recognition that clinical research requires an engagement with basic theory (e.g. physiological, genetic, biochemical) and a range of empirical techniques such as bedside observation, laboratory and animal studies. EBM works best when used in this pluralistic methodological context.

References

AHRQ. (2002). Systems to rate the strength of scientific evidence. *Evidence Report/Technology Assessment 47,*

Als-Nielsen, B., Chen, W., Gluud, C., & Kjaergard, L. L. (2003). Association of funding and conclusions in randomized drug trials: A reflection of treatment effect or adverse events? *JAMA: The Journal of the American Medical Association, 290*(7), 921–928.

Ashcroft, R. E. (2004). Current epistemological problems in evidence based medicine. *Journal of Medical Ethics, 30*(2), 131–135.

Baratz, S. R., Goodman, C., & Council on Health Care Technology. (1990). *Improving consensus development for health technology assessment: An international perspective*. Washington: National Academy Press.

Bekelman, J. E., Li, Y., & Gross, C. P. (2003). Scope and impact of financial conflicts of interest in biomedical research: A systematic review. *JAMA: The Journal of the American Medical Association, 289*(4), 454–465.

Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *The New England Journal of Medicine, 342*(25), 1878–1886.

Bero, L., Oostvogel, F., Bacchetti, P., & Lee, K. (2007). Factors associated with findings of published trials of drug-drug comparisons: Why some statins appear more efficacious than others. *PLoS Medicine, 4*(6), e184.

Berwick, D. M. (2005). Broadening the view of evidence-based medicine. *Qual. Saf. Health Care, 14*(5), 315–316.

Bluhm, R. (2005). From hierachy to network: a richer view of evidence for evidence-based medicine. *Perspectives in Biology and Medicine, 48*(4), 535–547.

Canadian Task Force on Preventive Health Care. *CTFPHC History/Methodology.* Retrieved 07/07, 2009, from http://www.ctfphc.org

Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford; New York: Clarendon Press; Oxford University Press.

Cartwright, N. (2007a). Are RCTs the gold standard? *Biosocieties, 2*(2), 11–20.

Cartwright, N. (2007b). Evidence-based policy: Where is our theory of evidence? *Center for Philosophy of Natural and Social Science, London School of Economics, Technical Report 07/07,*

Cartwright, N. (2009). Evidence-based policy: What's to be done about relevance. *Philosophical Studies, 143*(1), 127–136.

Cartwright, N. (2010). *Will this policy work for you? predicting effectiveness better: How philosophy helps.* Unpublished manuscript.

Charlton, B. G., & Miles, A. (1998). The rise and fall of EBM. *Quarterly Journal of Medicine, 91*(5), 371-371-374.

Cohen, A. M., Stavri, P. S., & Hersh, W. R. (2004). A categorization and analysis of the criticisms of evidence-based medicine. *International Journal of Medical Informatics, 73*(1), 35-35-43.

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *The New England Journal of Medicine, 342*(25), 1887–1892.

