Social Epistemology
Vol. 22, No. 4, October–December 2008, pp. 325–331

Routledge
Taylor & Francis Group

INTRODUCTION

The Rhetoric of Research Methodology

Jason Grossman and Joan Leach

How Can We Describe the Dominant Rhetoric that is Evidence-Based Medicine?

Evidence-Based Medicine (EBM) is the dominant rhetoric in health research. So effective is this dominant rhetoric that if you work outside EBM or question its practices your position is unlikely to be taken seriously. This is doubly so if you question EBM from within a complimentary or alternative medicine framework. You are likely to be seen as a complainer; as not as rigorous as you should be; as operating from the epistemic Paleolithic.

This issue of *Social Epistemology* explores what EBM is, what it should be, and why it has become the dominant rhetoric and everything else has become a counter-rhetorical force.

So What is EBM?

Most importantly, Evidence-Based Medicine (EBM) is *not* just medicine based on evidence. To make this clear, consider what it would look like if, counterfactually, EBM *was* medicine based on evidence. For one thing, it would then be restricted to medicine; but while EBM is, perhaps, restricted to the health disciplines, its scope runs way beyond the territory claimed by the gatekeepers of what is medical. In that sense, the scope of EBM is broader than the uninitiated might expect. But more importantly (and therefore receiving far more attention in this issue of *Social Epistemology*), there is a sense in which the scope of EBM is *narrower* than its name suggests. This is in the dimension of methodology. Does anything based on evidence count as evidence-based? Absolutely not.

This claim, that something based on evidence can fail to be evidence-based in the terms set by EBM, might seem to be a bold one. It requires two types of support: on one side of the equation, it requires us to know what we mean by "evidence"; on the other side it requires proper documentation of the claim that much that is evidence (once we know what that means) is *verboten* according to EBM.

ISSN 0269–1728 (print)/ISSN 1464–5297 (online)
DOI: 10.1080/02691720802567365

So what is evidence? That might seem to be a crucial question. But we think that in order to see the poverty of EBM, it is not necessary to answer it very fully. We take the following (actual) exchange as our guide:

> *The scene is a university seminar room circa 2001. A seminar on Lacan is in progress.*
> Audience member: "There's no evidence for that".
> Speaker: "You have too narrow a view of evidence".
> Audience member: "I don't have any view of evidence. Make some noises designed to get me to believe it".

The lesson we take home from this exchange (which we take to be a win for the opening player) is that any piece of language (or in principle any other semiosis) which supports an assertion is evidence for that assertion. And although we cannot argue this here in detail, it seems obvious that convincing noises (perhaps with some caveats about rationality or social norms or some such) are what make for *good* evidence.

EBM Takes a Very Different View

As we will see in more detail in the papers in this issue (especially those by Derkatch, La Caze, and Grossman), what counts as good evidence in EBM, and even in many cases what counts as evidence at all, depends on a particular view (a *very* particular view) of The Scientific Method. This Scientific Method (even capitals cannot do justice to the singularity and respect it is granted) is ossified in various competing yet highly similar sets of prescriptions enunciated by the various experts in EBM. As La Caze puts it:

> experimental evidence from clinical studies is distinguished according to *methodology* (La Caze, this issue, 356; author's emphasis)

If you run with EBM's methodological prescriptions, you are providing evidence. If you don't, you are not … no matter that you might be supporting a major conclusion with empirical data, a knock-down argument, a mathematical proof, and the Tablets of Moses.

Note that there are *several* contentious ingredients here: the very idea that there is a single Scientific Method; the ossification of that method into a set of low-level rules; a whole variety of problems with what exactly those rules are (see especially Dowe and Grossman); and the exclusion of anything else.

Among the types of evidence which do not count as evidence for EBM, the types most often disputed are various kinds of "observational" studies, a term of art used to denote and (to some extent) denigrate any empirical study which is not a randomised controlled trial. The papers by Derkatch, La Caze, and Grossman discuss problems with this particular exclusion. But another exclusion often goes unnoticed, although very clear in both the literature on EBM and the literature of EBM. This is the exclusion of "basic" science, which in the EBM lexicon means anything which is not a direct examination of the effects of a proposed intervention on humans. Most prominent in this category of basic science is laboratory research, including such varied disciplines as organic chemistry, molecular genetics, biophysics, physiology, pharmacology, histology and so on. As Adam La Caze explains:

> These sciences are primarily focused on developing a theoretical basis for how the body works (physiology), how drugs interact with physiological processes in the body (pharmacology) and the physiological abnormalities involved in disease (pathophysiology). (La Caze, this issue, 355)

Even randomised controlled trials performed on (non-human) animals count as basic science according to EBM, and therefore cannot count as good quality evidence.

La Caze is one of the few theorists who are exploring this strange categorisation of evidence and the resulting cognitive dissonance in a scientific community which takes basic research as its bedrock and yet denies that it gives us good evidence. In his paper in this issue, he considers an obvious move to solve this problem, namely restricting the scope of EBM so that basic research would be considered to give bad evidence *for some medical purposes* but good evidence for other purposes. He shows that although this move has merit it does not succeed in making sense of the evidence categorisations of EBM.

So What Exactly is EBM?

Now we know what EBM is not. We also know, in the abstract, that EBM is effectively defined by a set of rules and, of course, social practices around those rules. And we know a little about the rules. But what exactly are the rules? Why haven't we just listed them for you?

The attempt to answer these questions at the level of detail required to understand the rules and social practices of EBM constitutes a whole field of research in itself, one which is pursued in part in the papers by Lipworth et al., La Caze, and Grossman. While there is some research talking about medical practice using EBM (especially RCTs), and we hope there will be more, what we are doing in this edition of *Social Epistemology* is looking more abstractly at what questions such sociological research should attempt to answer.

A question which needs to be raised immediately is: who is best placed to comment on what EBM is and should be? And who is even qualified to do so? It is an enquiry in which doctors have so far been the presumed experts. Indeed, it is implicit in much EBM literature that only the leaders of the movement, a surprisingly small number of clinical epidemiologists, are fully qualified to define EBM. Lipworth et al. discuss this concentration of power in terms of *exceptionalism*, the idea being that it is primarily the claim that EBM is *special* which has resulted in the protection of its central doctrines by its leaders:

> it is conceivable that EBM exceptionalism has resulted in a similar concentration of power in the hands of those who accept and have expertise in the principles and practices of EBM, and has marginalized those who do not. (Lipworth et al., this issue, 427–8)

Lipworth et al. warn of the potential dangers of exceptionalism. Were EBM treated as less exceptional, they say, it might be more able to incorporate "novel" ideas (although this might come at some cost):

> non-exceptionalism allows practitioners to draw on insights from elsewhere rather than "re-inventing the wheel". … it could be very useful to draw on the rich insights and

> practical strategies of those who, throughout the history of medicine, have reflected on the role of evidence in medical practice. And it could be very useful to recognize the similarities between the evidence-generating and evidence-applying principles of EBM and those of law, engineering and politics. (Lipworth et al., this issue, 428)

Even theoretical statisticians are relatively disenfranchised by EBM. For the importance of this point, see the papers by Dowe and Grossman.

The question of the definition of EBM, taking into account social practices as well as rules, is one which the medical disciplines do not yet know how to take seriously. So far, the story of EBM is internalist history, history told by the victor, even though the victory isn't yet clear. However, the ramifications of EBM are large, too large for busy doctors. Could a broad coalition of those engaged in social research do a more thorough job than the purely internalist story being told so far?

We are proud to note that none of the papers in this special edition offer a definition of EBM. Rather, they work with the multifarious practices that go under the name of EBM.

Colleen Derkatch argues that it is the definition which is precisely at issue. She also shows us how EBM operates to exclude certain pratices, namely complementary and alternative medicine (CAM), precisely by encouraging CAM practitioners to submit themselves to the scrutiny of EBM.

The other papers in this issue deal with the definitional issue implicitly, showing once again the ways in which practice is infinitely richer than definition. As David Mercer points out, the definition of EBM has expanded to include a range of literary technologies involved in the production of medical knowledge, including clinical practice guidelines:

> The eminent clinician has not necessarily been simply replaced by the "scientific expert clinician", but by teams of librarians, economists, and bureaucrats with skills in informatics as well as basic medical knowledge. (Mercer, this issue, 412)

EBM has not only expanded its evaluative scope to incorporate more and more of medical and health practice, but has also produced new problems and new genres. In Mercer's view, EBM is not just a technical discipline; it is a social movement. This is evidenced by the complicity of EBM approaches with new legal standards in Anglo-American law, most notably the parallels between the use of EBM in medical negligence cases and the 1993 Daubert ruling which sets limits on the use of expert witnesses in US courts. As Mercer notes,

> Medical practitioners will find it more difficult to defend themselves against claims of medical negligence by suggesting they have simply followed standards of care set by the medical profession rather than a standard of care backed by an official clinical guideline. (Mercer, this issue, 417)

A scientist is no longer allowed to say, "Here is why I am making my claim"; neither in court, nor in the medical literature, nor in the clinic. A good reason for believing something is no longer good enough. Instead, she must say, "Here is a pre-existing rule which agrees with me". The advent of EBM has rendered it illegitimate to even attempt to overrule epistemic authority. An EBM hierarchy of evidence plays the role of a Papal

bull. A Protestant appeal to an epistemic conscience is not only likely to fail but is not even considered legitimate. Grossman (this issue) refers to this as a failure of methodological pluralism.

Continuing the metaphor, Papal bulls have at least the advantage that the Vatican will interpret them for you. The Bible, on the other hand, is famously hard to interpret. La Caze tells us that an EBM evidence hierarchy is less like an encyclical and more like the Bible. The EBM hierarchy appears to tell us what "type" of evidence a given study provides us with, where "type" connotes *valeur*. However, the interpretation of the evidence given by a study is sensitive to minute details of topic and statistical analysis (La Caze, Dowe, Grossman). The level of methodological fetish required to determine the "type" of evidence represented by a study is so counter-intuitive (according to La Caze and Grossman) that users of EBM, rather than appear mad, are often forced to override their own rules. This flexibility, commendable though it is, threatens the very structure of EBM as a set of rules. It threatens to open the field of discourse up in exactly the same sense that EBM has closed it down. It is much like the Christian response to the requirement in Exodus 35:2 to kill anyone who works on the Sabbath: namely, that the requirement is manifestly unacceptable but the notion of "living by the Bible" is not, even though it is the Bible which specifies the requirement. Self-contradiction and cognitive dissonance loom. At any rate, the silliest consequences of the rules of EBM are (very sensibly) ignored, while the set of rules as a whole is still praised. Whether this can be made coherent is open to question.

Medicine's Methodological Turn

Jeanne Daly's recent history of EBM, *Evidence-Based Medicine and the Search for a Science of Clinical Care* (Berkeley and Los Angeles: University of California Press, 2005), documents the early history of EBM. Before it was anything else, and before it began to raise contentious issues or hackles, EBM was about turning what had been issues of substance into issues of methodology.

EBM emerged from the overlap between:

A: clinical epidemiology—the application of certain statistical methods to certain social contexts (namely medicalised bodies in hospitals)

and

B: clinical epidemiologists—doctors trained in biostatistics.

In other words, EBM is the child of a marriage between methodology and professionalisation. And all its genes are dominant: in order to qualify as evidence-based, a study has to be right methodologically *and* right according to the canonical audience of clinical epidemiologists.

Despite this methodological upbringing, EBM remains uncritical of its own methodology in many ways. This special issue of *Social Epistemology* discusses many of these. Among other things, it describes a pincer movement in which EBM is under threat from complimentary and alternative medicine on the one hand (Derkatch) and

the "basic" biomedical sciences on the other (La Caze). Of course, it is important to realise that these two have traditionally been very differently motivated, and differently executed as well, so that CAM practitioners and laboratory scientists make an uneasy coalition. In any case, we are left with many questions about why EBM has reached its current state, which we hope will be answered in future issues among other places.

Normativity: The Return of the Repressed

So much for the traditional editorial in which we introduce the papers in this special edition of *Social Epistemology* dedicated to EBM. We questioned above, as our authors do, whether the medical community is the right community to decide what EBM is and what it should be. We would now like to begin an answer to that question.

We have already said that a broad coalition of social researchers is more competent to decide the major questions of EBM than any narrowly medical discipline. We would now like to suggest that sociologists in particular are more competent. This might suggest to some that science studies approaches will come to the rescue; and perhaps so. But we are dubious that some of the claims of science studies will be easily applicable to this area (regardless of their merits and deficiencies elsewhere). This can be fixed, and what we propose in this issue can be seen as a beginning in fixing it. What is needed is a critical rhetoric which takes the details of scientific methodology seriously. It is for this reason that we have included technical papers in this special issue (notably those by Dowe and Grossman).

David Dowe's paper is particularly noteworthy from the technical point of view. Dowe discusses a plausible general framework for assessing quantitative evidence from any source and in any discipline. But Dowe's framework is at odds with many of the detailed prescriptions of EBM. (The paper by Grossman highlights this conflict by discussing the case of confidence intervals in more detail, showing that the approach which EBM uses is problematic. Dowe's proposed approach avoids the problems which Grossman presents.) The central place of methodology in EBM means (arguably) that EBM simply is its detailed methodological prescriptions, or at least depends on them for its identity; so (arguably) Dowe's paper functions as a powerful argument against EBM not just as it is currently practiced but as it is and always has been fundamentally conceived. So who is right? We have called Dowe's framework 'plausible': what does this entail? Does it entail that we the readers must take Dowe's arguments on board, evaluate them, and decide on them? Of course this might be anathema to a science studies approach, in which all knowledge is social.

Suppose the disciplinary matrix of science and technology studies takes on board the technicalities raised by Dowe, La Caze, Grossman, and others, as it is well able in principle to do, but suppose it does this on its own terms. Then we will get an externalist story which is different from the doctors'. But who is right? How do we choose between the two (or, please, more) stories? Sadly for those, including us, who see the attraction of pluralism, the choice between alternative stories has to be based on some sort of normative considerations. In the case of EBM, what are they? The doctors at

least have an answer: rationality. However, their sort of rationality has been much criticised from without, and this edition of *Social Epistemology* introduces some new criticisms from within, most powerfully by Dowe and also by La Caze, Grossman, and Boumans. These criticisms from within are what we call the strategy of "hit 'em where it hurts".

Marcel Boumans makes our espousal of normativity more precise by supporting the authority of the normative statistician: someone "with skilled knowledge of statistical reasoning" (Boumans, 390) who isn't afraid to use it to draw normative conclusions. We would go further in the direction of interdisciplinarity and call for an army—or perhaps a rabble—of normative statisticians *with serious interest and expertise in social epistemology*. These normative statisticians would provide expertise in statistical methods, but would also need to be familiar with the vagaries of reasoning in conditions of uncertainty … not just "uncertainty" as modelled by mathematical decision theorists, but real uncertainty in all its methodological, social and political glory.

Such normativity is obviously hard to come by. Several of the papers in this issue ask readers to "take sides" in debates on statistical theory (Dowe, Boumans, La Caze, Grossman). Others ask readers to engage the questions of the role of statistical theory writ large in our theories of knowledge (Derkatch, Boumans, Lipworth et al.). Is it reasonable to believe in the existence of an audience that can do both? We think so. In an important sense, to discuss EBM properly just requires this. What dialogue requires is that social epistemologists and methodologists (especially, in this case, statistical theorists) pay attention to each other's arguments and create spaces where some kind of mutual understanding can occur.

Social Epistemology
Vol. 22, No. 4, October–December 2008, pp. 333–352

Routledge
Taylor & Francis Group

A Couple of the Nasties Lurking in Evidence-Based Medicine

Jason Grossman

The Evidence-Based Medicine (EBM) movement is an ideological force in health research and health policy which asks for allegiance to two types of methodological doctrine. The first is the highly quotable motherhood statement: for example, that we should make conscientious, explicit and judicious use of current best evidence (paraphrasing Sackett). The second type of doctrine, vastly more specific and in practice more important, is the detailed methodology of design and analysis of experiments. This type of detailed methodological doctrine tends to be simplified by commentators but followed to the letter by practitioners.

A number of interestingly dumb claims have become entrenched in prominent versions of these more specific methodological doctrines. I look at just a couple of example claims, namely:

1. *Any randomised controlled trial (RCT) gives us better evidence than any other study.*
2. *Confidence intervals are always useful summaries of at least part of the evidence an experiment gives us about a hypothesis.*

To offer a positive doctrine which might move us past the current conflict of micro-theories of evidence, I propose a mild methodological pluralism: in any local context in which none of a variety of scientific methodologies is clearly and uncontentiously right, researchers should not be discouraged from using any methodology for which they can provide a good argument.

Keywords: Evidence; Evidence-Based Medicine; Methodology; Statistics; Pluralism

Jason Grossman is a Lecturer in the School of Humanities, Australian National University. He has worked in both health policy and biostatistics. Correspondence to: Jason Grossman, School of Humanities, A. D. Hope Building, Australian National University, Acton ACT 0200, Australia. Email: Jason.Grosman@anu.edu.au

ISSN 0269–1728 (print)/ISSN 1464–5297 (online)
DOI: 10.1080/02691720802566961

What is Evidence-Based Medicine?

Evidence-Based Medicine (EBM) is an ideology (by which I mean a simple idea with an associated political programme—something which is no bad thing in itself) which originally stood for the extension of certain parts of the methodology of clinical epidemiology, with its strong emphasis on biostatistics, to the whole of the practice of clinical medicine. Then, from the late 1980s, EBM gradually became the view that those methods ought to be applied to all medical and health-related research and (importantly) health policy. For details of this history, see Daly (2005).

The most famous definition of EBM is the following, by Sackett, one of the two or three people most responsible for the current form of the EBM movement:

> 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)

This definition is inaccurate in at least two important respects. First of all, EBM is and almost always has been, since very shortly after its invention, applied to populations as well as individual patients and to health policy as well as patient care. In this respect, the definition may have been almost accurate when first suggested but is accurate no longer.

Secondly, while it is true that the EBM literature includes exhortations to use the best evidence, it is dominated not by this general rule but instead by a number of *extremely specific* rules. These rules are to be applied rigidly, as I will illustrate below, without any consideration in particular contexts of whether they fulfill Sackett's definition by helping us to use current best evidence. The EBM movement has justified to its own satisfaction (not to mine) the idea that its specific rules will help us to use current best evidence; but having done that, it tells us not to worry about the matter any further, and certainly not to evaluate its rules on a case-by-case basis. It gives us a Catholic doctrine of trusting the Church's rules, without the amelioration of the Catholic doctrine that we may also look to our consciences to judge the rules on a case-by-case basis. We will see this in action in the examples which I will present later.

Despite its inaccuracy, Sackett's definition is widely accepted, presumably on the grounds that one of the main originators of EBM ought to know what it is. It's worth pausing here to recall that Berkeley was tempted to define his subjective idealism as the merest common sense as held by the man in the street. Just as Sackett's definition embodies what Sackett believes to be the most salient facts about EBM, such a definition of subjective idealism as common sense embodies what Berkeley took to be the most salient facts about his theory. But it is clearly no good; at least, not as a definition. A much better definition of subjective idealism would be one which asked more specifically what we had to believe in order to believe it: such as, that objects are bundles of sensations. *Then* we could start arguing over whether it's true.

Similarly, Sackett's definition of EBM is neither good lexicography nor good history but merely Sackett's view of EBM's merits. Just as in the case of subjective idealism, a better definition would list the things that EBM tells us to believe, not in terms of vague success words like "best" but in terms of EBM's specific injunctions.

Unfortunately, EBM is a rag-bag of methodological imperatives with no unifying principle. A proper definition, taking into account everything which EBM actually stands for, would be gut-witheringly, boringly long, and excruciatingly disjunctive. So I will not give a definition. Instead, I will concentrate on two parts of EBM, based on the claims listed above. These are claims which would form part of any realistic definition.

Dumb Claim Number 1

Dumb claim of EBM number 1: *Any randomised controlled trial (RCT) gives us better evidence than any other study.*

Dumb claim 1 is not mentioned anywhere in the motherhood statements of EBM; but it is a central claim of *operationalised* versions of EBM, even when they are presented in the same publications as the motherhood statements.

EBM is operational through and through, because one of its central tenets is that "Systematic and explicit methods of making judgments can reduce errors and improve communication" (GRADE working group 2004, 1490). Consequently, the methods it recommends are more or less rigid and algorithmic. One of the most accessible ways in which the concepts of EBM have been operationalised is through evidence hierarchies. The EBM movement has produced a number of these, most of which are minor variants on each other. I will present four. For simplicity and brevity, I will start by concentrating on a particularly influential hierarchy produced by the US Preventive Services Taskforce. As well as being influential, the US Preventive Services Taskforce hierarchy is also typical.[1] Figure 1 shows the US Preventive Services Taskforce hierarchy alongside two Australian hierarchies, one of which I have analysed in Grossman and Mackenzie (2005). At the end of this section, I will briefly consider a more recent and more complex evidence hierarchy. I will show that it still upholds EBM's dumb claim 1.

The EBM movement actively promotes these evidence hierarchies (more or less interchangeably), and the terminology of "level 1 evidence" has long ago become entrenched in the language of medical research, as illustrated in this extract from the minutes of the National Health and Medical Research Council, the Australian government's most senior health policy committee:

> The draft document [on acute pain management] was introduced to the Council by Professor Michael Cousins, who chaired the working group which prepared the document. … Professor Cousins reported that much of the evidence within the document relates to the efficacy of one form of treatment versus another, there often being options. He also noted the general lack of level I evidence. (National Health and Medical Research Council 1998)

By implication, had level 1 evidence been available it would have been taken to exclude or override other evidence. The analysis in Grossman and Mackenzie (2005) gives more detailed reasons to think that this is the case. The ways in which the EBM literature recommends the hierarchies be used varies, and of course the ways in which they *are* used is another thing entirely. To document either or both of those exhaustively would be a very major undertaking which I have not undertaken fully; but to the extent that I have (both formally in Grossman and Mackenzie (2005) and informally), there

US Preventive Services Task Force (Preventive Services Task Force 1996)	MERGE (New South Wales Department of Health) (Liddle et al. 1996)	Medicare Services Advisory Committee (Medicare Services Advisory Committee 2000)
Level I: evidence obtained from at least one properly randomised, controlled trial. **Level II-1:** evidence obtained from well-designed controlled trials without randomisation. **Level II-2:** evidence obtained from well-designed cohort or case-control analytic studies, preferably from more than one center or research group. **Level II-3:** evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled experiments (such as the results of the introduction of penicillin treatment in the 1940s) could also be regarded as this type of evidence. **Level III:** opinions of respected authorities, based on clinical experience; descriptive studies and case reports; or reports of expert committees.	**Level I:** systematic review of all relevant randomised controlled trials; large multi-center randomised controlled trials **Level II:** one or more randomised controlled trials and studies **Level III:** controlled trials without randomisation; cohorts; case-control analytic studies; multiple time series; before and after studies (preferably from more than one center or research group) **Level IV:** other observational studies	**Level I:** evidence obtained from a systematic review of all relevant randomised controlled trials **Level II:** evidence obtained from at least one properly designed randomised controlled trial **Level III-1:** evidence obtained from well-designed pseudo-randomised controlled trials (alternate allocation or some other method) **Level III-2:** evidence obtained from comparative studies with concurrent controls and allocation not randomised (cohort studies), case-control studies, or interrupted time series with a control group **Level III-3:** evidence obtained from comparative studies with historic control, two or more single-arm studies, or interrupted time series without a parallel control group **Level IV:** evidence obtained from case-series, either post-test or pre-test and post-test

Figure 1 Sample Evidence Hierarchies.

seems to be consensus that the whole point of the hierarchies is to represent strict orderings of study types. Recall the point I made earlier that the EBM enterprise is largely algorithmic (operationalised), and deliberately so, so that decisions about (for example) which papers are worth reading can be automated. The strict separation of study types is part of this decision process.

Dumb Claim 1 is False

This brings us back to dumb claim 1, because dumb claim 1 is equivalent to the strict separation between the top levels of the evidence hierarchy.

What I want to show in the rest of this section is that dumb claim 1 is false. It is *not* the case that any randomised controlled trial (RCT) gives us better evidence than any other study.

The first thing we might note about dumb claim 1 is how strange it is to claim anything about how good the evidence from a study is based solely on the study's *design*.

A study can be designed well but just happen to give useless results, or vice versa; and yet dumb claim 1 contradicts these facts. So dumb claim 1 is in trouble right off the bat. It seems to me that that is a knock-down argument against dumb claim 1. I could stop there. But instead I will put that point to one side, because dumb claim 1 is going to get itself into trouble in quite a number of other ways.

The rest of my strategy will be something like this. I will show that a particular bad RCT was bad, and hence that bad RCTs can be bad. That is perhaps uncontentious. Then I will compare my bad RCT with a good study which was not an RCT. That will show pretty conclusively, I hope, that it is possible for evidence from level 1 of the evidence hierarchy to be worse than evidence from lower down in the hierarchy, contra dumb claim 1.

Each step in this argument will be blindingly obvious, and admitted to be so in the EBM literature whenever particular cases are considered; and yet the conclusion of the argument is contradicted by dumb claim 1, the evidence hierarchy continues to be taken seriously (and rigidly), and policymakers continue to prefer level 1 evidence without enquiring into the quality or outcomes of particular studies. Go figure.

All that remains is to look at a bad RCT and a good non-RCT.[2]

A Bad RCT

> Patients with HIV infection were randomized to received [sic] either a "real" patient held record or a "control" patient held record and prospectively followed. Twenty of the 41 patients randomized to the study had records available for analysis. Eight of the 20 available patients were randomized to the real records and 12 to control records. Both groups were well matched. (Meese, Vujovic, and Fairley 1997, 226)

In this study, 41 HIV-positive people attending a single hospital clinic were given a record which either did (intervention) or didn't (control) hold a summary of their medical record, which they used to facilitate their interactions with the many medical

professionals they subsequently dealt with. The literature suggested that the intervention group should have had improved health outcomes.

This was a bad RCT. One reason it was bad was that it was small. (I will come to why small RCTs are bad in a moment.) Contrary to what one might wish, small RCTs are common. This is for three main reasons: lack of funding, wishful thinking about the outcome, and confusion about the level of analysis. To start with lack of funding: first of all, many RCTs are student projects, which are essentially self-funded. Secondly, while *medical* funding structures encourage researchers to fund projects "properly" or not at all, the same is not true in all disciplines. That's all I can say about funding; much more could of course be said, but not briefly.

To turn to wishful thinking about the outcome: practically all studies have their sample sizes chosen by "power calculations" which depend on the expected outcome. But the expected outcome is often chosen by the experimenters themselves, who have a vested interest in running an affordable study. Consequently, the sample size can be, and presumably often is (and in my own experience certainly is), chosen to be precisely that value which allows the study to be run, even if this makes the study too small to really be useful. (Admittedly, constraints such as peer review can complicate this story.)

Finally, and most interestingly from the social point of view, there is the possibility of confusion about the level of analysis (Evidence-Based Medicine Working Group 2002). I'll illustrate this first with a hypothetical example. Imagine an anti-drug campaign in high schools. The evaluation of the campaign is an RCT which enrolls 6,000 school students, 600 of whom are in each of 10 schools. What is the sample size? 6,000? For some purposes, yes. But for the purposes that matter most, the sample size is 10, because we only sampled 10 schools. The point here is that the sampling which matters most is the sampling across the main type of variation that we're interested in. After all, we would not say that the sample size was 24,000 on the grounds that we had sampled 24,000 students' limbs; and that's because the variation between students' limbs is of no interest in this context. When there's a social variable of interest (as of course there often is), it's *that* variable which should determine the sample size of most interest. In this school campaign the variable of by far the most interest is the school.[3] And in terms of that variable we've only sampled 10 things.

This issue crops up in the real RCT by Meese, Vujovic, and Fairley. The claimed sample size was 20; but that was the number of patients. What Meese et al. were studying was the effectiveness of having the patients carry around summarised versions of their medical records. So there was a very important social variable, namely the local medical systems those patients found themselves in. The real sample size, in terms of that variable, was plausibly the number of hospitals studied. In terms of that variable, the sample size was one.

Even if the sample size was 20, as claimed, that was still rather small, so we had better look at the well-known reason why small randomised studies are bad. The reason is that *randomising, in a small study, is likely to be either pointless or counterproductive*. This is most obvious with a really tiny study (although it's also true with a merely small study like Meese's). Take a study with just four people in it, with an intervention and a control group, and suppose that two of the people are male and two female. Are you

going to randomise the people into the groups? If you do, there's a 50% chance that both males will be in one group and both females in the other. If two of them speak only English and two are bilingual, there's a 50% chance that you'll randomise both bilingual speakers into one group and both monolingual speakers into the other. And so on. You'd have a much better chance of balancing these variables if you did it deliberately, by choosing your groups, than if you left it to randomisation. Maybe you should randomise anyway, in a desperate attempt to balance variables you don't know about. But really you'd be much better off abandoning the study.

This illustrates that in a small RCT the intervention and control groups are likely to be wildly mismatched. When that is the case, it's unreasonable to draw any conclusions at all from the study (although one can amuse oneself by drawing conclusions about the authors instead or, more fairly, about their research culture).[4]

I'm claiming that the intervention and control groups in small RCTs are likely to be wildly mismatched. How can I square this with the statement by Meese et al. that "Both groups were well matched"? Perhaps surprisingly, this is easily done. Epidemiologists like Meese et al. are allowed to make claims of that sort when they've made significance tests of the main variables against the hypothesis that the intervention and control groups are exactly the same. There are two problems with that. (As usual in this paper, these are problems well known to theoretical statisticians.) First of all, we are not interested in whether the intervention and control groups are exactly the same, but rather in whether they are wildly different. Secondly, the very same small sample size which caused me to go into conniptions in previous paragraphs causes these significance tests to be extremely weak, and therefore to fail to reject almost any hypothesis. Putting these two technicalities together, here is what we get. Meese et al.'s significance tests failed to reject the hypothesis that the two groups were exactly the same; they would also (I bet) have failed to reject the hypothesis that they were wildly different.[5] This illustrates the fact that the standard statistical tests to see whether small groups of subjects are well matched are completely meaningless.[6]

That summarises the reasons why small RCTs are bad. And there are other reasons why the study by Meese et al. is a bad RCT. One is that it is a social intervention treated as if it were an individual intervention. Almost whenever this happens, we should expect the effects of the study to "bleed" from the intervention group into the control group. In a little more detail: this is a social intervention because it has its effects (if at all) by affecting a whole social system, consisting not just of a patient but also of a number of medical professionals in interaction with the patient and with each other. If that were not the case, then the patient-held record would not be efficacious. (At least, this is a plausible theory of the basis on which one might expect the intervention to work.) Whatever the effect on the social system is, it may be presumed to work in favour of the people in the control group too, thus (a) distorting the results and (b) weakening any beneficial effects of randomising and further weakening dumb claim 1.

We have not yet finished with Meese et al. Note that they started with 41 subjects but ended up with only eight in the intervention group and 12 in the control group, so there was very substantial dropout from the study. The most interesting and worrying aspect of this dropout is that the analysis of the results treated this dropout as if it

occurred evenly between the intervention and the control groups. But if the dropout from an RCT is substantially different between the groups (note: *not* if it's statistically significantly different, which in a small trial it almost never will be, but if it's substantially different) and if the difference is inexplicable (which it almost always is) then all bets are off: the study is a goner. The problem is somewhat hidden (at least to my eye) by the small sample size, which means that one group has only four more people than the other; but those four people represent a huge proportion of the sample, so the dropout really was substantially different between the two groups, for some unknown reason. But the results of the study consist *entirely* of a comparison between the two groups. And that comparison may claim to be telling us something about the intervention … but really (for all we know or have any right to guess) it's telling us just as much about whatever it was that caused the differential dropout.

So much for an example of a bad RCT, which, despite being bad, was still an RCT, and therefore produced level 1 evidence.

Now let's look at a study which fits into the evidence hierarchy much lower down, giving us only level 2 evidence—according to the hierarchy, much less reliable stuff.

A Good Non-RCT

MacGowan et al. (1997) studied clients of seven methadone maintenance clinics in two States in New England (northeastern USA) in the 1990s. These were people who were being given the drug methadone in the hope that it would reduce their dependence on heroin; approximately 20% were HIV positive. MacGowan et al. wanted to know what the effects of the clinics' counseling and treatment programmes were on the clients' health behaviours, and they wanted to find this out prospectively, by conducting interviews which tracked individual clients over a period of a year. This was an ambitious study, since people who are or have been addicted to heroin are often unwilling to be tracked by researchers over time.

MacGowan et al.'s study was not an RCT. In terms of the US Preventive Services Task Force hierarchy, it *cannot* provide level 1 evidence, and this is something we could know about it *beforehand*, no matter how well the study turned out. (Indeed, it can't even provide evidence at level II-1.) And yet it provides excellent evidence, much better than the results of the bad RCT.

The first interesting difference between the two studies, given my criticisms of the bad RCT we saw earlier, is that MacGowan et al. studied 1,053 people.[7] Indeed, MacGowan et al.'s study seems to have been a response to previous studies which were too small (Calsyn et al. 1982; Landis, Earp, and Koch 1992).

The second interesting difference is that MacGowan et al. obtained detailed ethnographic information from their study subjects in their interviews, including demographic, sexual, and health information relevant to HIV. They were able to match this with clinical records for many subjects. This enabled them to analyse their study results according to a variety of important factors which could have skewed the outcomes. Contrast this with Meese et al.'s simple claim that their intervention and control group were well-matched which, as we have seen, was meaningless.

Neither of these benefits of MacGowan et al.'s study over Meese et al.'s earns them any credit in terms of the evidence hierarchy.

MacGowan et al. discovered that people who remained in methodone treatment and counselling were decreasingly likely, over time, to:

- inject illegal drugs at all;
- inject with dirty equipment; and
- have sexual partners without condoms.

Importantly, these results were obtained using multiple regression, which made some (admittedly fallible) attempt to take into account all the relevant social variables which MacGowan et al. (but not Meese et al.) had measured.

So, to summarise this section and the previous one: we've seen a bad RCT which didn't teach us anything, and we've seen a good study which wasn't an RCT but which taught us a lot. Perhaps not surprisingly, dumb claim 1 turned out to be false.

Does EBM Really Support Dumb Claim 1?

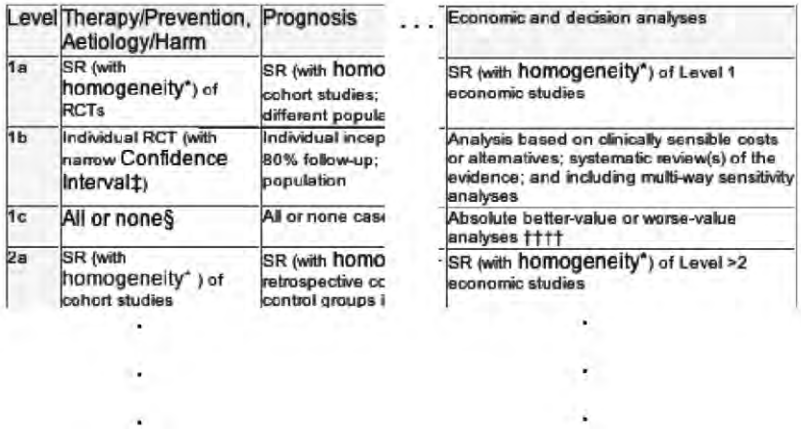
Selective quoting from the vast EBM literature, like selective quoting from the Bible and Shakespeare, can be made to support almost any conclusion. Having shown how silly dumb claim 1 is, one might reasonably guess that I would have to have been selectively quoting in order to make it seem like anyone would ever have supported it.

But unlike Shakespeare (if not the Bible), the EBM literature reaches consensus about a number of things. This is so on a number of levels (most of which I cannot investigate here). One way in which the literature forms a consensus on dumb claim 1 is by returning to it repeatedly over time. I will illustrate this here with a slightly more recent evidence hierarchy.[8]

The Oxford Centre for Evidence-Based Medicine (OCEBM) is perhaps the most thoughtful and careful large group of EBM scholars. It has produced one of a number of evidence hierarchies which improve on the first generation shown above (See Figure 2).

The column headed "economic and decision analyses" is perhaps the one which should interest us the most. But it turns out that the studies which are praised by that column are those which combine economic analyses with "SR's" (systematic reviews)[9] of level 1 studies. Looking across the page, we see that level 1 studies (for the OCEBM) are RCTs of a special type, or alternatively a very rare study called "All or none" which produces an effect in literally 100% of people exposed to an intervention and literally 0% of people denied the intervention.[10] To all intents and purposes, then, level 1 studies are what they were in the simpler hierarchy we've been using all along, and the best economic and decision analyses are level 1 studies with economic analyses tacked on. Almost nothing has changed (although to be sure, what little has changed has changed for the better).

The main change (for our purposes) that's taken place between the first-generation evidence hierarchies and the OCEBM hierarchy is the increase in columns, from one column (all studies) to five columns, dealing separately with studies about "therapy/



Level	Therapy/Prevention, Aetiology/Harm	Prognosis	. . .	Economic and decision analyses
1a	SR (with homogeneity*) of RCTs	SR (with homo cohort studies; different popula		SR (with homogeneity*) of Level 1 economic studies
1b	Individual RCT (with narrow Confidence Interval‡)	Individual incep 80% follow-up; population		Analysis based on clinically sensible costs or alternatives; systematic review(s) of the evidence; and including multi-way sensitivity analyses
1c	All or none§	All or none case		Absolute better-value or worse-value analyses ††††
2a	SR (with homogeneity*) of cohort studies	SR (with homo retrospective cc control groups i		SR (with homogeneity*) of Level >2 economic studies

Figure 2 The OCEBM Hierarchy (excerpt).
Source: Centre for Evidence-Based Medicine, undated.

prevention/aetiology/harm", "prognosis", "diagnosis", "differential diagnosis/symptom prevalence study", and "economic and decision analyses". The diagnosis and differential diagnosis/symptom prevalence columns are similar (for our purposes) to the prognosis column.

Note the medical, clinical focus in the OCEBM terminology (as compared to the earlier terminology of the US Preventive Services Taskforce). The clinical terminology might make it seem at first glance as though the OCEBM hierarchy is limited to hospital interventions. This is not the case: almost every study that's of interest in the broader health field is covered by the OCEBM table … specifically, by the first and last columns, which are (as I have just shown) still committed to dumb claim 1 (except for the *very* minor caveat about all-or-nothing studies and the improvement in sometimes recognising sample size issues).[11]

The clinical focus of the OCEBM table suggests an important point. It would be much easier to argue for an analogue of dumb claim 1 in a very limited clinical context, such as in large drug trials conducted entirely in non-ambulatory patients (where the drawbacks of bad RCTs discussed above are limited because social covariates are *likely* to be *relatively* unimportant, effects are *unlikely* to bleed from the intervention group into the control group, and so on). I would like to thank an anonymous referee of this paper for the idea that when EBM advocates make dumb claim number 1 they actually intend only to make some such narrower claim. Unfortunately, regardless of what they intend, that is not what they say; and, more importantly, they draw *inferences* from dumb claim 1 which belong to the broad claim which I am examining in this paper rather than to the hypothetical narrower one.[12]

One final change which has occurred in the short time since the first generation of evidence hierarchies can be seen in level 1b of the OCEBM evidence hierarchy. Here we see an explicit mention of RCTs which produce narrow confidence intervals. This leads nicely on to dumb claim 2.

Dumb Claim Number 2

Dumb claim of EBM number 2: *Confidence intervals are always useful summaries of at least part of the evidence an experiment gives us about a hypothesis.*

As we saw at the beginning of this paper, at one level of doctrine the EBM movement does not admit to believing in anything much (only "the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients"). But there is a level of nitty-gritty at which EBM really does stand for something. This paper is about a couple of the things it really stands for.

At the nitty-gritty, operational level, EBM stands for the application of certain specific statistical methods. In the eyes of its founding fathers, this presumably does not contradict the idea that we should use the best evidence, since presumably they think that only those specific statistical methods can give us the best evidence. However, there is no more reason to think that they are right about that than there would be to think that any other small group of doctors had once and for all established god's own truth about the best set of specific statistical methods to use for all purposes.

Here I will examine one particularly important set of research methods: confidence intervals. These are recommended by EBM proponents as god's own method of parameter estimation and hypothesis testing.

The EBM literature is full of injunctions to use confidence intervals under all circumstances. And yet it has been known to theoretical statisticians since long before the birth of EBM that confidence intervals sometimes give results which are known *in advance* to be completely useless.

Claims That We Should Use Confidence Intervals

The consolidation of the EBM movement in the late 1980s coincided with a major consolidation among the world's most influential biostatisticians. Almost all at once (in an act of unanimity which is not surprising given the close-knit nature of the international biostatistical community), the major medical journals published papers and editorials asking medical researchers to replace their previous methods of measuring statistical uncertainty (all based on significance tests) by confidence intervals (Langman 1986; Gardner and Altman 1986; Berry 1986; Anonymous 1987). They reasoned that confidence intervals, which associate a pair of numbers with each result instead of the single number given by a significance test, allowed one to estimate the size of a result at the same time as deciding on its statistical significance.[13]

The recommendation was by no means new (confidence intervals are very nearly as old as the significance tests they replace (Neyman 1937)), but confidence intervals had not been very widely adopted up until that point; and so the concerted attempt by journals to make their authors use confidence intervals was something of a watershed. The EBM movement absorbed the new consensus, and made it a sine qua non of "best practice". Since then, the battle for confidence intervals has been more or less won in the mainstream EBM literature, and continues in a set of ever-widening disciplinary

ripples: for example, it is currently being fought in the field of forecasting studies (Armstrong 2007).