Daly, J. (2005). *Evidence-based medicine and the search for a science of clinical care*. Berkeley; New York: University of California Press; Milbank Memorial Fund.
Davidoff, F., Haynes, B., Sackett, D., & Smith, R. (1995). Evidence based medicine. *BMJ (Clinical Research Ed.), 310*(6987), 1085–1086.
Denny, K. (1999). Evidence-based medicine and medical authority. *Journal of Medical Humanities, 20*(4), 247–263.
Doll, R., Peto, R., & Clarke, M. (1999). First publication of an individually randomized trial. *Controlled Clinical Trials, 20*(4), 367–368.
Elstein, A. S. (2004). On the origins and development of evidence-based medicine and medical decision making. *Inflammation Research, 53*(Supplement 2), S184-S184-S189.
Epstein, S. (1996). *Impure science: AIDS, activism, and the politics of knowledge*. Berkeley: University of California Press.
Evidence-Based Medicine Working Group. (1992). Evidence-Based Medicine. A new approach to teaching the practice of medicine. *JAMA: The Journal of the American Medical Association, 268*(17), 2420–2425.
Feinstein, A. R., & Horwitz, R. I. (1997). Problems in the "evidence" of "evidence-based medicine". *The American Journal of Medicine, 103*(6), 529–535.
Goodman, N. (1955). *Fact, fiction and forecast*. Cambridge: Harvard University Press.
Goodman, S. N. (2002). The mammography dilemma: A crisis for evidence-based medicine? *Annals of Internal Medicine, 137*(5 Part 1), 363–365.
Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., et al. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ (Clinical Research Ed.), 336*(7650), 924–926.
Hampton, J. R. (2002). Evidence-based medicine, opinion-based medicine, and real-world medicine. *Perspectives in Biology and Medicine, 45*(4), 549–568.
Harari, E. (2001). Whose evidence? lessons from the philosophy of science and the epistemology of medicine. *Australian and New Zealand Journal of Psychiatry, 35*(6), 724-724-730.
Hopewell, S., Clarke, M. J., Stewart, L., & Tierney, J. (2007). Time to publication for results of clinical trials. *Cochrane Database of Systematic Reviews,* (1) doi:10.1002/14651858.MR000011.pub2
Howick, J. (2008). Against *a priori* judgments of bad methodology: Questioning double-blinding as a universal methodological virtue of clinical trials. *PhilSci Archive*
Howick, J. (2011). *The philosophy of evidence-based medicine* Wiley-Backwell.
Ioannidis, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA: The Journal of the American Medical Association, 294*(2), 218–228.
Juni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA: The Journal of the American Medical Association, 282*(11), 1054–1060.
Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.
Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change*. Chicago: University of Chicago Press.
Kuhn, T. S., Conant, J., & Haugeland, J. (2000). *The road since structure: Philosophical essays, 1970–1993, with an autobiographical interview.* Chicago: University of Chicago Press.
Lakatos, I., & Musgrave, A. (1970). *Criticism and the growth of knowledge*. Cambridge: University Press.
LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *The New England Journal of Medicine, 337*(8), 536–542.
Lexchin, J., Bero, L. A., Djulbegovic, B., & Clark, O. (2003). Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. *BMJ (Clinical Research Ed.), 326*(7400), 1167–1170.
Marks, H. M. (1997). *The progress of experiment: Science and therapeutic reform in the united states, 1900–1990*. Cambridge: Cambridge University Press.
Meldrum, M. L. (1998). A calculated risk: The salk polio vaccine field trials of 1954. *British Medical Journal, 317*, 1233-1233-6.
Meldrum, M. L. (2000). A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/Oncology Clinics of North America, 14*(4), 745-745-760.
Montgomery, K. (2006). *How doctors think: Clinical judgment and the practice of medicine*. Oxford: Oxford University Press.
Pocock, S. J., & Elbourne, D. R. (2000). Randomized trials or observational tribulations? *The New England Journal of Medicine, 342*(25), 1907–1909.

Rawlins, M. (2008). De testimonio: On the evidence for decisions about the use of therapeutic interventions. *Lancet, 372*(9656), 2152–2161.

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *BMJ (Clinical Research Ed.), 312*(7023), 71–72.

Sehon, S. R., & Stanley, D. E. (2003). A philosophical analysis of the evidence-based medicine debate. *BMC Health Services Research, 3*(14).

Shrier, I., Boivin, J. F., Steele, R. J., Platt, R. W., Furlan, A., Kakuma, R., et al. (2007). Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. *American Journal of Epidemiology, 166*(10), 1203–1209.

Solomon, M. (2008). Epistemological reflections on the art of medicine and narrative medicine. *Perspectives in Biology and Medicine, 51*(3), 406–417.

Solomon, M. (in progress). *Beyond the art and science of medicine.* Unpublished manuscript.

Straus, S. E., & McAlister, F. A. (2000). Evidence-based medicine: A commentary on common criticisms. *Canadian Medical Association Journal, 163*(7), 837–841.

Tonelli, M. R. (1998). The philosophical limits of evidence-based medicine. *Academic Medicine, 73*(12), 1234–1240.

Tonelli, M. R. (2006). Integrating evidence into clinical practice: An alternative to evidence-based approaches. *Journal of Evaluation in Clinical Practice, 12*(3), 248–256.

Woolf, S. H. (2008). The meaning of translational research and why it matters. *JAMA: The Journal of the American Medical Association, 299*(2), 211–213.

Worrall, J. (2002). What evidence in evidence-based medicine? *Philosophy of Science, 69*(S3), 316–330.

Worrall, J. (2007a). Evidence in medicine and evidence-based medicine. *Philosophy Compass, 2*(6), 981–1022.

Worrall, J. (2007b). Why there's no cause to randomize. *The British Journal for the Philosophy of Science, 58*(3), 451–488.

Yank, V., Rennie, D., & Bero, L. A. (2007). Financial ties and concordance between results and conclusions in meta-analyses: Retrospective cohort study. *BMJ (Clinical Research Ed.), 335*(7631), 1202–1205.