The papers cited above introduced confidence intervals to the bulk of medical researchers, either directly or indirectly: they were paraphrased and cited by less prestigious journals as the innovation diffused. (Gardner and Altman (1986) has been cited at least 790 times.) Perhaps importantly, they failed to instill a good understanding of what confidence intervals are in their readers, starting with their very own journal editors. One of the first and presumably most influential editorials, in the British Medical Journal, went so far as to give an interpretation and implicit definition of confidence intervals which owes more to what we'd *like* a confidence interval to be than to what one is:

> …consider a difference in blood pressure between groups of diabetics and non-diabetics of 6 mm Hg with a 95% confidence interval of 1.1 to 10.9 mm Hg.
>
> …we also say using the confidence interval that *there is only a 2.5% chance that the true difference in the population at large is greater than 10.9 mm Hg.* …the conclusion must be that we are unlikely to be missing a large and clinically important difference. (Langman 1986, 716; emphasis added)

Assigning probabilities to hypotheses like that is something which all theoretical statisticians agree confidence intervals can't do. We will see below an illustration of why they can't (at least, an illustration of one of several reasons why they can't). You may wonder what else statistical methods could possibly be for if not to allow us to assign probabilities to hypotheses. Wonder away; parts of the statistical literature wonder with you (notably Savage and discussants 1962). In the meantime, such assignments cannot be made using confidence intervals, and were never meant to be (Neyman 1937; Stuart, Ord, and Arnold 1999).[14]

Confidence Intervals Don't Always Work

In this section, I will use a toy example to show that confidence intervals don't always work. Toy examples are pernicious, of course. But this is a rare case in which the point that needs to be made is purely mathematical, so a toy example is relatively harmless. This example has been well known to theoretical statisticians for many decades. It shows how there can be a perfectly formed confidence interval which makes no sense.[15]

Here is the toy example. Suppose we need to write a protocol for the following common scenario. A patient is admitted to hospital having swallowed three toy academics. A machine that goes beep tells a surgeon that the three toys are stuck in a 2 cm length of the patient's colon. The surgeon would like to operate to remove the three toys. But there is a constraint. One of the toys is a toy Dean. This one has to be removed first. It's guarded on both sides by toy social epistemologists which (fortuitously) protect the colon from toxic paint which flakes off the toy Dean. If the surgeon accidentally removes one of the toy social epistemologists first the patient will die and the surgeon will be struck off. There's a 50% chance that the surgeon can feel the position

of one of the social epistemologists by palpation of the colon, and a 50% chance that she can feel the positions of both of them. She definitely can't feel the Dean.

Assuming the three toys are equally spaced with the Dean in the middle, we can get a 75% confidence interval for the position of the Dean,[16] where x_1 and x_2 are the positions at which the surgeon felt something:

$$\begin{aligned} &x_1 - 1 \text{ to } x_1 - 1 \text{ if } x_1 = x_2 \\ &(x_1 + x_2)/2 \text{ to } (x_1 + x_2)/2 \text{ otherwise} \end{aligned}$$

Why is this a 75% confidence interval? First of all consider the first line of the confidence interval, and suppose that the surgeon could only feel one toy, at position x_1. There's a 50% chance of this happening. Since we're assuming that the toys are equally spaced over 2 cm, with the Dean exactly in the middle, the Dean has to be at $x_1 - 1$ or $x_1 + 1$. We may as well guess that it's at $x_1 - 1$. 50% of the time we'd be right. Multiplying the two 50%s we've got so far, that gives us a 25% chance so far of the confidence interval covering the right answer. The rest comes from the second line. If the surgeon manages to feel both toy social epistemologists then bingo! The toy Dean is definitely in the middle. But there's only a 50% chance of that happening. So altogether there's a 75% chance of the confidence interval covering the right answer. This means it's a valid 75% confidence interval, using the words "valid", "covering" and "confidence interval" the way they're meant to be used in Neyman–Pearson confidence interval theory.

Unfortunately, this 75% confidence interval is completely useless. By the time the surgeon's calculated the confidence interval, she knows for sure whether she's managed to feel one toy or two. (She must, or she wouldn't know the numbers to put into the confidence interval.) If she's only managed to feel one then she knows there's only a 50% chance that the confidence interval contains the position of the toy Dean, not a 75% chance. She should be looking for the missing toy on both sides of the one she's felt, not on the side which the confidence interval seems to recommend. And if she's managed to feel two toys then she knows exactly where the toy Dean is (right in the middle). In that case, to act on the basis of the 75% confidence interval (as if it only had a 75% chance of containing the position of the toy Dean, when actually it has a 100% chance) would be stupid and possibly fatal. Any reader who doesn't like complications should focus on the second possibility. *If the surgeon feels two toys, then she knows for sure where the third toy is, but confidence interval theory, because of the way it sets up its formulas in advance (which is an essential part of the theory), gives her only a 75% confidence interval for its position.*

Recall Langman's incorrect claim in the Lancet that "there is only a 2.5% chance that the true difference in the population at large" is to the right of a 95% confidence interval (Langman 1986, 716). If Langman's interpretation of confidence intervals had been right, there would be a 12.5% chance that the toy Dean was to the right of the 75% confidence interval we've calculated. But by the time the surgeon calculates her confidence interval she knows *for sure* that there is a 25% chance or a 0% chance (depending on whether she felt one toy or two) that the remaining toy is to the right of

the confidence interval. In the case we're considering, Langman's interpretation is not only wrong on theoretical grounds but also numerically way out.

My analysis of this is example is not contentious (although, for some reason, my claim that it has implications for EBM is more contentious). According to one popular reading of the original theory of confidence intervals (a reading which I endorse, by the way, although the death of the author in both senses makes its truth irrelevant for present purposes), confidence intervals were never meant to tell us anything about the value of the parameter in a particular experiment. This is the reading of the theory which statisticians tend to fall back on when pressed to explain examples like this one. All a confidence interval can tell us, they say, is that in experiments similar to this one a certain result will happen 75% of the time, even though we may know *for sure* that it has not happened this time. A fat lot of use that is to the individual patient whom Sackett claimed to have in mind in the quote with which I opened the paper.

One more technical thought is necessary before we finish with the toy example. The problem can be avoided by calculating two separate intervals: one for the possibility that the surgeon feels one toy, and one for the possibility that she feels two toys. That solves the problem in this case, in an ad hoc way; can it be generalised? There are (surprisingly, not very many) more thoughtful advocates of confidence intervals who consider adapting confidence intervals in some such way; but it doesn't work, at least not in general. *Some* such counterexamples go away if one uses a technique known as conditioning on ancillary statistics, but many do not (Leslie 1998). If we avoid all of these sorts of counterexamples by only using confidence intervals when we find out, retrospectively, that they give sensible answers, the whole of confidence interval theory falls down; not only for the plausible reason that theories should not depend on such gerrymandering but also, demonstrably, because the theory essentially depends on not calculating its probabilities retrospectively (Berger and Wolpert 1988).

Confidence Intervals: Conclusions

It has been known to statisticians since at least the 1940s that confidence intervals *sometimes* give results which are known in advance to be completely wrong. I have to suppose that the few supporters of EBM who are aware of this problem take the attitude that only *sometimes* is not a problem. But there is an interesting catch here. There is no general theory which can tell us when a confidence interval will be subject to the sort of problem illustrated above … except for very powerful theories (such as Bayesianism) which not only tell us that but also provide us with replacement estimates and significance tests. I do not have space here to go into these alternative theories in detail, but the details of Bayesianism are well known; work on other theories which might turn out to be equally powerful is in its infancy but an outline is given in (Barnett 1999). So it seems to me (for this and, in fact, for other more contentious reasons which I explore elsewhere) that we ought to be using these other theories to find out when confidence intervals will let us down … and that we then might as well go the whole hog and use those theories as full replacements for confidence intervals. In other words, it seems to

me that any suspicion which the above example might instill that confidence intervals are not the best method of analysing experimental data is entirely correct.

My recommendation of somewhat vague alternative methods is necessarily a little wishy-washy, given the current state of statistical theory and of the philosophy of statistics. But that is not such a terrible thing, nor really such a surprising thing; and to have made a lack of clear-cut methodological rules seem surprising might be the worst thing EBM has done to us.

Methodological Pluralism

I have diagnosed two specific problems with the methodological doctrines of EBM. Perhaps more interestingly, I have diagnosed a problem with the fact that EBM *has* such specific doctrines, or rather that EBM involves applying such specific methodological doctrines so uniformly and over such an immense field.[17]

One possible antidote to the sort of situation we find ourselves in here is to let those whose tendencies run to methodological generalisation put together proposals like the worst excesses of EBM, and then stand on the sidelines and pick holes in them, as I have done in this paper so far. That can be fun, but it seems likely to be ineffective.

Another possible antidote is to propose a more positive stance which can pre-empt methodological overgeneralisations. An obvious antidote along such lines would be methodological pluralism of some sort: that is, some sort of understanding that there is no single best methodology.

In principle there is, presumably, a best methodology if not for a whole field then at least for each research situation. Or so most of us suppose, and so I will allow for the sake of argument. So in principle we should not need to be methodological pluralists.

However, this is not much more use than it would be to say that in principle there is, presumably, a design somewhere in Platonic heaven for a Panacea Machine which would choose perfect treatments for all diseases. What can we conclude from that? Nothing very useful. We don't have such a machine, and in the absence of one we are treatment pluralists.

What do I mean by saying that we are treatment pluralists? First of all, I mean that we believe we should allow different treatments in different situations. That is obvious and uncontentious. But in fact, we are treatment pluralists in a broader sense than that: we (and by "we" I mean all major organised health regulatory systems) allow different doctors to recommend different treatments in what is essentially the *same* situation.

Evidence-Based Medicine is something of a move against treatment pluralism, but it is not completely opposed to it. Even according to the staunchest EBMers, there are many health problems for which it is irredeemably contentious what the best treatment is, at least in the short term, either because nobody knows what the best treatment is or, slightly more controversially but not much, because a variety of people are giving reasons for preferring one treatment or another and there is no clear process for deciding between them. (Not even the staunchest EBMer thinks that there is *always* a clear process for deciding between competing medical arguments, even though EBM is an attempt to find a clear process whenever possible.)

So we (and now by "we" I mean not just regulators but surely all of us) are treatment pluralists in the fairly strong sense that we approve of a social epistemology in which a variety of treatments can be recommended by a variety of authorities even when addressing the same set of medical circumstances ("same" either in the type or the token sense; if you like, the same disease in the same patient).

This suggests what I think is the right sort of methodological pluralism. By analogy with treatment pluralism, we can use "methodological pluralism" to refer to approval of a social epistemology in which different methodologists can recommend different methodologies in what is essentially the same situation. I think that is right. But it would be useful to have a more easily applicable definition, for which I suggest:

> **methodological pluralism** is the view that in any local context in which none of a variety of scientific methodologies is *clearly and uncontentiously right*, researchers should not be discouraged from using any methodology for which they can provide a good argument (using "argument" in the broadest sense).

Note that this is not currently the case: a researcher who wishes to use a methodology not approved by EBM has no social opportunity to present a good argument for doing so. It is certainly not possible, for example, to have a paper published in a medical journal in which one uses a substantially unorthodox statistical methodology to analyse, for example, a clinical trial. At least, so it is widely believed and, of course, the perception is itself part of the problem. In any case, no journal has ever published a major analysis of, for example, a large clinical trial which used a statistical methodology substantially different from p-values or confidence intervals (such as a Bayesian analysis).[18]

This statement of methodological pluralism mentions a judgement of a methodology as being either *clearly and uncontentiously right* for a specific situation or not. This raises the question of who should make such a judgement or, perhaps more interestingly, who should be allowed to enforce such a judgement in the areas in which methodological pluralism matters (editorial policy, funding policy, health service policy etc.). This is a tricky question which unfortunately I do not have space to discuss here, except to mention the obvious option, which is that the judgement as to whether a methodology is *clearly and uncontentiously right* could be made by the same social institutions which currently judge methodological issues in health and medicine (editors, funding bodies, governmental organisations and academic reviewers and consultants). In that case, it would be an obvious requirement of methodological pluralism that the people holding the relevant roles in those institutions, and the mechanisms which those institutions instantiate, do their best to take the arguments in favour of a variety of methodologies at face value, rather than (as they presently do, insofar as EBM is now the norm) privileging one set of methodologies once and for all.

Under this proposal, some improvement in the direction of methodological pluralism would be fairly easy. Slightly unorthodox positions such as the anti-EBM positions I take in my discussions of the examples above are fairly easy to find in the literature if one looks.

However, *some* positions would inevitably be ignored; and to the extent that potentially good methodologies are ignored even by people who are trying to be

methodological pluralists, we would have to judge methodological pluralism to be a failure. A more thorough methodological pluralism could no doubt be achieved by some reform of the institutions and social structures within which methodological judgements are made, perhaps along mildly Feyerabendian lines; but a reasonable discussion of that option would take me too far afield.

I do not know how to defend methodological pluralism in principle. I only know how to defend it in practice. This is like saying that I do not know how to defend treatment pluralism in principle, because I would not be able to defend it if we actually had a Panacea Machine. Lacking a Panacea Machine, in the here and now, we are all treatment pluralists; and all I want to do is defend methodological pluralism for the here and now (and, admittedly, for the forseeable future) in which we do not have a methodology which is clearly and uncontentiously right even in specific, well defined situations, never mind one which is clearly and uncontentiously right across the whole of the type of large category into which EBM likes to sort experimental situations. The bulk of this paper has been an illustration of two ways in which we fail to have any *clearly and uncontentiously right* methodology. In both examples, I have canvassed two methodological views: the view which I claim is held by most proponents of EBM, and my own. Neither view is *clearly and uncontentiously right*. I have shown that the EBM methodology is not *clearly and uncontentiously right*, because it is wrong. But my own view is not *clearly and uncontentiously right* either, not because it is wrong but because it is contentious. In such cases, methodological pluralism is self-evident: all it says is that any methodology which has a good argument going for it should get a fair hearing. In particular, EBM should get a fair hearing, if only on the grounds that anything so widely believed has some chance of being right, no matter how strong my arguments that it is wrong seem to be; and alternatives to EBM should get fair hearings, because of the criticisms of EBM mentioned above and elsewhere. So methodological pluralism is what we need.

Acknowledgements

I would like to thank Adam La Caze, Joan Leach, Fiona Mackenzie, Alison Moore, Dick Parker, an anonymous reviewer for *Social Epistemology*, and the participants in the University of Queensland's Fourth Biohumanities Workshop, for helpful comments on this paper. The University of Queensland's Fourth Biohumanities Workshop was generously supported by the University of Queensland's Biohumanities Program and the University of Sydney's Centre for Time.

Notes

[1] One of the few major differences between the various evidence hierarchies is that some of them have split what they call level 1 evidence into sub-levels, one of which corresponds to "meta-analyses", i.e., statistical amalgamations of a number of experiments on a single issue. This is a good idea. It happens not to affect any of the topics discussed in this paper.

[2] To find my RCT and non-RCT, I have gone back in time a few years. As EBM has taken hold, it has become increasingly hard to find non-RCTs in the peer-reviewed medical literature.

(This in itself tells a story about the success of EBM, of course.) There are still some, and there is some hope that the tide is currently turning (thanks to Fiona Mackenzie for reminding me of this); but in any case, by going way back, I have been able to more or less match my RCT and my non-RCT in subject matter, which aids the comparison.

[3] I'm taking a few things about how school anti-drug campaigns work for granted here; for example, that school students talk to each other.

[4] Sometimes a large number of small RCTs can be combined in a statistical meta-analysis to produce worthwhile results. That does not affect the conclusion of this section, which is merely that some RCTs are bad. Idiosyncratic RCTs, or RCTs which have been conducted sufficiently badly, will never become part of a meta-analysis.

[5] I say "I bet" because I cannot actually test this without access to the raw data.

[6] In any case, as far as I can tell from their paper, the only variables which Meese et al. tested were age and CD4 lymphocyte count, so no amount of statistical analysis could tell whether the intervention and control groups were matched on such important variables as native language. This shows the importance of measuring the right things, something which tends not to be mentioned in the hierarchies of evidence.

[7] As I argued above for Meese et al., MacGowan et al.'s sample size may effectively have been the number of social units, not the number of people, for some purposes. In that case, their sample size was only seven. But even then, they did much better than the bad RCT, because unlike the RCT they were not relying on their sample size to balance variables between the intervention and control groups. If an RCT's sample size is too small because of a confusion between individual and social levels of analysis, all of its results are suspect, because the analysis of an RCT depends on variables being balanced in this way. But if MacGowan et al.'s sample size was too small for the same reason, no such assumption was made, and all that happened was that they may have confused the effectiveness of methodone clinics in general (based on the individual level of analysis, assuming a large sample size) with the correct conclusion, namely that they'd tested some methodone clinics which may have been idiosyncratically good (the social level of analysis, assuming a small sample size of seven clinics with a large sample size within each clinic). What was a fatal mistake in an RCT is a relatively minor mistake here.

[8] The OCEBM hierarchy is hardly more recent than the US Preventive Services Taskforce hierarchy, but the latter is a very late (although otherwise typical) member of an earlier generation of hierarchies, as Figure 1 demonstrates. I did not use the OCEBM hierarchy in my analyses above because it is much more complicated and uses non-standard labelling of its categories. We will see that it tells essentially the same story in any case, except for one important improvement: it does take sample size into account, at least sometimes, by requiring narrow confidence intervals in some cases. See below for a discussion of confidence intervals.

[9] "Systematic reviews" are statistical agglomerations of smaller studies: what I earlier called meta-analyses. There has always been a tendency in England to avoid the word "meta-analysis", which some English people find pretentious. For a while, the English tried to popularise the alternative term "overview"; that failed, so now they are trying "systematic review".

[10] There is a matching category of economic analysis called "Absolute better-value or worse-value analyses" which would apply when, for example, an intervention made something cost more without incurring any benefit.

[11] The OCEBM hierarchy is promoted by a web site which is focused on clinical health care, so the fact that the hierarchy is also to some extent focused on clinical health care is by no means a criticism in itself.

[12] Incidentally, I do not believe that even the narrow claim is entirely correct, although clearly it is vastly more defensible than dumb claim 1 (so much so that I would not call the narrow claim dumb).

[13] In criticising confidence intervals, I do not wish to imply that researchers should return to previous methods of significance testing. Most, and possibly all, of the problems with

confidence intervals apply *mutatis mutandis* to standard hypothesis tests, statistical significance, p-values etc., since (e.g.) p-values are strictly intertranslatable with (isomorphic to) the end-points of confidence intervals, with only a few exceptions. When these exceptions occur, there is great (and interesting) uncertainty in the medical community about which is right, the p-value or the confidence interval. See Grossman et al. (1994) for an example.

[14] There are statistical procedures which *can* assign probabilities to hypotheses, although they have their own drawbacks. The best known are Bayesian procedures. These days Bayesian methods are usually criticised for being excessively subjectivist, but non-subjectivist Bayesian methods exist and, in fact, were one of the main foils against which confidence intervals were originally intended to compete (Neyman 1937).

[15] This example has been discussed to death in the mathematical literature (Welch 1939; Robinson 1975, 1977, 1979; Berger and Wolpert 1988), so I can be sure that there are no lurking infinities or other gotchas.

[16] I say "*a* 75% confidence interval" rather than "*the* 75% confidence interval" to avoid wasting space by discussing the fact that confidence intervals are not unique. Even though they are not unique, the reader can trust me that there is no other confidence interval which makes more sense of this example. Note by the way that there is *no* 95% confidence interval in this case, unless we define one which is even more artificial and problematic than the 75% confidence interval.

[17] I do not claim that it is the fault of the progenitors of EBM that it has ended up covering such an immense field. To some extent, the colonies of EBM may have sucked the colonists onto themselves. This interesting and important question is not my current topic.

[18] Personal communication, Mahesh K. B. Parmar, Cancer Division, Medical Research Council Clinical Trials Unit, 13 July 2007; Donald A. Berry, M. D. Anderson Cancer Center, University of Texas, 15 July 2007

References

Anonymous. 1987. Report with confidence. *The Lancet* 329 (8531): 488.

Armstrong, J. Scott. 2007. Significance tests harm progress in forecasting. *International Journal of Forecasting* 23: 321–7.

Barnett, Vic. 1999. *Comparative statistical inference,* 3rd edition. New York: John Wiley.

Berger, James O., and Robert L. Wolpert. 1988. *The likelihood principle,* 2nd edition. Hayward, CA: Institute of Mathematical Statistics.

Berry, Geoffrey. 1986. Statistical significance and confidence intervals. *Medical Journal of Australia* 144: 618–9.

Calsyn, Donald A., Andrew J. Saxon, George Freeman, and Stephen Whittaker. 1982. Ineffectiveness of AIDS education and HIV antibody testing in reducing high risk behaviors among injection drug users. *American Journal of Public Health* 82: 573–5.

Centre for Evidence-Based Medicine. u.d. Levels of evidence [cited 1 June 2008]. Available from http://www.cebm.net/index.aspx?o=1025; INTERNET.

Daly, Jeanne. 2005. *Evidence-based medicine and the search for a science of clinical care.* Berkeley and Los Angeles: University of California Press.

Evidence-Based Medicine Working Group. 2002. *Users' guide to the medical literature: A manual for evidence-based clinical practice.* Chicago: AMA Press.

Gardner, M. J., and D. G. Altman. 1986. Statistics in medicine: confidence intervals rather than p values: estimation rather than hypothesis testing. *British Medical Journal* 292: 746–50.

GRADE working group. 2004. Grading quality of evidence and strength of recommendations. *British Medical Journal* 328: 1490–7.

Grossman, Jason, and Fiona J. Mackenzie. 2005. The randomised controlled trial: Gold standard, or merely standard? *Perspectives in Biology and Medicine* 48 (4): 516–34.

Grossman, Jason, M. K. B. Parmar, D. J. Spiegelhalter, and L. S. Freedman. 1994. A unified method for monitoring and analysing controlled trials. *Statistics in Medicine* 13: 1815–26.

Landis, S. E., J. L. Earp, and G. G. Koch. 1992. Impact of HIV testing and counseling on subsequent behavior change. *AIDS Education and Prevention* 4: 61–70.

Langman, M. J. S. 1986. Towards estimation and confidence intervals. *British Medical Journal* 292: 716.

Leslie, Claire F. 1998. Lack of confidence: A study of the suppression of certain counter-examples to the Neyman-Pearson theory of statistical inference with particular reference to the theory of confidence intervals. Master's thesis, University of Melbourne.

Liddle, J., M. Williamson, and L. Irwig. 1996. *Method for evaluating research and guideline evidence.* Sydney: NSW Department of Health.

MacGowan, Robin J., Robert M. Brackbill, Deborah L. Rugg, Nancy M. Swanson, Beth Weinstein, Alfred Couchon, John Scibak, Susan Molde, Paul McLaughlin, Thomas Barker, and Rich Voigt. 1997. Sex, drugs and HIV counseling and testing: A prospective study of behavior-change among methadone-maintenance clients in New England. *AIDS* 11: 229–35.

Medicare Services Advisory Committee. 2000. *Funding for new medical technologies and procedures: Application and assessment guidelines.* Canberra: AusInfo.

Meese, Peter, Olga Vujovic and Christopher Fairley. 1997. Randomised trial of a patient held record on the communication with general practitioners in patients with HIV infection. *Venereology* 10 (4): 226–7.

National Health and Medical Research Council. 1998. Report of the 129th Session Sydney, 16 November 1998 [cited 1 June 2008]. Available from http://www.nhmrc.gov.au/publications/reports/129sess.htm; INTERNET.

Neyman, Jerzy. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London, Series A* 236 (767): 333–80.

Preventive Services Task Force. 1996. *Guide to clinical preventive services: Report of the US Preventive Services Task Force.* Baltimore: Williams and Wilkins.

Robinson, G. K. 1975. Some counterexamples to the theory of confidence intervals. *Biometrika* 62 (1): 155–61.

Robinson, G. K. 1977. Corrections and amendments: Some counterexamples to the theory of confidence intervals. *Biometrika* 64 (3): 655.

Robinson, G. K. 1979. Conditional properties of statistical procedures. *The Annals of Statistics* 7 (4): 742–55.

Sackett, David L., William M.C. Rosenberg, J.A. Muir Gray, R. Brian Haynes, and W. Scott Richardson. 1996. Evidence based medicine: What it is and what it isn't. *British Medical Journal* 312: 71–2.

Savage, L. J., and discussants. 1962. *The foundations of statistical inference.* London: Methuen.

Stuart, Alan, J. Keith Ord, and Steven Arnold. 1999. *Kendall's advanced theory of statistics vol. 2A: Classical inference and the linear model,* 6th edition. London: Arnold.

Welch, B. L. 1939. On confidence limits and sufficiency, with particular reference to parameters of location. *Annals of Mathematical Statistics* 10: 58–69.

Social Epistemology
Vol. 22, No. 4, October–December 2008, pp. 353–370

Routledge
Taylor & Francis Group

Evidence-Based Medicine Can't Be…

Adam La Caze

Evidence-based medicine (EBM) puts forward a hierarchy of evidence for informing therapeutic decisions. An unambiguous interpretation of how to apply EBM's hierarchy has not been provided in the clinical literature. However, as much as an interpretation is provided proponents suggest a categorical interpretation. The categorical interpretation holds that all the results of randomised trials always trump evidence from lower down the hierarchy when it comes to informing therapeutic decisions. Most of the critical replies to EBM react to this interpretation. While proponents of EBM can avoid some of the problems raised by critics by suitably limited the claims made on behalf of the hierarchy, further problems arise. If EBM is to inform therapeutic decisions then a considerably more restricted, and context dependent interpretation of EBM's hierarchy is needed.

Keywords: Evidence-Based Medicine (EBM); Randomised Controlled Trials; Frequentist Statistics; Therapeutic Decision Making

1. Introduction

Evidence-based medicine (EBM) proposes that medical decisions be based on the best available evidence. Despite achieving a level of orthodoxy over the past 15 years, EBM continues to be intensely debated in sections of the medical literature (Miles, Polychronis, and Grey 2006). EBM has also recently gained the attention of philosophers of science (Worrall 2002, 2007b,a; Bluhm 2005; Grossman and Mackenzie 2005; Upshur 2005). This has been led in part by an interest in the epistemological claims of EBM, and in part by recognition, from both practitioners and philosophers, that there is much philosophical work to do (Haynes 2002; Worrall 2007a). While there is little to disagree with the claims of EBM at the general level—*of course* medical decisions should be based on the best evidence—the proposal is vacuous without also elucidating precisely what you mean by this evidence and how you propose that it be used. To the extent EBM fills in these philosophical details, it does so by proposing a "hierarchy of

Adam La Caze is a PhD candidate in the Philosophy Department at The University of Sydney. He is completing his PhD on the philosophical foundations of Evidence-Based Medicine. Correspondence to: Adam La Caze, Philosophy Department, Main Quad, University of Sydney, Sydney 2006, Australia. Email: alacaze@mac.com

ISSN 0269–1728 (print)/ISSN 1464–5297 (online)
DOI: 10.1080/02691720802559438

evidence". EBM's hierarchy of evidence is based on how the studies that provide the evidence are designed. EBM suggests that medical decisions be informed by evidence from as high up the methodological hierarchy as possible. Given the extensive practical and political influence of EBM in a wide range of medical decisions, perhaps the most surprising (and worrying) area of philosophical work that is yet to be done is the provision of a clear interpretation and defence of EBM's hierarchy of evidence.

This is not to suggest that aspects of EBM have not been debated. Many aspects of trial methodology have been extensively discussed within clinical epidemiology, statistics and philosophy. The role of randomisation provides one prominent example.[1] There is also an abundance of literature in which EBM is advocated or taught—as opposed to philosophically justified—including, for example, the well known EBM "guidebooks" (Straus et al. 2005; Guyatt and Rennie 2002). Indeed, the social aspects of EBM are important. As Reilly (2004, 991) somewhat foggily remarks, despite "the lack of consensus and clarity about what EBM is", "anyone in medicine today who does not believe it is in the wrong business". While it is easy to parody statements such as this, they represent an underlying reality: many in the medical community are convinced of the merits of EBM, even if it is not clear yet precisely what EBM is. What is missing is a systematic justification of EBM's methodological hierarchy that survives critical analysis.[2]

EBM puts forward the methodological hierarchy as a tool for making good medical decisions. Ideally, any justification of EBM needs to, first, describe how the hierarchy should be applied, and, second, justify how this application of the hierarchy improves medical decision making. This paper focuses of the first part of this task for EBM. Proponents have made bold claims about what can be achieved by making decisions in accordance with the EBM hierarchy. Specifically, randomised trials are seen to provide an especially secure form of evidence.

> Because the randomised trial, and especially the systematic review of several randomised trials, is so much more likely to inform us and so much less likely to mislead us, it has become the "gold standard" for judging whether a treatment does more good than harm. (Sackett et al. 1996)

However, as the philosophical criticisms show, not all the claims made by proponents of EBM on behalf of randomised trials can be justified (Grossman and Mackenzie 2005; Worrall 2007a, 2002). I wish to extend this criticism in a particular way. Advocates of EBM propose that medical decisions—and even more specifically *therapeutic* decisions—are better informed by reference to the evidence hierarchy. I show that the interpretation of EBM's hierarchy that is most often put forward by proponents cannot be justified.

An unambiguous interpretation of the hierarchy has not been provided. Early papers, and the EBM "guidebooks", provide the clearest account. On this account, the hierarchy is interpreted categorically. The categorical interpretation of the hierarchy holds that evidence from higher up the hierarchy trumps evidence from lower down. I describe this interpretation in section 2. The philosophical treatments of EBM are examined in section 3. These accounts respond to the clear, but somewhat simplistic, view of EBM

that has been provided, and expose its problems. Ambiguity about how the hierarchy should be interpreted, however, gives proponents of EBM some "wriggle room". Restricting the claims of EBM, by explicitly narrowing the domain of application, and accepting that randomised trials are fallible, avoids some of the criticisms that have been raised. In the final section, I show that even if these moves are made, the categorical interpretation cannot be justified. And moreover, that imposing any further limits impedes the application of the hierarchy to therapeutic decisions. Hence, the paper is predominately negative. If EBM is to inform therapeutic decisions, the hierarchy cannot be interpreted as proposed by advocates.

2. EBM according to the advocates

EBM's history is recent, and localised. It developed as a distinct approach to medical practice and education within the Department of Medicine and Clinical Epidemiology and the Department of Biostatistics at McMaster University, Canada, during the 1980s and 1990s. The McMaster faculty involved in disseminating the central ideas of EBM, including David Sackett, Gordon Guyatt, David Haynes, and Deborah Cook, continue to be prominent among EBMs leading proponents. Guyatt, who first coined the term, suggests "Evidence-Based Medicine" is a development of David Sackett's notion of "bringing critical appraisal to the bedside", referring to the application of clinical epidemiological skills—in particular an understanding of the strengths and weaknesses of research methods—to the problems experienced by patients presenting at the clinic (Guyatt and Rennie 2002).[3]

The first paper to outline EBM in detail perhaps best illustrates how proponents conceive EBM.

> A new paradigm for medical practice is emerging. Evidence-based medicine deemphasizes intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient grounds for clinical decision making and stresses the examination of evidence from clinical research. Evidence-based medicine requires new skills of the physician, including efficient literature searching and the application of formal rules of evidence evaluating the clinical literature. (Evidence-Based Medicine Working Group 1992)

EBM is seen as a move from basing medical decisions on the "unsystematic" judgment of an individual clinician, based on experience or the findings of the bench or basic sciences, to the more "systematic" and "relevant" outcomes of patient-related clinical research. The "basic" or "bench sciences" are physiology, pharmacology, and related disciplines such as pathophysiology. These sciences are primarily focused on developing a theoretical basis for how the body works (physiology), how drugs interact with physiological processes in the body (pharmacology), and the physiological abnormalities involved in disease (pathophysiology). Theory-based sciences such as these provide a contrast to the empiricism of EBM.

Proponents insist on labeling EBM a Kuhnian paradigm shift in medicine.[4] But they are using "paradigm" more informally than Kuhn.[5] The continued insistence that EBM is a paradigm shift simply illustrates the conviction of proponents that there is a marked distinction between EBM and the pre-EBM process of medical decision

making. EBM's key claim is that good medical decisions involve the appropriate interpretation of evidence:

> Understanding certain rules of evidence is necessary to correctly interpret literature on causation, prognosis, diagnostic tests, and treatment strategy. (Evidence-Based Medicine Working Group 1992)

The "rules of evidence" are provided by EBM's methodological hierarchy.

EBM puts forward different hierarchies for different types of medical decisions. Hierarchies have been provided for decisions relating to therapeutic decisions, prognosis, diagnosis, symptom prevalence, and economic and decision analyses; each relying on similar methodological distinctions (Guyatt and Rennie 2002; Phillips et al. 2001). I focus on the hierarchy provided for treatment and harm. EBM's largest influence has been on therapeutic decision making. (Later, I show there are good reasons to restrict EBM's claims to therapeutic decisions).

Being specific about what therapeutic decisions entail is important to this analysis. By "therapeutic decisions" I mean both population and individual therapeutic decisions. Population therapeutic decisions rely on answering the question of whether the benefits of a particular medical therapy outweigh its harms in a defined population of patients. Such a population typically being defined in terms of average age, condition being treated, and presence of co-morbidities. Individual therapeutic decisions, by contrast, focus on the question of whether the proposed benefits of a particular medical therapy outweigh the possible harms in an individual patient, given his or her unique characteristics.

A number of hierarchies have been proposed for therapeutic decisions, but the differences between them are primarily in the level of detail. See Table 1 for the hierarchy provided by Guyatt and Rennie (2002, 12)[6] (for a more detailed version, see Phillips et al. (2001)).

The hierarchy highlights the distinctions important to EBM. "Systematic" *experimental* evidence is valued higher than "unsystematic" clinical experience. Of the experimental evidence, patient-related *clinical* evidence—that is, direct experimental evidence of the effects of treatments on patients—is valued higher than experimental evidence from the basic sciences. And finally, experimental evidence from clinical studies is distinguished according to *methodology*: randomised studies, and systematic reviews of

Table 1 The Hierarchy of Evidence Supplied by Guyatt and Rennie (2002, 12)

A Hierarchy of Strength of Evidence for Treatment Decisions
N of 1 randomised controlled trial
Systematic reviews of randomised trials
Single randomised trial
Systematic review of observational studies addressing patient-important outcomes
Single observational study addressing patients-important outcomes
Physiologic studies (studies of blood pressure, cardiac output, exercise capacity, bone density, and so forth)
Unsystematic clinical observations

randomised studies, are claimed to provide better evidence than non-randomised, or "observational" studies.

Randomised studies permit investigators to impose an intervention on participants in the study. Participants are recruited and then randomised to treatment or control and monitored for differences in outcomes. By contrast, observational studies follow subjects who are going about their lives, choosing (as much as is possible) which medicines they take and to what risk factors they expose themselves. Observational studies may be prospective or retrospective with respect to the events under investigation. I will refer to a study as "prospective" if patients are entered into the study and observed as events occur, and "retrospective" if all the events under investigation occurred prior to the start of the trial. The two main forms of observational studies are cohort, and case-control. Cohort studies observe two groups of participants: one group exposed to the risk under investigation (such as a treatment or environmental pollutant), and a second group not exposed. These "cohorts" are then followed to see if the outcomes differ between the groups. Case-control studies begin at the other end of the timeline, that is, once an event (or "outcome") has occurred (for example, a heart attack or a diagnosis of cancer). The group for which the outcome has occurred, the "case" group, is compared to a control group—a group for whom the outcome under investigation has not occurred. The two groups are compared according to their exposure to the risk factors (or treatments) under investigation in an attempt to isolate the cause of the event.

According to proponents of EBM, *experimental* evidence is superior to *non*-experimental evidence, *clinical* experimental evidence is superior to *non*-clinical experimental evidence, and *randomised* clinical experimental evidence is superior to *non*-randomised clinical experimental evidence. But how is this superiority achieved? To answer this question it is first necessary to examine how EBM applies the methodological hierarchy. In the account provided by the EBM guidebooks the notion that randomised studies trump evidence from lower down the hierarchy is central.

> If the study wasn't randomised, we 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 randomised; if it isn't we can bin it.) Only if you can't find any randomised trials should you go back to it. (Straus et al. 2005, 118)

> The hierarchy implies a clear course of action for physicians addressing patient problems: they should look for the highest available evidence from the hierarchy. The hierarchy makes clear that any statement to the effect that there is no evidence addressing the effect of a particular treatment is a non sequitur. The evidence may be extremely weak—it may be the unsystematic observation of a single clinician or a generalisation from physiologic studies that are related only indirectly—but there is always evidence. (Guyatt and Rennie 2002, 14–15)

These quotes show that EBM has a broad concept of "evidence"; results of randomised trials do not constitute the only source of evidence. But, equally, EBM has a narrow conception of what provides the "best evidence". According to EBM, when it comes to therapeutic decisions the "best evidence" is provided by the results of randomised studies. Thus, the EBM guidebooks suggest a *categorical* interpretation of the hierarchy.

On the categorical interpretation, randomisation is seen to provide an incontrovertible epistemic good. The results of randomised studies are epistemologically superior to the results of non-randomised studies, and the superiority is absolute. *All* the results of a randomised study are *always* superior to the results of studies from lower down the hierarchy—at least, for all those studies that are conducted that meet the standards of publication. How else could it be appropriate to "bin" all non-randomised studies relating to the therapeutic question we are investigating?

3. The Critic's View of EBM

The philosophical criticisms of EBM have focused on different aspects of the approach, but each respond to a similar view of the hierarchy (Bluhm 2005; Grossman and Mackenzie 2005; Worrall 2007a, 2002, 2007b). Not surprisingly, the shared view is the one most clearly articulated in the EBM guidebooks. That is, that EBM's hierarchy should be interpreted categorically. It is possible to summarise the critical response into a number of broad themes. How some, but not all, of these criticisms may be avoided by proponents of EBM is discussed in the sections that follow.

Worrall (2007b, 452), examines the notion that randomised studies provide especially secure knowledge in medicine.

> It is widely believed that RCTs carry special scientific weight—often indeed that they are *essential* for any truly scientific conclusion to be drawn from trial data about the effectiveness or otherwise of proposed new therapies or treatments. This is especially true in the case of clinical trials: the medical profession has been overwhelmingly convinced that RCTs represent the "gold standard" by providing the only "valid", unalloyed, genuinely scientific evidence about the effectiveness of any therapy. (Worrall 2007b, 452)

Worrall shows that the benefits of randomisation fall short of making randomised trials "essential" in the sense EBM often takes them to be. Contrary to what is often claimed, Worrall shows that randomisation does *not* ensure that all confounding factors, known and unknown, are equally balanced in the experimental groups. While randomisation has some benefits, such as preventing some types of selection bias, it certainly does not ensure infallibility. Nor, Worrall argues, does randomisation justify the *very special* scientific weight proponents of EBM place in randomised trials.

Jason Grossman and Fiona Mackenzie (2005, 523) also highlight the fallibility of randomised trials. In addition they illustrate the problems of measuring the quality of evidence according to a single methodological criteria:

> […][W]hen one attempts to follow the guidelines, one discovers that whether or not the intervention in question is amenable to RCTs, if no RCTs have been performed the evidence obtained can never be better than level III. That is, even the most well-designed, carefully implemented, appropriate observational trial will fall short of even the most badly designed, badly implemented, ill-suited RCT. (Grossman and Mackenzie 2005, 523)

Clearly, when evaluating evidence, much more needs to be considered in addition to whether a trial was randomised. (Notably, this is one criticism that is increasingly recognised in the medical literature (Glasziou, Vandenbroucke, and Chalmers 2004; Guyatt et al. 2008a; The GRADE Working Group 2004)).

Robyn Bluhm (2005) also reacts to a categorical interpretation of EBM's hierarchy. But her focus is directed towards the question of how broadly the hierarchy should be applied. In particular, Bluhm is concerned that *epidemiology* relies on the basic sciences. If EBM's hierarchy is applied broadly (say to all of science), then the basic or bench sciences are seen to be "lower" forms of evidence. But the basic sciences are essential for discovering "effective causal interventions in the course of a disease in individual patients"—most certainly a key aim of epidemiology (Bluhm 2005, 543). Grossman and Mackenzie are also concerned about how broadly EBM's hierarchy is thought to apply. In particular, Grossman and Mackenzie are concerned about the application of EBM's hierarchy to public health policy.

> In recent years this preference for RCTs has extended beyond medicine, with researchers swept up with the ideals and methods of EBM in the promise of scientific recognition and increased funding. One important area in which this has happened is the evaluation of public health interventions, where (to take one example) a food policy program, evaluated observationally, has little chance of being accepted as effective, no matter how effective it actually is, and consequently has no chance of securing the sort of government funding available to phase III drug trials, even though food policy is probably more important to population health than all of these drug trials put together. (Grossman and Mackenzie 2005, 517)

The categorical interpretation of EBM's hierarchy also creates problems for external validity.[7] "External validity" refers to the extent results of a clinical trial can be generalised to patients other than those involved in the study. The problem arises because of the importance of the basic sciences in *interpreting* (and thus generalising) the results of randomised trials.

> Because RCTs tend to report only average results in the treatment and control groups, the extent and sources of within group variability are not known. Both extrapolation of the results of an RCT to other patient groups and an understanding of the reasons for differences in outcomes within the study group require a knowledge of biological factors that may influence the effectiveness of a drug. This type of information, however, cannot come from epidemiological studies alone. Rather, it is often first discovered in the context of physiological studies on humans or animals (the second lowest level of evidence in the hierarchy) and of unstructured clinical observation (the lowest level). (Bluhm 2005, 537)

Randomised trials examine the effects of a therapy in a very small sample of the patients who will eventually receive the drug. Often, though not always, the sample of patients that are included in trials are highly selected; they are considerably younger and suffering less comorbid illness. Applying the results of randomised trials to individual patients raises questions of extrapolation and interpolation. If the trial was highly selective in its sample population it can be difficult to know whether the results of the trial extends to patients in routine care. And, for less selective trials, it can be difficult to know whether an individual patient, who resembles the individuals in the trial, would have been among the proportion of patients who benefited from the therapy under investigation.

To the extent these questions can be answered, they rely on the basic sciences. Extrapolating the findings of a randomised trial to a patient under routine care often

relies on a judgment of whether the patient's physiological characteristics are similar in relevant respects to patients included in the trial sample. If the patient under routine care is judged to be similar to the sample population, then there is a good chance the results of the trial can be extended to this patient. A judgment that the results of the trial do not extend to an individual in the clinic is often due to the physiological characteristics of the individual—for instance, the patient may suffer a comorbid illness that will reduce the effectiveness of the therapy (or increases the risk of adverse effects).[8] While the problem of external validity is acknowledged within EBM, interpreting the hierarchy categorically makes it intractable. Extrapolating the findings of a randomised trial requires a comprehensive understanding of the basic sciences. If the evidence provided by the basic sciences is as poor as their place in EBM's hierarchy suggests, then there is no principled way to apply to the results of these trials to patients.[9]

In general, proponents of EBM have elected not to engage with criticism directly (Buetow et al. 2006). Instead the account of EBM provided by proponents has subtly shifted over time.[10] Because of the lack of direct debate, and the absence of a rigourous defence of EBM's epistemological claims, pinning down EBM's "current" view is difficult. There is certainly enough "wriggle room" within EBM to avoid some of the criticisms discussed in this section. The view of EBM that would result, however, is considerably more complex, and yet to be adequately explicated by proponents. I now examine how proponents of EBM can legitimately avoid some criticisms by suitably restricting their claims (while maintaining the primary aim of EBM as informing therapeutic decisions). This can be done by refining how the hierarchy is interpreted. Importantly, while some criticisms can be avoided by suitably restricting EBM's claims, the resolution of other problems comes at a cost to EBM's central aim of informing therapeutic decisions.

4. EBM Can't Be: How the Hierarchy Can't Be Interpreted

Two criticisms of EBM can be addressed, at least in part, by recognising that the hierarchy does not provide general epistemological rules. The domain of application for the hierarchy should be limited to the context for which it was developed: therapeutic decisions. This addresses Bluhm's concerns, and, less directly, provides an avenue for proponents to respond to Worrall's concerns regarding the *very special* weight EBM places in randomised trials. Restricting EBM's claims to therapeutic decisions, however, is not enough. As much as EBM proposes an interpretation of the hierarchy, it is a *categorical* interpretation. According to this view, when looking for evidence to inform a therapeutic decision if it doesn't come from a randomised study "bin it". This view fails to acknowledge the complexity of the results provided by randomised trials.

4.1 The EBM Hierarchy Does Not Provide General Epistemological Rules

Much rhetoric about EBM gives the impression that the hierarchy provides some general epistemological rules for all of science. It is the implicit assumption that the hierarchy provides such rules that fuels claims that the highest levels of the hierarchy

provide especially secure evidence, and gives the impression that the hierarchy can be broadly applied. If EBM's hierarchy provides general epistemological rules, it would be expected to hold independent of context (or at least hold in a range of contexts defined by some general principles). On this view, randomised studies would provide superior evidence to that of the "basic sciences" in all (or at least many) scientific disciplines, not just clinical science.

It only takes a moment's reflection to see that this is simply false. Many sciences progress, in whole or in part, without randomised studies. Much of physics, for instance, does just fine without randomised studies. Rather, if it makes any sense, EBM's hierarchy makes sense in the context of *therapeutic* decisions. While I do think a philosophical account of the hierarchy can be provided, it is far from general. Any account of evidence in medicine will be highly dependent on the specific context of the clinical sciences. Importantly, EBM proponents, when pushed, accept this limitation on the range of application of the evidence hierarchy.

"Evidence-Based Medicine: What it is and what it isn't", is a reply by proponents of EBM to criticisms of the approach (it is one of the few papers in which proponents engage, indirectly at least, with criticism) (Sackett et al. 1996). This paper responds to claims that EBM focuses exclusively on randomised trials and meta-analyses. The reply is telling. It makes clear that many types of medical decisions do not require randomised trials. Questions of prognosis or the accuracy of a diagnostic test, for instance, are answered by non-randomised studies. It is only *therapeutic* questions that require randomised studies.

> It is when asking questions about therapy that we should try to avoid the non-experimental approaches, since these routinely lead to false positive conclusions about efficacy. Because the randomised trial, and especially the systematic review of several randomised trials, is so much more likely to inform us and so much less likely to mislead us, it has become the "gold standard" for judging whether a treatment does more good than harm. However, some questions about therapy do not require randomised trials (successful interventions for otherwise fatal conditions) or cannot wait for the trials to be conducted. And if no randomised trial has been carried out for our patient's predicament, we must follow the trail to the next best external evidence and work from there. (Sackett et al. 1996, 71–2)

This quote reinforces the categorical interpretation of the hierarchy, but makes clear that the focus of the hierarchy is therapeutic decisions. While some therapeutic decisions may occasionally have to be made on the basis of alternative evidence, if the results of a randomised study are available, then the decision should be based on them.

Given the rhetoric that is sometimes employed, it is not surprising that some have interpreted proponents of EBM to view the hierarchy as providing general epistemological rules, but it is an over-reach. EBM's central claim is that evidence from study designs featured higher up the hierarchy more reliably informs *therapeutic decisions. If* experimental results from the basic sciences or observational studies are inferior to evidence from randomised studies, it is only in terms of therapeutic decision making.

This limits EBM's claims considerably. And, it provides a response to Bluhm's concern regarding EBM undermining the importance of the basic sciences to epidemiology. The hierarchy simply does not extend that far. It should be applied only when

considering the question of whether a particular therapy benefits a patient, or group of patients, more than it is likely to harm. The task of documenting the incidence, and discovering the cause, of a disease need not refer to EBM's hierarchy of evidence. This, of course, is implied by the differing hierarchies provided by proponents of EBM (Phillips et al. 2001). But is not made explicit enough in many discussions of EBM.

Recognising that EBM's hierarchy does not provide general epistemological rules also opens some avenues for proponents of EBM to respond to Worrall's concerns. Worrall (2002, 2007b,a) shows that randomisation does not provide any guarantee of the results of a randomised trial. Randomisation does not ensure experimental groups are equally balanced for all confounding factors. Recognising that randomisation is not essential generally opens the way for a much more limited—and thus more plausible—defence of randomisation in the context of therapeutic trials. Indeed, limiting EBM's claims in this way underlines the need for a positive account of why randomised trials are needed for therapeutic questions (Worrall (2007a) has shown this is yet to be provided by advocates of EBM).

Limiting EBM's claims to the context of therapeutic decisions also provides a response to the emerging epidemic of "evidence-based" disciplines. If a clear, and justifiable, interpretation of EBM is yet to be provided in the very context it was designed for then the plight of these second-generation "evidence-based" disciplines is not promising.

The "evidence-based" moniker has been extended to other areas of practice, such as nursing and pharmacy, other areas of health decision making, such as public health interventions, as well as a quickly increasing number of disciplines outside healthcare, including evidence-based policy making. Though some do, not all of these second-generation "evidence-based" disciplines explicitly import EBM's hierarchy along with its moniker. Whether or not they import EBM's hierarchy, the "evidence-based" claims of these disciplines are either problematic, or at best, unclear. When they do import the EBM hierarchy, such as in evidence-based nursing, pharmacy and public health, it is usually a case of "if it is good for medical decision making, then it is good for us". In these situations, the EBM hierarchy is being extended to the decisions of interest to the discipline. Any use of EBM's hierarchy of evidence outside of therapeutic decision making is going to need an independent justification for the scientific context to which it is to be applied. This is not to say it can't be done. Some areas of these other disciplines may be similar enough to therapeutic questions so as to justify use of the hierarchy. But, a justification is needed. Furthermore, for some questions within these new "evidence-based" disciplines the hierarchy is simply inappropriate. As already discussed, one example is the application of EBM's hierarchy to some public health interventions. Grossman and Mackenzie (2005) show that randomised trials are ill-suited to appropriately address some research questions within public health. But, due to the hegemony of EBM's hierarchy, methodologies that are well suited to address the research question are being ignored, or automatically and inappropriately downgraded. Conversely, when disciplines take on the evidence-based moniker without importing EBM's hierarchy, such as the way "evidence-based policy" is often used, then it is difficult to see what work the moniker is doing (other than sounding vaguely reassuring).[11] EBM without its

hierarchy is meaningless. So too are other uses of the moniker without some explicit expression of what "evidence" is being referred to, and how it is being used.

4.2 The EBM Hierarchy Can't Be Interpreted Categorically

Recognising the hierarchy does not provide general epistemological rules, and limiting application of the hierarchy to therapeutic decisions, provides an avenue of response to some criticisms of EBM. But not all. EBM's account of how the hierarchy should be put into action relies on a categorical interpretation. When searching for evidence to inform a therapeutic decision:

> If the study wasn't randomised, we'd suggest that you stop reading it and go on to the next article in your search. (Straus et al. 2005, 118)

This does not suggest an interpretation of the hierarchy where only particular well defined questions are best answered by randomised studies. The categorical interpretation suggests that when it comes to therapeutic decisions *all* of the results of a randomised trial *always* trump evidence from lower down the hierarchy. Evidence from observational studies may sometimes be needed to help inform therapeutic decisions, but only in the absence of a randomised trial, and only then, when the considerably "weaker" strength of this evidence is emphasised.

> Without clear confirmatory evidence from large-scale randomised trials or their meta-analyses, reports of moderate treatment effects from observational studies should not be interpreted as providing good evidence of either adverse or protective effects of these agents (and, contrary to other suggestions, the absence of evidence from randomised trials does not in itself provide sufficient justification for relying on observational data). (Collins and MacMahon 2007, 24)

While the categorical interpretation is relatively straightforward, it simply can not be sustained.

First, as has already been discussed, the categorical interpretation equates quality of evidence with a single aspect of methodology. Many aspects of clinical trials affect the quality of the evidence they produce, not simply whether or not they are randomised (Grossman and Mackenzie 2005). This is one criticism that the medical literature is responding to. The recently developed GRADE system for evaluating the quality of evidence explicitly recognises that randomisation is only one measure of quality (Guyatt et al. 2008b,a).

The second problem for the categorical interpretation holds even for the best designed (and implemented) randomised trials. The categorical interpretation fails to distinguish between the different types of "results" furnished by randomised trials. Randomised trials supply many "results"; however, the warrant for each of these results is far from equal—even by EBM's reckoning. Randomised studies are designed (statistically and methodologically) with a particular question in mind. Most often (in the studies of interest in EBM) the question is whether a given therapy will have a beneficial effect on a defined outcome in a defined group of patients. For example, a randomised study might examine whether aspirin reduces the rate of death in patients who are

admitted to hospital suffering from acute coronary syndrome. The question for which the trial has been designed is called the primary hypothesis, and the outcome of interest to this hypothesis, the primary outcome or endpoint. In addition to the primary hypothesis there are usually two to three secondary hypotheses and related endpoints. These secondary hypotheses often relate to other benefits the therapy may have, as well as harms the therapy may cause. For example, with regard to the aspirin trial, secondary hypotheses may relate to whether aspirin reduces angina pain, and whether it increases the risk of bleeding. Therapeutic decisions often rely on (or at least need to incorporate) the results of secondary endpoints. After all, any therapeutic decision requires an *overall* assessment of both the benefits and harms of the therapy.

The results of an intervention on subgroups within the trial are also important to therapeutic decisions. For instance, regarding the aspirin trial above, a clinician with an elderly female diabetic patient will be particularly interested in the results of the intervention in the relevant subgroups; the female patients, the elderly patients and the diabetics. Subgroup analyses raise a number of thorny issues for the appropriate analysis and interpretation of randomised trials, and there is a range of views on the matter.[12] However, whichever view is taken with regard to the appropriate analysis of subgroups, it is undeniable that they provide evidence of importance to therapeutic decisions. This results in an

> …unavoidable conflict between the reliable subgroup-specific conclusions that doctors and their patients want, and the unreliable findings that subgroup analyses of clinical trials might offer. (Collins and MacMahon 2007, 13)

Subgroup analyses and analyses of secondary endpoints, together with the results from the primary hypothesis test make up the "results" of randomised studies. Any interpretation of the hierarchy needs to acknowledge the different warrant these results provide.

Randomised trials are analysed according to frequentist statistics. The methods for hypothesis testing and estimation proposed by Jerzy Neyman and Egon S. Pearson are particularly influential.[13] Within these methods, *power* plays a vital role in establishing the warrant of the statistical test. From a pre-trial perspective the role of power is not contentious. "Power" is the pre-test probability that the statistical test will "reject" the null hypothesis, on the assumption that the null hypothesis is false. Much effort is taken to ensure that the primary hypothesis test is sufficiently powered. Trials that are not sufficiently powered to test the primary hypothesis are often refused funding, or not given ethical approval. This is because underpowered trials are less likely to provide "definitive" results according to the dictates of frequentist statistics—that is, a result that "rejects" the null hypothesis. Statistical tests on secondary hypotheses and subgroup analyses, however, are often underpowered.[14]

Once the results of a trial have been observed the role of power is considerably more contentious. However, it is well recognised that the observed results of a trial are less reliable when the size of the trial is small relative to the true size of effect under investigation. Underpowered tests can result in false negative results—that is, fail to reject a false null hypothesis. After all, low power predicts—from a pre-trial perspective on the

assumption the null hypothesis is false and a given size of trial—that observing a statistically significant result is unlikely. Somewhat less well recognised, but just as important, a low powered test can also result in false positive results. If the true effect size is small, and the power of the test for this small effect is low, then any result that is statistically significant will over-estimate the effect size. Land (1980) provides a description of this phenomenon, and uses it to explain 100-fold discrepancies in estimation of cancer risks due to low-dose radiation. In this sense—that is, the possibility of false negative, or false positive, results—the results of subgroup analyses and analyses of secondary endpoints are unreliable (when they are underpowered). The unreliability of the results of subgroups and secondary endpoints, coupled with the importance of these results to therapeutic decisions, undermines a simplistic categorical interpretation of the EBM hierarchy.

To be clear, I am not suggesting that proponents of EBM do not recognise that the results of primary hypothesis tests have a different warrant to the results of secondary hypotheses and subgroups analyses. On the contrary, they will be the first to point out the differences. It is not contentious that these different types of results have a different epistemic standing.[15] Moreover, EBM has a fairly standard reply to the problems of subgroups analyses and analyses of secondary endpoints: await the results of meta-analyses. In the ideal case when you have a number of high quality randomised trials that include similar enough patients and test the same treatment, meta-analysis will improve the reliability of subgroup analyses and secondary endpoints. The results of meta-analyses, however, are not always available, and when they are the realities of clinical research can undermine the improved reliability achieved in ideal circumstances (Egger et al. 1997; Egger and Smith 1995). More importantly, for our purposes, none of this is recognised by proponents of EBM when they describe how the hierarchy should be applied. The categorical interpretation is the most straightforward interpretation of EBM's hierarchy that is provided by proponents. But it fails to acknowledge that some results of randomised studies are unreliable—and because subgroup analyses and secondary endpoints are of particular interest to therapeutic decision makers, the unreliability of these results presents a particular problem for EBM.[16]

This suggests a second limitation is needed to further restrict application of EBM's hierarchy of evidence. Not only does the hierarchy need to be restricted to therapeutic questions, but within therapeutic questions, application of the hierarchy should (at best be) limited to the results of primary hypothesis tests and well-conducted meta-analyses, as it is only for these tests that the optimal warrant of frequentist statistics is provided. While this is a positive move for EBM, as it provides a more justifiable interpretation of the hierarchy, there is a cost. EBM no longer fulfils its aim of informing therapeutic decisions. Recall, therapeutic decisions rely on assessing the benefits and harms of therapies for groups of patients and individuals. Limiting the hierarchy to the results of primary hypothesis tests impedes this interpretation of the hierarchy informing therapeutic decisions in two ways.

First, the primary hypothesis under test in the vast majority of clinical trials is a "benefit" hypothesis. That is, trials are set-up, and powered to test whether a therapy produces a proposed *benefit* in a defined group of patients. Outcomes regarding the

safety of the therapy are almost always relegated to a secondary hypothesis. Whereas the possibility of benefits and harms are symmetrically important to therapeutic decisions, the quality of evidence provided within EBM for benefits and harms is asymmetrical; according to frequentist methods the benefits of therapies are tested more rigourously than the harms. The categorical interpretation of EBM's hierarchy obscures this asymmetry by proposing that therapeutic decisions be informed by reference to a hierarchy that fails to recognise the differing warrant provided by the results of primary and secondary analyses. Again, while in other sections of the literature proponents of EBM acknowledge that randomised trials are not the best method for establishing *unsuspected* adverse effects, and recognise that the results of secondary endpoints and subgroup analyses can be unreliable, there is no recognition of any of this in what proponents of EBM say about applying the hierarchy of evidence.

Second, the results of secondary endpoints and subgroup analyses play a role in informing therapeutic decisions in an *individual* (Horwitz et al. 1998; Rothwell 2005). The results of the primary hypothesis test gives information on whether the therapy benefits a defined population of patients. As discussed earlier, while the appropriate analysis of secondary endpoints and subgroups is highly contentious, these results play a role in decisions regarding individual patients. By comparing the unique characteristics of the patient in the clinic with the appropriate subgroups within the trial, therapeutic decisions can be refined so as to be more relevant to the individual. The categorical interpretation of the hierarchy fails to acknowledge the reduced warrant for findings from subgroup analyses. Further limiting the hierarchy to the results of primary hypothesis tests and the results of meta-analyses rectifies this failure, but rules out using analyses of subgroups and secondary endpoints to refine therapeutic decisions.

The categorical interpretation of the hierarchy provides a simple message for decision makers: Base your decisions on the results of randomised trials and meta-analyses. The message however is too simple; the results furnished by randomised trials are considerably more complicated. The interpretation of EBM's hierarchy can be further restricted to avoid this problem, but this more restricted interpretation severs the direct link between EBM's hierarchy and therapeutic decisions.

5. Conclusion

Proponents of EBM do not provide an unambiguous interpretation of the hierarchy of evidence. But as much as an interpretation is provided, the categorical interpretation of EBM's hierarchy is the interpretation most often put forward by advocates (either explicitly, or implicitly). The categorical interpretation holds that the results of randomised studies more reliably inform therapeutic decisions than the results of observational studies. This interpretation, however, cannot be justified without considerable caveat. Any successful interpretation of EBM's hierarchy of evidence will have to limit the claims of EBM. Two such limits are proposed. First, the application of the hierarchy should be limited to therapeutic decisions. EBM proponents, in their more careful moments, admit that the evidence hierarchy under consideration does

not apply to other medical decisions, for example, decisions relating to prognosis, or unsuspected side effects of drugs. But, the reasons for this have not been documented, and as a result are forgotten, or under-emphasised in much of the EBM literature. Further, even once the application of the hierarchy has been limited to therapeutic decisions the categorical interpretation still does not hold. The second limit further restricts application of EBM's hierarchy to the results of primary hypothesis tests and meta-analyses. The second limit is proposed because findings regarding secondary hypotheses, and subgroup analyses, are less reliable according to frequentist statistics.

As promised, this paper has been mostly negative. It has shown that the dominant (and most clear) interpretation of EBM's hierarchy that has been provided by proponents cannot be justified. And while amendments can be made to how the hierarchy is interpreted to avoid some of the criticisms this cannot be done without also restricting EBM's claims to be able to inform therapeutic decisions. In as much as there is a positive payoff to the conclusions of this paper, it will be found in clearing the way for the possibility of a considerably more restricted, and context dependent interpretation of EBM's hierarchy of evidence.

Acknowledgements

I would like to thank Mark Colyvan, Jason Grossman, and the participants of the workshop on the concept of evidence in the biohumanities held at The University of Queensland January 2007 for helpful discussion and comments.

Notes

[1] See, for instance, Armitage (1982); Lindley (1982); Suppes (1982); Urbach (1985); Worrall (2007a,b)

[2] A good recent attempt to collate some of the key arguments at the heart of EBM's claims is provided by Rothwell (2007).

[3] The "methods" referred to by proponents of EBM are certainly not new, and neither is their direct application to the bedside; clinical epidemiology pre-dates EBM. What proponents of EBM have done, however, is disseminate these ideas, and successfully convince many in the medical community for the application of clinical epidemiological ideas to be the "benchmark" when making, or justifying, medical decisions.

[4] The original evocation of Kuhn is provided in Evidence-Based Medicine Working Group (1992, 2420); the continued insistence is provided in Guyatt and Rennie (2002, 8).

[5] EBM is most certainly not a paradigm shift in the Kuhnian sense; there is no incommensurability between the new and old theories of medical decision making. Further, the shift to the EBM model of medical decision making has been (and continues to be) piecemeal—this would not be possible if EBM really was a Kuhnian paradigm shift.

[6] Guyatt and Rennie (2002) place N of 1 randomised trials at the top of their hierarchy of evidence. N of 1 trials are conducted with a single patient. In these studies, the patient is randomly allocated to a period of treatment with the intervention under investigation (the "active" treatment) or control. Once the period has ended the patient receives the alternative treatment (either active, or control). The patient's outcomes are monitored in each period. Both the patient and clinician are blinded to whether the patient is receiving active treatment or control. The set up mimics the very common "unsystematic" clinical practice of giving a

patient treatment and monitoring their outcome. N of 1 trials do not play a large role in medical research, and do not assist answering population therapeutic questions. I will not consider them here.

[7] Both Bluhm (2005) and Upshur (2005) recognise the problem of external validity in some form.

[8] Another factor important to external validity, but not related to the basic sciences, are the circumstances under which the trial was performed. If patients included in the trial are treated in ways that are importantly different to how they are treated in routine care, the then external validity of the trial will be low. Assuming the trial treated patients under realistic conditions, then the reliance on the basic sciences to inform judgments about external validity is increased.

[9] In addition, different parts of pathophysiology and pharmacology will have different levels of plausibility. EBM, by placing all of the basic sciences low on the hierarchy, fails to differentiate those parts of the basic sciences in which we have a high degree of confidence with those parts that are currently more speculative in nature.

[10] Worrall (2007a, 983) recognises this point

[11] It might be argued that "evidence-based" is doing some work in "evidence-based policy". Specifically, demarcating policy decisions based on emotion, or tabloid press, from policy decisions based on some form of "evidence". But, this use of "evidence" is much too vague. To do something more than sound vaguely reassuring "evidence-based policy" needs to be much more clear about what this "evidence" is, and how it is being used.

[12] See Feinstein (1998); Horwitz et al. (1998), and Rothwell (2005) for discussion. Bluhm (2005) also highlights some of the problems that subgroup analyses hold for EBM.

[13] See Neyman and Pearson (1933) and Neyman (1937).

[14] It should be noted that "power" as defined within hypothesis testing does not play a direct role in estimation theory. However, the conceptual framework for hypothesis testing and estimation are similar, and the influence of a concept similar to power could be outlined within estimation theory. While there are calls within medical statistics, for estimation to completely replace hypotheses testing, p values retain an important role in the analysis of clinical trials Ware et al. (1992).

[15] Although, precisely what should be done about this different epistemic standing is highly contentious. I will, however, leave the details of this debate for another time.

[16] I am also not suggesting that the results of outcomes within trials that have low pre-trial power are unimportant, or irrelevant (on the contrary these results are very important). My point is simply that the warrant provided for these results according to frequentist statistics is different to the warrant provided for the results of a well-powered primary hypothesis test. And, that the categorical interpretation of EBM's hierarchy fails to adequately recognise this difference.

References

Armitage, P. 1982. The role of randomization in clinical trials. *Statistics in Medicine* 1: 345–52.

Bluhm, R. 2005. From hierarchy to network: A richer view of evidence for evidence-based medicine. *Perspectives in Biology and Medicine* 48 (4): 535–47.

Buetow, S., R. Upshur, A. Miles, and M. Loughlin. 2006. Taking stock of evidence-based medicine: opportunities for its continuing evolution. *Journal of Evaluation in Clinical Practice* 12 (4): 399–404.

Collins, R., and S. MacMahon. 2007. Reliable assessment of the effects of treatments on mortality and major morbidity. In *Treating Individuals: From randomised trials to personalised medicine,* edited by P. M. Rothwell. Philadelphia: Elsevier.

Egger, M., and G. D. Smith. 1995. Misleading meta-analysis. *British Medical Journal* 310 (6982): 752–4.

Egger, M., G. D. Smith, M. Schneider, and C. Minder. 1997. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 315 (7109): 629–34.

Evidence-Based Medicine Working Group. 1992. Evidence-based medicine: A new approach to teaching the practice of medicine. *Journal of the American Medical Association* 268 (17): 2420–5.

Feinstein, A. R. 1998. The problem of cogent subgroups: A clinicostatistical tragedy. *Journal of Clinical Epidemiology* 51 (4): 297–9.

Glasziou, P., J. Vandenbroucke, and I. Chalmers. 2004. Assessing the quality of research. *British Medical Journal* 328 (7430): 39–41.

Grossman, J., and F. J. Mackenzie. 2005. The randomised controlled trial: Gold standard, or merely standard? *Perspectives in Biology and Medicine* 48 (4): 516–34.

Guyatt, G. H., and D. Rennie, ed. 2002. *Users' guide to the medical literature: Essentials of evidence-based clinical practice.* Chicago: American Medical Association Press.

Guyatt, G. H., A. D. Oxman, R. Kunz, G. E. Vist, Y. Falck-Ytter, H. J. Schunemann, for the GRADE Working Group. 2008a. What is "quality of evidence" and why is it important to clinicians? *British Medical Journal* 336 (7651): 995–8.

Guyatt, G. H., A. D. Oxman, G. E. Vist, R. Kunz, Y. Falck-Ytter, P. Alonso-Coello, H. J. Schunemann, for the GRADE Working Group. 2008b. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal* 336 (7650): 924–6.

Haynes, R. B. 2002. What kind of evidence is it that evidence-based medicine advocates want health care providers and consumers to pay attention to? *BMC Health Services Research* 2 (3) [cited 28 December 2007]. Available from http://www.biomedcentral.com/1472-6963/2/3.

Horwitz, R. I., B. H. Singer, R. W. Makuch, and C. M. Viscoli. 1998. Clinical versus statistical considerations in the design and analysis of clinical research. *Journal of Clinical Epidemiology* 51 (4): 305–7.

Land, C. E. 1980. Estimating cancer risks from low doses of ionizing radiation. *Science* 209 (4462): 1197–203.

Lindley, D. V. 1982. The role of randomization in inference. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1982: 431–46.

Miles, A., A. Polychronis, and J. E. Grey. 2006. The evidence-based health care debate - 2006. Where are we now? *Journal of Evaluation in Clinical Practice* 12 (3): 239–47.

Neyman, J. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236 (767): 333–80.

Neyman, J., and E. S. Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A* 231: 289–337.

Phillips, B., C. Ball, D. L. Sackett, D. Badenoch, S. E. Straus, R. B. Haynes, and M. Dawes. 2001. Oxford Centre for Evidence-Based Medicine Levels of Evidence, May 2001.

Reilly, B. M. 2004. The essence of EBM. *British Medical Journal* 329 (7473): 991–2.

Rothwell, P. M. 2005. Treating individuals 2: Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet* 365 (9454): 176–86.

Rothwell, P. M., ed. 2007. *Treating individuals: From randomised trials to personalised medicine.* Philadelphia: Elsevier.

Sackett, D. L., W. Rosenberg, J. A. M. Gray, B. Haynes, and W. S. Richardson. 1996. Evidence based medicine: What is it and what it isn't. *British Medical Journal* 312 (7023): 71–2.

Straus, S. E., W. S. Richardson, P. Glasziou, and R. B. Haynes. 2005. *Evidence-Based Medicine: How to Practice and Teach,* 3rd edition. London: Elsevier Churchill Livingstone.

Suppes, P. 1982. Arguments for randomizing. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1982: 464–75.

The GRADE Working Group. 2004. Grading quality of evidence and strength of recommendations. *British Medical Journal* 328: 1490–8.

Upshur, R. E. 2005. Looking for rules in a world of exceptions: Reflections on evidence-based practice. *Perspectives in Biology and Medicine* 48 (4): 477–89.

Urbach, P. 1985. Randomization and the design of experiments. *Philosophy of Science* 52: 256–73.

Ware, J. H., F. Mosteller, F. Delgado, C. Donnelly, and J. A. Ingelfinger. 1992. P values. In *Medical Uses of Statistics,* edited by J. C. I. Bailar and F. Mosteller. Boston: NEJM Books.

Worrall, J. 2002. What Evidence in Evidence-Based Medicine? *Philosophy of Science* 69: S316–S330.

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

———. 2007b. Why There's No Cause to Randomize. *British Journal for the Philosophy of Science* 58 (3): 451–88.

Social Epistemology
Vol. 22, No. 4, October–December 2008, pp. 371–388

Routledge
Taylor & Francis Group

Method as Argument: Boundary Work in Evidence-Based Medicine

Colleen Derkatch

In evidence-based medicine (EBM), methodology has become the central means of determining the quality of the evidence base. The "gold standard" method, the randomised, controlled trial (RCT), imbues medical research with an ethos of disinterestedness; yet, as this essay argues, the RCT is itself a rhetorically interested construct essential to medical-professional boundary work. Using the example of debates about methodology in EBM-oriented research on complementary and alternative medicine (CAM), practices not easily tested by RCTs, I frame the problem of method as a fundamentally rhetorical problem, situated within a boundary drama, and deeply rooted in the discursive practices of science and medicine. The genre of the RCT report, for example, idealises the research process and can tilt the course of arguments about CAM, while the notion of efficacy can function as a rhetorically mobile boundary object that can redefine the very terms of debate. I suggest herein that arguments about method in CAM debates can productively be read, metonymically, as expressions of more general anxieties in medicine about knowledge and evidence, community values, and professional boundaries; as such, these debates can illuminate some of the rhetorical dimensions of EBM.

Keywords: Evidence-Based Medicine; Rhetoric of Science, Health, and Medicine; Scientific Methodology; Complementary and Alternative Medicine; Clinical Trials

In his 2001 article "Message to complementary and alternative medicine [CAM]: Evidence is a better friend than power", Andrew J. Vickers depicts evidence-based medicine as the great leveller in biomedical boundary disputes. He argues that, while not many "conventional medical personnel…are prepared to go on the record to defend CAM", supporters of evidence-based medicine (EBM) are "a notable exception" in debates about the validity of practices such as chiropractic, acupuncture, homeopathy, and herbal medicine (2001, 1). Mainstream–alternative labels mean

Colleen Derkatch is a doctoral candidate in English at the University of British Columbia, where she studies rhetoric of science, health, and medicine. Correspondence to: Colleen Derkatch, University of British Columbia, English, 397-1873 East Mall, Vancouver, British Columbia, V6T1Z1 Canada. Email: derkatch@interchange.ubc.ca

ISSN 0269–1728 (print)/ISSN 1464–5297 (online)
DOI: 10.1080/02691720802559412

nothing, he says: "what matters in EBM is evidence, not how a treatment is currently categorized". Similarly, Phil Fontanarosa and George Lundberg baldly assert in their introduction to the 1998 CAM special issue of the *Journal of the American Medical Association*:

> There is no alternative medicine. There is only scientifically proven, evidence-based medicine supported by solid data, or unproven medicine, for which scientific evidence is lacking. Whether a therapeutic practice is "Eastern" or "Western", is unconventional or mainstream, or involves mind–body techniques or molecular genetics is largely irrelevant except for historical purposes and cultural interest. (1998, 1618)

If a health intervention "works", these commentators, and others (e.g. Angell and Kassirer 1998, 839), suggest, its philosophical and professional orientation should not matter. However, that orientation *does* matter in debates about the validity of CAM practices, I argue in this essay, because the evidence produced by EBM methods can tilt the course of arguments variously to defend or expand professional boundaries. Methodology itself, that is, serves a central argumentative function in debates about the legitimacy of health practices and practitioners.

The development of EBM in the late 20th century depended in large part on the intuitive appeal of the randomised controlled trial (RCT) as a means of determining which medical interventions work and which do not. Innovations such as blinding, randomisation, and placebo controls imbued the RCT with an ethos of disinterestedness that appealed to proponents of EBM, who sought to establish a new kind of medical practice based on "evidence not eminence" (as the cliché goes). Much of the rhetoric of EBM is an epideictic rhetoric, a celebratory rhetoric of praise and blame, where *evidence* is taken as a good in itself and the central premise seems simply to be "the more, the better". But it is a slippery rhetoric, too, because, while evidence in the abstract is lavishly praised, certain kinds of evidence are more praiseworthy in certain contexts than others. In the ubiquitous evidence hierarchies of the EBM literature, the quality of evidence derived from RCTs and meta-studies is ranked at the top and that derived from qualitative and observational studies at the bottom (see, e.g. Devereaux and Yusuf 2003; Committee 2005, 97–98). While the RCT design has significant and well-known shortcomings, as the ranking methodology within the larger context of evidence-production it gains a kind of concrete authority. This shift is not surprising because, as Uffe Jensen has noted of EBM, "what is accepted as evidence always depends on ontologies enacted in a particular context. Different ontologies will embody or imply different standards" (2007, 104). In rhetorical theory, this contingency might be framed as *kairos*, the fitness of an argument to time, place, and circumstance. Different evidence is differently persuasive in different contexts.

Rhetoric appears frequently as a keyword in debates about EBM but its use in this context (as in many others) is generally pejorative, understood as a counterpoint to knowledge rather than as constitutive of it.[1] For example, Charlton and Miles argue that "[b]eyond EBM there lies a whole world of good practice and real evidence which has been largely forgotten or obfuscated by rhetoric" (1998, 373). However, although the term does include this everyday meaning of empty or deceptive speech ("mere rhetoric"), it refers more generally to the study of human discourse and the webbed

relations among knowledge, belief, language, argument, speakers, and audiences. While a full-bodied theory of rhetoric is not typically foregrounded in the discourses surrounding EBM (but see Dysart-Gale 2008), behind the scenes, it permeates them. For example, a 2007 special issue of *BioSocieties* focuses on how RCTs "mobiliz[e] opinions" in EBM about drug safety and efficacy (Wahlberg and McGoey 2007, 4), while a 2006 issue of *Social Science and Medicine* examines some of the ways the idea of evidence can be marshalled to achieve specific ends (Lambert, Gordon, and Bogdan-Lovis 2006, 2620). Both of these special issues isolate political and professional processes that hinge on the idea of persuasion, and many of the papers contained within them operate in a rhetorical mode. However, this sort of work could benefit from the specificity and theoretical orientation that a rhetorical perspective can afford; my aim in this paper, then, is to add an explicitly rhetorical dimension to current debates on EBM by explicating some of the ways in which boundary-work is effected through the deployment of the randomised, controlled trial as the "gold standard" research methodology.[2]

The conventional RCT poses considerable difficulty in testing CAM practices because they are not easily randomised, standardised, or controlled; this difficulty of testing is a central issue around which participants in biomedical debates over CAM align themselves. I argue here that the problem of methodology in evidence-based analyses of CAM is a fundamentally rhetorical problem, situated within a boundary drama, and deeply rooted in the discursive practices of both science and medicine. As Kirstin Borgerson reports, an estimated 80% of biomedical interventions in North America have not been tested by RCT. "Research into alternative medicine is required to meet the highest standards", she notes, "even though many currently accepted medical practices have not met (and may never meet) those same standards" (2005, 506). Haavi Morreim similarly notes that "[a]ny attempt to throw out or discredit CAM on grounds of scientific inadequacy is sure to toss out large portions of conventional medicine alongside. To 'hold' both to 'the same' standards appears to bode far worse for medicine than for CAM" (2003, 228). In evidence-based evaluations of CAM, then, the RCT methodology seems to function as a rhetorical *topos,* or line of argument, which can be invoked variously to position the work within or beyond scientific borders. I develop this claim in two parts, which are preceded by a brief introduction to the methodological issues as the centre of the debate.

The essay begins by contextualising the *topos* within the overlapping orbits of EBM, RCT design, and CAM, whose points of contact give the argument from method shape. It then explicates how the genre of the experimental article is mobilised, as a major constituent of the *topos*, in EBM-oriented CAM research. Reports of clinical trials are modelled on the conventional publication genre of the sciences, the IMRaD format (Introduction, Methods, Results, and Discussion); because this structure conditions our beliefs about how science and medicine are conducted (and, importantly, *not* conducted), it can illuminate how method can be invoked rhetorically. The essay, finally, isolates the concept of *efficacy*—whether or not a health intervention "works"—as a central organising principle of both CAM research and EBM more broadly construed. While the determination of efficacy appears to follow inevitably from the

RCT method, this section suggests that the process of determination itself is loaded with rhetorical potential that can have significant implications for biomedical boundary work. Research on CAM, I want ultimately to suggest, precipitates what Bertolt Brecht called, in theatre, the *Verfremdungseffekt*[3]: it makes the normally tacit procedures of medical research "strange" and, as a consequence, more readily open to inspection.[4]

EBM, RCT, CAM: Ingredients of an Argument

The method *topos* gains its warrant in medical–scientific boundary work through the RCT's entrenchment as the principal source of evidence in EBM. Methodology has long functioned as a "rhetorical resource" in scientific boundary work (Yeo 1986, 261), but the positioning of CAM within an EBM paradigm further expands its argumentative scope as it reaches across the diverse conceptual geographies of science, medicine, and alternative health. This section explicates how CAM practices resist the RCT method, and how researchers remedy that resistance to frame their work within the EBM model.

First advanced formally in 1992 (Evidence-Based 1992), EBM was defined in Sackett et al.'s landmark 1996 essay as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (1996, 71). Sackett et al.'s definition has served as a rallying cry for proponents of EBM, who have built vast research networks on the premise that "best practice" in medicine is that rooted in an empirical evidence base, not in the experience and intuition of individual practitioners. EBM's rise has been controversial but widespread. For example, physicians Alvan Feinstein and Ralph Horwitz worry about the implications of EBM for practice but note nevertheless that it has "acquired the kind of sanctity often accorded to motherhood, home, and the flag" (1997, 529). Similarly, sociologists Eric Mykhalovskiy and Lorna Weir write: "Recently, a colleague, remarking on the enthusiasm for evidence within health care, noted that we live in a time of 'evidence-based everything" (2004, 1060). In principle, EBM makes perfect sense: who would not want to practice (or receive) medicine that is shown to work? However, EBM is problematic in multiple ways, ranging from upstream, epistemic concerns (e.g. What counts as "current best evidence"? Where does that evidence come from? Who judges it?) to downstream concerns in the realm of practice (e.g. Will EBM lead to an algorithmic or "cookbook" medicine? How ought one to proceed in the absence of evidence? What was medicine before it became "evidence-based"?).

Appealing to what Ted Porter (1995) calls the "cult of impersonality", the RCT holds a nearly sacred status in EBM, its aura cast so widely that it has come to seem both natural and inevitable as a research methodology. Developing in the 1940s and 1950s out of both the agricultural and statistical sciences, the RCT's acceptance in the medical world was slow, only gaining widespread adherence in recent decades (Devereaux and Yusuf 2003; Porter 1995). Many medical practitioners initially resisted the new method, viewing it as a machination of legislators to curb their clinical authority (Matthews 1995; Porter 1995). As the RCT gained favour in the 1980s and 1990s, expanding both in

scope and in its ability to draw firm conclusions, research became, in Ted Kaptchuk's view, less about "emphasizing outcomes" than "the purity of the means" of obtaining them (1998, 1724). Shifting the question of *what* we know in medicine to *how* we know it, the RCT has thus come to be seen, both within the scientific community and beyond, as "medicine's most reliable method for 'representing things as they really are'" (Kaptchuk 2001, 541). Yet the role of the RCT in evidence production in medicine can easily be overestimated, largely because of its high position on the hierarchy; this is particularly true in debates about CAM, where Sackett et al.'s reference to the "best available evidence" comes to mean almost exclusively that derived from RCTs, despite the authors' own insistence that the highest-level *available* evidence at any time is grounds enough for evidence-based practice.[5]

In their ideal form, RCTs randomly divide experimental participants into groups that receive the therapy under study ("intervention" groups) and groups that do not ("control" groups). All participants in the intervention group receive the same treatment, regardless of any variations among the individuals (e.g. health/personal history, concomitant conditions), and participants across both groups are matched to some extent for age, sex, and any other characteristics triallists deem necessary. Typically, participants and researchers are blinded to intervention assignment, a precaution meant to reduce biases in outcomes evaluations and/or placebo effects—when participants in the control group report, or experimenters note, improvement, despite their having received only an inert simulation of the study intervention (for more on RCT design, see Held, Weddle, and Wilhelmsen 2003; Jadad and Enkin 2007).

Studies of CAM pose significant methodological questions because the practices do not generally translate well into the "gold standard" RCT. Randomisation and standardisation are foreign, and often incommensurate, concepts in CAM practices, which, in contrast to biomedicine, tend to view patients as fundamentally unique, so two people with the same ailment might be treated altogether differently, depending on their unique constellation of symptoms and/or personal characteristics (Barry 2006, 2647; Degele 2005, 118). Corollary to this emphasis on uniqueness is that treatments can be difficult to standardise in experimental settings: while biomedical treatment is largely symptomatic in the sense that a person may be treated separately for different conditions (even by separate specialists), many CAM practitioners aim to address all symptoms together.[6]

Controlling and blinding studies of manual practices such as acupuncture or chiropractic are also difficult because they involve unmistakable physical actions that are difficult to simulate (e.g. piercing the skin and moving the spine with an often audible popping sound). In a practice such as acupuncture, there is no available control that is both realistic and definitely inert (*a la* sugar pill), and practitioners usually cannot be blinded. These methodological problems leave researchers to puzzle out how such studies ought to be conducted, interpreted, and incorporated into practice because the practices do not fit easily with conventional scientific procedures or genres. The question of how, then, to proceed with testing CAM in an EBM framework is nested in a complex web of factors[7]—disciplinary, professional, epistemic, generic, philosophical, commercial, regulatory, and more besides—all of which are, in part, persuasive

because the ways the practices are lined up with the RCT methodology can enhance their fit with biomedical evidence standards, or, crucially, not.

The problem with RCT-derived evidence on CAM is that the data are often difficult to interpret because their methodologies tend not to be straightforward. As Margolin, Savants, and Kleber observed in the 1998 *Journal of the American Medical Association* (*JAMA*) theme issue on CAM, RCTs depend foremost on community membership and shared assumptions, methods, and knowledge but "[t]hese conditions may not be satisfied with respect to considerations of alternative medicine, which typically involve heterogeneous groups comprising incommensurate cultural and evaluative frameworks" (1626). Without an overarching common standard between biomedicine and most CAM practices, designers of CAM RCTs must assess carefully whether the design is compatible both with accepted biomedical methodologies and with the principles of the practice under study; when the two do not match up, the solution, at least for CAM studies published in prominent biomedical journals (e.g. *JAMA*, *British Medical Journal*), tends to be exactly as we might expect: to re-shape the unconventional health practices so that they fit better into a conventional biomedical mould.

For example, to meet the biomedical norm of treatment standardisation, Shlay et al. (1998) invented what they called a "Standardized Acupuncture Regimen" (SAR) for the control of HIV-related peripheral neuropathy. This standardised approach meant that all participants from the intervention group received exactly the same acupuncture treatment—needles applied to a fixed set of acupoints—rather than a tailor-made treatment in accordance with the procedures of Traditional Chinese Medicine. Although the SAR was selected by a panel of practicing acupuncturists, the practice as manifest in the study bears little resemblance to its clinical counterpart; moreover, such a design runs a high risk of returning a false negative result (Paterson and Dieppe 2005). The triallists' decision to adopt the standardised regimen despite their own admission of its irrelevance to practice (2005, 1594) betrays their awareness of the need to fit their study within the prescribed genre of medical research, the RCT report, because the notion of standardisation, though certainly conventional, is by no means essential to trial design—nursing and psychotherapy trials, for example, frequently feature individualised interventions without compromising their quality of evidence. This kind of generic over-reaching, I describe next, illustrates the potential for method to serve argumentative ends in biomedical boundary work because the RCT is an idealised model that, even in EBM itself, frequently falls short of its reputation as the "gold standard".

RCT Reports as Idealising Genres

A key constituent of the method *topos* is the genre of the scientific report. The authors of CAM RCTs in the biomedical literature appear to employ the genre to enhance their ethos as scientists in the face of their work on subjects not fully compatible with a scientific purview. The genre itself embodies an idealised model of research that seems, from the start, to set the testing of CAM up for failure; by tracing some of the ways in which even biomedicine fails to meet these standards, we can learn more about the rhetorical

workings of RCT-derived evidence, the centrepiece of EBM. The arguments about method in CAM debates can productively be read, metonymically, as expressions of more general anxieties in biomedicine about knowledge and evidence, community values, and professional boundaries.

As Carolyn Miller's seminal formulation of genres as "social action" has demonstrated (1995),[8] genres accomplish important rhetorical work: they shape the production of discourses and condition their reception; they are instrumental in processes of identification, particularly in academic/professional settings, where writers must demonstrate a thorough understanding of their field's genres to assert their community membership and establish themselves as authorities; and they perform important epistemic functions, both in the discourses they enact and among the discursive communities that use them. Berkenkotter and Huckin note that "[g]enres are intimately linked to a discipline's methodology, and they package information in ways that conform to a discipline's norms, values, and ideology" (1995, 1)—they are, in short, "the intellectual scaffolds on which community-based knowledge is constructed" (24). However, genres can be difficult to isolate and describe because, as Anthony Paré explains, "[t]he automatic, ritual unfolding of genres makes them appear normal, even inevitable; they are simply the way things are done" (2002, 59). This sense of generic "inevitability" lends itself to a sort of generic *invisibility*, particularly in science, a world shaped by genres meant to disappear, seeming to leave readers with only objective facts. By embodying unconscious, tacit social actions, genres are inherently ideological; in science, that ideology inheres in its institutionalised genres—grant proposals, lab reports, conference papers, experimental articles, and so on.

In medicine, the RCT report, modelled on the genre of the experimental article in science, is one of its central "intellectual scaffolds". Ted Porter argues that the "relative rigidity" of the rules about conducting and publishing experiments and their results (i.e. the IMRaD model) "ought to be understood in part as a way of generating a shared discourse, of unifying a weak research community" (1995, 228). The idea that medicine, as a professional community, could be called "weak" does not mesh well with the social histories of medicine supplied by, for instance, Paul Starr (1982), who identifies medicine as the exemplary profession. However, in Porter's lexicon, weak communities are simply those, such as medicine and psychology, that rely on science but are not themselves scientific (cf. physics, the quintessential "pure" science). From this perspective, the adoption of the IMRaD model in medicine could be seen as a persuasive move on the community level, where users of the genre actively (though not necessarily consciously) foster identification between their methods and the valorised methods of science in order to secure the community's boundaries—and their position within them. This anxiety is particularly intense in an EBM context, which frequently frames the need for greater production and consumption of evidence not as a matter of medicine *becoming* more scientific but of catching up with the *other* disciplines of science. Note, for example, the tenor of R.M. Califf's admonishment: "The failure of the medical professions to develop and implement data standards has left medicine far behind most other major enterprises" (2003, 427).

Other fields of study have capitalised on the mobilising force of the scientific article, particularly the social sciences, by appropriating it in their own realms of publication (Bazerman 1988; Berkenkotter and Ravotas 1997); this "borrowing" of scientific genres is part of the communally instantiated action of genre. The notion of genre borrowing is salient to evidence-based studies of CAM because, while they are ostensibly in the mainstream (published under the imprimatur of biomedical journals), they can also usefully be thought of, in a sense, as borrowed genres: they apply the habits of mind and modes of communication of one field (medicine) to another disparate, possibly incommensurate field (acupuncture, chiropractic, etc.). Not surprisingly, then, many CAM RCT reports exhibit a kind of hyper-performance of the experimental genre through an exaggerated empiricity—a strategic overdescription of salient trial features that increase their association with scientific methods (e.g. Cardini and Weixin 1998; Shlay et al. 1998).

In the Shlay et al. study, for example, the explicitly detailed description of the needle insertion protocol is perhaps most revealing in what it does not say: the authors describe carefully the needle insertion depths—"between 1.28 to 2.54 cm (0.5 to 1.0 in) for spleen point 9, 2.54 to 3.81 cm (1.0 to 1.5 in) for spleen point 7, and 1.5 to 3.05 cm (0.6 to 1.2 in) for spleen point 6", and so on (1998, 1591)—as well as the type, frequency, and duration of the needles' manipulation, but at no point do they describe *why* the needles are inserted or manipulated. Traditional acupuncture theory is wholly absent in the report and the authors only gesture toward the intervention's potential mechanism in the Discussion, where they cite several biochemical possibilities. The absence of any underlying theory in the Methods section makes the authors' careful description of the needle manipulation protocol seem strangely undermotivated. One of the effects of this description-without-explanation is that it downplays the study's association with acupuncture as an independent professional and philosophical practice. And yet it also seems to serve a higher purpose, related to the premium placed in biomedicine on method—the means to knowledge, as Kaptchuk notes (1998), rather than knowledge itself. Method is one of the primary ways that scientists identify themselves as members of their communities and persuade readers to take their results seriously; in boundary disputes, method can be invoked rhetorically to position a given health practice or study within or beyond the borders of science.

Methods sections of experimental articles ostensibly exist for the sake of replication, to enable other researchers to repeat a study's methods to confirm or discount its findings. (Note that Shlay et al. cite replication as the primary motive behind their standardised regimen.) However, rhetoric and genre scholars Charles Bazerman (1988) and John Swales (1990), and sociologists Gilbert and Mulkay (1984) before them, have shown that methods sections are poorly suited to duplication because they are so "elliptical" that their methodologies are not explicit enough for other researchers to follow (Swales 1990, 169). Given their abstractness, I would argue, after Bazerman and Swales (and others, e.g. Giltrow 1995), that methods sections might usefully be thought of as primarily arguments from *prolepsis* (i.e. anticipating and heading off objections by indicating the appropriateness of the study's methodology) and from *ethos* (the researcher's character *qua* researcher). In CAM research, because the ability to test a

practice according to scientific principles is a central issue around which stakeholders align themselves, the methods sections of CAM RCT reports can be useful places to trace some of the ways in which methodology can be marshalled persuasively.

The problem with this emphasis on method in CAM research is that the chances of a practice such as acupuncture meeting the methodological standards instantiated by the IMRaD format *while* remaining consonant with the core tenets of its own philosophical practice *and* demonstrating efficacy seem poor indeed. The RCT genre fixes an ideal of research that most CAM practices cannot meet, particularly not in an evidence-based model of practice, where the consequent shortage of evidence of efficacy is often taken as "evidence of the lack of an effect" (Stener-Victorin et al. 2002). But the RCT genre idealises standards of evidence that even biomedical studies frequently fail to meet; in this way, I think we can read demands that CAM meet or exceed the same standards as indicative of higher-order concerns in the medical profession about its methods of study and practice.[9]

Although the EBM literature has largely valorised the RCT, it has, at the same time, fostered a thriving culture of critique both of its methods and its ideology. These critiques helpfully illuminate the rather uneven ability of the RCT to demonstrate the safety and efficacy of health interventions; they also illuminate its own self-authenticating logic. Catherine Will, for example, studies "the ways in which modifications of the 'pure' world of the experiment may also be seen as strengthening the evidence it is intended to produce" (2007, 85); referring to this process as the "alchemy" of the clinical trial, she notes that the various contingencies of research—the many different sorts of people, technologies, institutions, funding bodies, and more involved in a particular project—are transformed, in part through "the ritual invocation of randomization and control", into objective, hard facts about the world (97). Evidence and the methods of obtaining it are reified through this transformation process, where they become no longer *ideals* but always attainable and necessary *constituents* of research. A critical metadiscourse has emerged in the medical literature, pointing to a deeply penetrating instability in the evidence base—an instability that can largely be traced back to the RCT as one of EBM's organising genres. Feinstein and Horwitz point out, for example, that one review of methodological quality showed that less than half of the studies analysed met basic scientific standards (1997, 533). Similarly, others have isolated blinding (e.g. Fergusson et al. 2004), randomisation (Chalmers 1998), assessment scales (Jüni et al. 1999), and placebo controls (Lakoff 2007) as key areas of weakness in biomedical research. (See, also, Angell 2005, who argues that nearly every aspect of medical research is suspect, particularly in pharmaceutical research.) Given that biomedical studies so routinely fail to meet the same standards to which CAM studies are held, we might usefully think of Fontanarosa and Lundberg's distinction between proven and unproven therapies (quoted above) instead as a reflection of the RCT's *symbolic* value in medicine—as representative of the aims and ideals of research. This is how the rhetorical study of CAM research can illuminate biomedicine itself, by opening up its evidentiary methods to scrutiny through a kind of Brechtian alienation—that is, by making them "strange".

It is not just the RCT's claims to rigour that are problematic, though: its relevance, too, is often hard to demonstrate. For example, the assumption that adequate testing

would necessarily ensure the safety and efficacy of any medical treatment in practice, alternative or otherwise, is also faulty: we know, from extensive research (e.g. Denis et al. 2002; Dopson et al. 2002; Grimes and Shulz 2002; Will 2007), that there is often a poor association between the evidence base and clinical behaviour. Some characterise that weak association as an "implementation gap", where practitioners simply lag behind their research counterparts (see Will 2007), while others see it as the product of a more fundamental mismatch between the theory and practice of medicine (e.g. Feinstein and Horwitz 1997). Part of the disconnection between research and practice certainly has to do with time: Sackett et al. report, for example, that UK doctors for have significantly less than an hour per week available to review the literature (1996, 72).

A greater problem with the usefulness of RCT results is that, to meet the standards of internal consistency, they must sacrifice their external consistency—Nancy Cartwright calls this the problem of "front-end rigour vs. back-end rigour" (2007, 19). This focus on internal consistency limits the studies' clinical relevance because the selection criteria are so constrained and the trial populations so homogeneous that the results are difficult to extrapolate to general populations. Much of the art of medical research lies in striking an appropriate balance between front-end and back-end rigour for a specific research question—we want the results to be as accurate as possible (i.e. internally valid) but also as useful as possible (i.e. externally valid). What it means to call a practice "effective" is largely a product of rhetorical negotiation, which is of especial importance to boundary work in EBM; I turn to this final point next.

Efficacy as a Rhetorically Mobile Boundary Object

Efficacy, and its sister term, *safety*, are cited, mantra-like, throughout the medical literature as the chief motivations behind research—we want to know which health behaviours and interventions are going to help us and not hurt us. In EBM, clear determinations of safety and efficacy are assumed to be the natural and necessary outcomes of research, as Sackett et al. claim: "External clinical evidence both invalidates previously accepted diagnostic tests and treatments and replaces them with new ones that are *more powerful*, *more accurate*, *more efficacious*, and *safer*" (1996, 72; emphasis added). These two terms are the primary touchstones in determining the legitimacy of health interventions, particularly regarding CAM: they provide the terms of the debate and condition its outcome. Note, for example, Fontanarosa and Lundberg's assertion that "[t]here is no alternative medicine", only proven and unproven medicine. "[A]s believers in science and evidence", they conclude, "we must focus on fundamental issues [such as] the need for convincing data on safety and therapeutic efficacy" (1998, 1618). The implication here is that any CAM practice proven both safe and effective will be integrated seamlessly into biomedical practice. It sounds simple enough but it just is not the case because, as keywords, safety and efficacy are so flexible that they can function as gatekeepers: if an intervention meets one implied standard of efficacy, for example, sceptics can (and often do) invoke a more rigorous, and more exclusive, meaning.[10]

In this section, I argue that biomedical researchers employ a variable principle of efficacy in studies of CAM in the service of the method *topos*, to reconcile their own disciplinary allegiances with both accepted methodologies and the practices under study. In this sense, efficacy seems to serve as a kind of rhetorically mobile boundary object that both *enables* research across disparate fields and can be used strategically, if unconsciously, to advance certain argumentative ends. This claim modifies Susan Leigh Star and James Griesemer's well-known original sense of a boundary object, which they define as "an analytic concept of those scientific objects which both inhabit several intersecting social worlds…*and* satisfy the informational requirements of each of them" (1989, 393; original emphasis).[11]

Star and Griesemer's study of the multidisciplinary Berkeley Museum of Vertebrate Zoology assumes an "ecological approach", where no single viewpoint takes primacy; they focus on how divergent groups of actors in the museum's creation were able to reconcile their various and often conflicting perspectives enough to facilitate cooperation, an ability the authors attribute to the stakeholders' exchange of several classes of boundary objects. Both Joan Fujimura (1992) and Greg Wilson and Carl Herndl (2007) take up the boundary object concept and reframe it: in Fujimura's case, it becomes part of the "standardized package", an idea that both facilitates work across collectives (which she says boundary objects are good at) and stabilises facts (which she says they are not good at); while in Wilson and Herndl's case, the boundary object becomes a "rhetorical exigence" that leads to the integration, rather than the demarcation, of social-professional boundaries. The multidisciplinary situations these authors describe depend on the actors' sincere interest in collaboration and a sense of mutual respect, even for members of groups historically ranked lower than the others.

In the case of CAM research, the principle of efficacy—the idea that a given practice or intervention "works"—unites the researchers and enables their work, but the situation itself is not, in all circumstances, marked by a sense of equality. Even for those biomedical researchers that work earnestly with CAM practitioners, their relative hierarchies remain always on the horizon and the principle of efficacy evoked at a particular moment can either unite or divide the participants, depending on the researchers' orientations. There seems, then, to be a rhetorical potential within the boundary object concept that has not yet been fully explored.

Take, for example, the distinction between studies of efficacy and of effectiveness. To an outsider, the two may appear indistinctive—both, for instance, are varieties of RCT—but the difference is crucial in the EBM realm, where "[c]urrent best evidence" has come to mean, almost exclusively, evidence obtained through efficacy studies. These studies feature rigid inclusion criteria, homogeneous populations and, ideally, unambiguous endpoints to minimise statistical "noise"; they consequently have high internal validity but possibly limited applicability to real-life populations. (In Cartwright's terminology (2007), they have high front-end rigour but low back-end rigour.) Effectiveness studies are large, community-based studies of more heterogeneous groups that trade methodological fastidiousness for applicability in what Steve Maguire calls the "real-world messiness" of clinical medicine (2002, 79); featuring

lighter inclusion criteria, more varied treatment settings, "softer" endpoints, and the allowance of concurrent treatments, they produce less reliable results due to the greater statistical noise. (These studies have lower front-end rigour and higher-back end rigour.) The Institute of Medicine's Committee on CAM distinguishes them in teleological terms: "Efficacy refers to what a treatment *can* do under ideal circumstances; effectiveness refers to what a treatment *does* do in routine daily use" (2005, 104; original emphasis. See also Jadad and Enkin 2007, 13–15).

CAM practices do not fit well within an efficacy model; they are much more amenable to effectiveness studies because such studies can better accommodate the sorts of patients, symptoms, treatments, and outcomes typical of CAM. For instance, users of CAM tend to use other modalities concurrently so there would be ethical questions about restricting their use, which an efficacy study would require for the sake of causality. Likewise, endpoints of CAM studies usually need to be softer (i.e. more subjective, usually patient-reported) than those of efficacy studies because patients typically seek CAM for chronic, intractable conditions; those lacking clear prognoses and/or treatment; and those associated with hard-to-measure symptoms such as pain and fatigue. But many critics of CAM hold efficacy, not effectiveness, up as the criterion for evaluating CAM (e.g. Delbanco 1998; Happle 1998; Smolle, Praise, and Kerl 1998), even though much significant biomedical research is effectiveness-based (and still more is based on observational studies, which feature neither randomisation nor controls).

There is a sense, even among CAM-friendly biomedical personnel (e.g. Fontanarosa and Lundberg 1998; Ernst 2004; Margolin, Savants, and Kleber 1998), that we must hold CAM practices up to an even higher standard—Haavi Morreim calls it a double standard (2003, 228)—than biomedicine. And so, not surprisingly, when CAM efficacy studies are available (e.g. Cardini and Weixin 1998; Shlay et al. 1998), their methods tend to be subjected to greater scrutiny than their biomedical counterparts (Borgerson 2005; Morreim 2003). Of course, given the studies' often complex challenges in design, it seems only fair to exercise some caution when evaluating their results. But critics such as Delbanco (1998), Happle (1998), and Smolle, Praise, and Kerl (1998) invoke a stricter set of criteria for proof of efficacy that reveals the kind of double-standard that Morreim notes is characteristic of most biomedical research on CAM. So efficacy seems, in this sense, to be a shared multidisciplinary concept that enables research on CAM—that is, it is a boundary object—but one that can also be invoked in a more restricted sense by some participants to re-shape issues into a framework more amenable with their own perspective.

There is also a more insidious side to efficacy: measures intended to ensure objectivity can be actively manipulated to produce desired results; in such circumstances, *efficacy* takes on new meaning, literally, because it comes to mean whatever the triallists engineer it to mean. By altering inclusion criteria, for example, studies of interventions designed for elderly patients can "engineer out" untoward side effects by using younger participants, who tend to experience fewer (Angell 2005; Petryna 2007). Likewise, studies can "engineer up" a study drug's efficacy by administering the comparator treatment at half the normal effective dose or in a nonstandard format (e.g. tablet form

rather than injected). The most extreme altering of efficacy, of course, is results suppression, where sponsoring companies bury negative studies, present only partial evidence, or spin negative results to highlight, for example, subpopulations of the trial for which the drug did work (Angell 2005). To be sure, these sorts of manipulative strategies are not carried out in the name of boundary work: as John Abraham (2007) and others have illustrated, they result directly from pharmaceutical companies' involvement in research. (If the companies are paying for the studies, the thinking seems to go, then they should have a say in what sorts of results they produce.) But if we think of *efficacy* as a boundary object with rhetorical potential, then this highly specialised, if frequently deceptive, notion of "proving" health interventions is pertinent because it demonstrates how ephemeral a concept efficacy can be. So dismissals of CAM that hinge on the idea of efficacy seem to be motivated, at least in part, by a demarcation exigence, to use Wilson and Herndl's phrase (2007), because those dismissals use a concept that is *flexible* within biomedicine to draw relatively *inflexible* boundary lines around it.

Ted Delbanco's damnation of CAM, generally, as a glorified placebo is a good case in point. Although he laments that "the [US] public should not stand for spending tax revenues on studies not worth doing", he expresses relief that such studies will at least "shatter claims for activity beyond placebo" (1998, 1561). Because biomedical studies of CAM practices are so often difficult to blind and control, however, their potential placebo effects can be difficult to assess. Moreover, many CAM practices do not distinguish among effects, placebo or otherwise (any effect counts; see Paterson and Dieppe 2005), so the real/placebo distinction would be moot in practice. I would argue here, then, that the criteria adopted in answering the question of whether or not a particular CAM intervention has real, or merely placebo, effects directly inform the ways in which efficacy can be invoked as a rhetorical boundary object.

For CAM sceptics, such as Delbanco, these problems with identifying appropriate placebo controls in CAM can be traced to defects in the practices themselves—they cannot be tested against placebos because they *are* placebos. For those with a more moderate view of CAM, however, the placebo control plays a murkier role in evaluating efficacy because, while it is an integral component of the RCT as a profession-defining methodology, is not nearly the safeguard against bias that the EBM literature would have one believe. Placebos can be manipulated to rig trials, as Lakoff shows with the case of "targeted efficacy" in antidepressant trials (2007, 65), and their use can even be considered unethical in some contexts, particularly when alternatives with known or suspected efficacy are available, as in some HIV research (see Maguire 2002). Given that patients often seek CAM for chronic pain (as in the Shlay et al. study), the possibility that half the trial population might be given nothing at all could be ethically troubling. The relative weight commentators assign to the use (or lack of) placebo controls in debates about CAM seems to inform their assessment of a practice's efficacy.

I want to suggest with these examples that we might usefully think of the meaning of boundary objects as having a certain degree of mobility for some of the actors involved—a mobility that is context-based, dependent on kairos.[12] Wilson and

Herndl's (2007) study of "knowledge mapping" at the US Los Alamos National Laboratory is informative here: it recasts Star and Griesemer's boundary object (1989), along with Peter Galison's notion of a trading zone (a temporary site for the "local coordination" of distinct groups with "vast global differences" (1997, 783)), into a rhetorical framework for interdisciplinary cooperation. But the examples I cite here do not fit exactly within Wilson and Herndl's model either because, while cooperation is the order of the day, there are specific, if limited, instances wherein biomedical actors can reshape the boundary object to alter favourably the terms of the debate. There seems, then, to be some overlap between the boundary object trading zone and what Michael Gorman calls the "élite" trading zone, where a group of experts use their specialized knowledge to dictate how a socio-technical system will function. "The expertise of such an élite", he says, "is black-boxed for other participants in the network" (2002, 933). This is a zone in which no meaningful trading takes place. Access to the notion of efficacy in CAM research appears to be partially black-boxed to some of the participants some of the time, which allows more powerful participants to control, if temporarily, what it means to say a given health practice "works". The rhetorical mobility of efficacy as a boundary object has important implications for both CAM-related research and EBM generally because it shapes what we know about health interventions and how we know it.

Studying some of the ways in which arguments from method in CAM debates do not match up with the role of methodology in the everyday work of medical research can magnify problems that have always been central to medical research but have been largely unarticulated within its rigid methodological and generic boundaries. That is, studies on CAM can offer clearer ways of seeing how certain forms of evidence can be marshalled rhetorically, to draw professional and epistemic boundary lines. Of course, methodological rigour is the centrepiece of medical research and, as such, studies cannot be designed ad-hoc. But, since genres such as RCT reports are community-based—they represent communally held values about what counts as proper evidence—trials of CAM can be illuminating because those communal values do not quite hold. CAM practices thrive outside of the biomedical sphere and, while the major ones (such as chiropractic, acupuncture, and massage) have increasingly come to work in concert with biomedicine, their practitioners remain strangers in the community of medical scientists. Their very strangeness within biomedical borders can bring into relief biomedicine's own idealised model of research, manifest in the EBM model, wherein the best evidence for a particular health practice seems to have little to do with patients themselves.

Acknowledgements

Portions of this essay were presented at the 2005 National Communication Association and the 2007 Society for Social Studies of Science conferences; I am grateful for the generative commentary at those sessions. I am also indebted to Judy Segal for her feedback over the course of this project.

Notes

[1] This is perhaps why Yves Gringas (2007), in a recent issue of this journal, expressed such reticence about associating his study of academic misunderstandings with rhetoric, choosing instead to situate his work with a related, but differently inflected, term: *argumentation*.

[2] In his well-known theory of boundary-work, Thomas Gieryn describes science in cartographic terms, where "[t]he epistemic authority of science is…, through repeated and endless edging and filling of its boundaries, sustained over lots of local situations and episodic moments, but 'science' never takes on exactly the same shape or contents from contest to contest" (1999, 14). This "edging and filling" of boundaries makes up much of the day-to-day life in labs, observatories, and the field. All of the data that scientists produce are interpreted, sorted, and sifted, results are tabulated and deemed significant or not, and conclusions are devised; all of these are, *inter alia*, rhetorical processes (see, e.g. Bazerman 1988, Myers 1990).

[3] Brecht's well-known "alienation effect" entails intentionally breaking the theatrical illusion so that viewers, unable to lose themselves within the fiction, remain critical of what they see.

[4] I ought to note at the outset that, while rhetoric, as a critical-theoretical practice, is not about unmasking per se, this essay does a little unmasking to illuminate some of the disconnections between how medical researchers and practitioners (and policymakers, the media, and the public) think about the conduct and value of medical research, and how that research tends actually to unfold. I do not mean to set up a rhetoric-reality binary through these theory-practice disconnections, however; rather, I want to show that the RCT's attendant value system is part of the professional fabric of medicine itself.

[5] Systematic reviews, though highest on the hierarchy, "can aggregate and evaluate but cannot change the basic information" furnished by RCTs (Feinstein and Horwitz 1997, 530).

[6] I should clarify that my claim here is not that biomedical theory sponsors a pedantically one-size-fits-all approach to health care, or that CAM therapies are necessarily as individualised as their proponents suggest; these differences may be more apparent than real, although it is beyond the scope of this paper to speculate on the rhetoric of *alternative* medicine.

[7] Evelleen Richards' (1991) study of the Vitamin C-cancer controversy pre-dates EBM's rise but provides a detailed historical analysis of such factors at work in research and medical cultures.

[8] Miller defines genres as "typified rhetorical actions based in recurrent situations" (1995, 31).

[9] Swales (1990) usefully describes how experimental genres idealise models of research.

[10] In rhetorical terms, *safety* and *efficacy* function in debates about CAM as *god-terms*, defined in Kenneth Burke's lexicon as powerful, indeterminate terms that "[sum] up a manifold of particulars under a single head" (1970, 2). (*Freedom* and *love* are quintessential god-terms.) As summary terms, they carry within them various, even conflicting, interpretations—they contain, Burke says, the resources of ambiguity, the fertile ground for persuasion.

[11] Star and Griesemer note that "[t]he creation and management of boundary objects is a key process in developing and maintaining coherence across intersecting social worlds" (1989, 393).

[12] Other pertinent examples of *efficacy*'s "mobility" could usefully be explored, such as outcomes measures, whose "hardness" can be scaled up or down to raise or lower the efficacy threshold.

References

Abraham, John. 2007. Drug trials and evidence bases in international regulatory context. *BioSocieties* 2 (1): 41–56.

Angell, Marcia. 2005. *The truth about the drug companies: How they deceive us and what to do about it.* New York: Random House.

Angell, Marcia, and Jerome P. Kassirer. 1998. Alternative medicine: The risks of untested and unregulated remedies. *New England Journal of Medicine* 339 (12): 839–41.

Barry, Christine Ann. 2006. The role of evidence in alternative medicine: Contrasting biomedical and anthropological approaches. *Social Science and Medicine* 62 (11): 2646–57.

Bazerman, Charles. 1988. *Shaping written knowledge: The genre and activity of the experimental article in science.* Madison: University of Wisconsin Press.

Berkenkotter, Carol, and Thomas N. Huckin. 1995. *Genre knowledge in disciplinary communication: Cognition/culture/power.* Northvale, NJ: Erlbaum.

Berkenkotter, Carol, and Doris Ravotas. 1997. Genre as tool in the transmission of practice over time and across professional boundaries. *Mind, Culture, and Activity* 4: 256–74.

Borgerson, Kirstin. 2005. Evidence-based alternative medicine? *Perspectives in Biology and Medicine* 48 (4): 502–15.

Burke, Kenneth. 1970. *The rhetoric of religion: Studies in logology.* Berkeley: University of California Press.

Califf, R. M. 2003. Issues facing clinical trials of the future. *Journal of Internal Medicine* 254 (5): 426–33.

Cardini, Francesco, and Huang Weixin. 1998. Moxibustion for correction of breech presentation: A randomized controlled trial. *JAMA* 280 (18): 1580–84.

Cartwright, Nancy. 2007. Are RCTs the gold standard? *BioSocieties* 2 (1): 11–20.

Chalmers, Iain. 1998. Unbiased, relevant, and reliable assessments in health care: Important progress during the past century, but plenty of scope for doing better. *British Medical Journal* 317 (7167): 1167–68.

Charlton, B. G., and A. Miles. 1998. The rise and fall of EBM. *Quarterly Journal of Medicine* 91 (5): 371–74.

Committee on the Use of Complementary and Alternative Medicine. 2005. *Complementary and alternative medicine in the United States.* Washington, DC: The National Academies Press.

Degele, Nina. 2005. On the margins of everything: Doing, performing, and staging science in homeopathy. *Science, Technology, and Human Values* 30 (1): 111–36.

Delbanco, Ted. 1998. A piece of my mind. Leeches, spiders, and astrology: Predilections and predictions. *JAMA* 280 (18): 1560–62.

Denis, Jean-Louis, Yann Hebert, Ann Langley, Daniel Roseau, and Louise-Helene Trotter. 2002. Explaining diffusion patterns for complex health care innovations. *Health Care Management Review* 27 (3): 60–73.

Devereaux, P. J., and S. Yusuf. 2003. The evolution of the randomized controlled trial and its role in evidence-based decision making. *Journal of Internal Medicine* 254 (2): 105–13.

Dopson, Sue, Louise FitzGerald, Wean Ferlie, John Gabby, and Louise Locock. 2002. No magic targets! Changing clinical practice to become more evidence based. *Health Care Management Review* 27 (3): 35–47.

Dysart-Gale, Deborah. 2008. Lost in translation: Bibliotherapy and evidence-based medicine. *Journal of Medical Humanities* 29 (1): 33–43.

Ernst, Edward. 2004. Equivalence and non-inferiority trials of CAM. *Evidence-based Complementary and Alternative Medicine* 1 (1): 9–10.

Evidence-Based Medicine Working Group. 1992. Evidence-based medicine: A new approach to teaching the practice of medicine. *Journal of the American Medical Association* 268 (17): 2420–25.

Feinstein, Alvan R., and Ralph I. Horwitz. 1997. Problems in the "Evidence" of "Evidence-based Medicine". *The American Journal of Medicine* 103 (6): 529–35.

Fergusson, Dean, Kathleen Cranley Glass, Duff Waring, and Stan Shapiro. 2004. Turning a blind eye: The success of blinding reported in a random sample of randomised, placebo controlled trials. *British Medical Journal* 328 (7437): 432–34

Fontanarosa, Phil B., and George D. Lundberg. 1998. Alternative medicine meets science. *JAMA* 280 (18): 1618–19.

Fujimura, Joan H. 1992. Crafting science: Standardized packages, boundary objects, and "Translation". In *Science as practice and culture,* edited by Andrew Pickering, pp. 168–214. Chicago: University of Chicago Press.

Galison, Peter Louis. 1997. *Image and logic: A material culture of microphysics.* Chicago: University of Chicago Press.

Gieryn, Thomas F. 1999. *Cultural boundaries of science: Credibility on the line.* Chicago: University of Chicago Press.

Gilbert, G. Nigel, and Michael Mulkay. 1984. *Opening Pandora's box: A sociological analysis of scientists' discourse.* New York: Cambridge University Press.

Giltrow, Janet. 1995. Genre and the pragmatic concept of background knowledge. In *Genre and the new rhetoric,* edited by Aviva Freedman and Peter Medway, pp. 155–78. London: Taylor and Francis.

Gorman, Michael E. 2002. Levels of expertise and trading zones: A framework for multidisciplinary collaboration. *Social Studies of Science* 32 (5/6): 933–38.

Grimes, David A., and Kenneth F. Schulz. 2002. An overview of clinical research: The lay of the land. *Lancet* 359 (9300): 57–61.

Gringas, Yves. 2007. "Please, don't let me be misunderstood": The role of argumentation in a sociology of academic misunderstandings. *Social Epistemology* 21 (4): 369–89.

Happle, R. 1998. The essence of alternative medicine. A dermatologist's view from Germany. *Archives of Dermatology* 134 (11): 1455–60.

Held, P., H. Weddle, and L. Wilhelmsen. 2003. Clinical trials: Introduction. *Journal of Internal Medicine* 254 (2): 103–04.

Jadad, Alejandro R., and Murray Enkin. 2007. *Randomized controlled trials: Questions, answers, and musings.* Malden, MA: Blackwell.

Jensen, Uffe Jull. 2007. The struggle for clinical authority: Shifting ontologies and the politics of evidence. *BioSocieties* 2 (1): 101–14.

Jüni, Peter, Anne Witschi, Ralph Bloch, and Matthias Egger. 1999. The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association* 282 (11): 1054–60.

Kaptchuk, Ted J. 1998. Powerful placebo: The dark side of the randomised controlled trial. *Lancet* 351 (9117): 1722–25.

Kaptchuk, Ted J. 2001. The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology* 54 (6): 541–49.

Lakoff, Andrew. 2007. The right patients for the drug: Managing the placebo effect in antidepressant trials. *BioSocieties* 2 (1): 57–71.

Lambert, Helen, Elisa J. Gordon, and Elizabeth A. Bogdan-Lovis. 2006. Introduction: Gift horse or Trojan horse? Social science perspectives on evidence-based health care. *Social Science and Medicine* 62 (11): 2613–20.

Maguire, Steve. 2002. Discourse and adoption of innovations: A study of HIV/AIDS treatments. *Health Care Management Review* 27 (3): 74–88.

Margolin, Arthur, S. Kelly Savants, and Herbert D. Kleber. 1998. Investigating alternative medicine therapies in randomized controlled trials. *Journal of the American Medical Association* 280 (18): 1626–28.

Matthews, J. Rosser. 1995. *Quantification and the quest for medical certainty.* Princeton, NJ: Princeton University Press.

Miller, Carolyn R. 1995. Genre as social action. In *Genre and the new rhetoric,* edited by Aviva Freedman and Peter Medway, pp. 23–42. London: Taylor and Francis.

Morreim, E. Haavi. 2003. A dose of our own medicine: Alternative medicine, conventional medicine, and the standards of science. *Journal of Law, Medicine, and Ethics* 31 (2): 222–35.

Myers, Greg. 1990. *Writing biology.* Madison: University of Wisconsin Press.

Mykhalovskiy, Eric, and Lorna Weir. 2004. The problem of evidence-based medicine: Directions for social science. *Social Science and Medicine* 59 (5): 1059–69.

Paré, Anthony. 2002. Genre and identity: Individuals, institutions, and ideology. In *The rhetoric and ideology of genre,* edited by Richard Coe, Lorelei Lingard, and Tatiana Teslenko, pp. 57–71. Cresskill, NJ: Hampton.

Paterson, Charlotte, and Paul Dieppe. 2005. Characteristic and incidental (placebo) effects in complex interventions such as acupuncture. *British Medical Journal* 330 (7501): 1202–05.

Petryna, Adriana. 2007. Clinical trials offshore: On private sector science and public health. *BioSocieties* 2 (1): 21–40.

Porter, Theodore M. 1995. *Trust in numbers: The pursuit of objectivity in science and public life.* Princeton, NJ: Princeton University Press.

Richards, Evelleen. 1991. *Vitamin C and cancer: Medicine or politics?* London: Macmillan.

Sackett, David L., William M. Rosenberg, J. A. Muir Gray, R. Brian Haynes, and W. Scott Richardson. 1996. Evidence based medicine: What it is and what it isn't. *British Medical Journal* 312 (7023): 71–72.

Shlay, Judith C., Kathryn Challenger, Mitchell B. Max, Bob Flaws, Patricia Reichelderfer, Deborah Wentworth, Shauna Hillman, Barbara Brizz, and David L. Cohn. 1998. Acupuncture and amitriptyline for pain due to HIV-related peripheral neuropathy: A randomized controlled trial. *Journal of the American Medical Association* 280 (18): 1590–95.

Smolle, Josef, Gerhard Praise, and Helmut Kerl. 1998. A double-blind, controlled clinical trial of homeopathy and an analysis of lunar phases and postoperative outcome. *Archives of Dermatology* 134 (11): 1368–70.

Star, Susan Leigh, and James R. Griesemer. 1989. Institutional ecology, "translations" and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science* 19 (3): 387–420.

Starr, Paul. 1982. *The social transformation of American medicine.* New York: Basic Books.

Stener-Victorin, Elisabeth, Matts Wikland, Urban Waldenstrom, and Thomas Lundeberg. 2002. Acupuncture—a method of treatment in reproductive medicine: Lack of evidence of an effect does not equal evidence of the lack of an effect. *Human Reproduction* 17 (8): 1942–46.

Swales, John. 1990. *Genre analysis: English in academic and research settings.* New York: Cambridge University Press.

Vickers, Andrew. 2001. Message to complementary and alternative medicine: Evidence is a better friend than power. *BMC Complementary and Alternative Medicine* 1 (1): 1.

Wahlberg, Ayo, and Lindsey McGoey. 2007. An elusive evidence base: The construction and governance of randomized controlled trials. *BioSocieties* 2 (1): 1–10.

Will, Catherine M. 2007. The alchemy of clinical trials. *BioSocieties* 2 (1): 85–99.

Wilson, Greg, and Carl G. Herndl. 2007. Boundary objects as rhetorical exigence: Knowledge mapping and interdisciplinary cooperation at the Los Alamos National Laboratory. *Journal of Business and Technical Communication* 21 (2): 129–54.

Yeo, Richard R. 1986. Scientific method and the rhetoric of science in Britain, 1830–1917. *The politics and rhetoric of scientific method,* edited by John A. Schuster and Richard A. Yeo, pp. 259–297. Boston: D. Reidel.

Social Epistemology
Vol. 22, No. 4, October–December 2008, pp. 389–404

Routledge
Taylor & Francis Group

Battle in the Planning Office: Field Experts versus Normative Statisticians

Marcel Boumans

Generally, rational decision-making is conceived as arriving at a decision by a correct application of the rules of logic and statistics. If not, the conclusions are called biased. After an impressive series of experiments and tests carried out in the last few decades, the view arose that rationality is tough for all, skilled field experts not excluded. A new type of planner's counsellor is called for: the normative statistician, the expert in reasoning with uncertainty par excellence. To unravel this view, the paper explores a specific practice of clinical decision-making, namely Evidence-Based Medicine. This practice is chosen, because it is very explicit about how to rationalize practice. The paper shows that whether a decision-making process is rational cannot be assessed without taking into account the environment in which the decisions have to be taken. To be more specific, the decision to call for new evidence should be rational too. This decision and the way in which this evidence is obtained are crucial to validate the base rates. Rationality should be model-based, which means that not only the isolated decision-making process should take a Bayesian updating process as its norm, but should also model the acquisition of evidence (priors and tests results) as a rational process.

Keywords: Expert; Rationality; Bayesian Statistics; Evidence-Based Medicine

1. Introduction

Most planners consider maximizing free choice to be consistent with economic efficiency and, thus, the most effective means of promoting or enhancing social welfare. This link between the augmentation of choices and increase of social welfare is based on the assumption that decisions are made rationally. However, new behavioural

Marcel Boumans is associate professor of history and methodology of economics at the University of Amsterdam. His research domain is marked by three M's: Models, Measurement and Mathematics. He is author of *How Economists Model the World into Numbers* (Routledge 2005) and editor of *Measurement in Economics: A Handbook* (Elsevier 2007). With Anne Beaulieu, he was guest editor of a special issue of *Social Epistemology* (Vol. 18 Number 2–3) on Objects of Objectivity. Correspondence to: Marcel Boumans, Department of Economics, University of Amsterdam, Roetersstraat 11, Amsterdam 1018 WB, The Netherlands. Email: m.j.boumans@uva.nl

ISSN 0269–1728 (print)/ISSN 1464–5297 (online)
DOI: 10.1080/02691720802559453

economists and psychologists, notably 2002 Nobel prize laureate Daniel Kahneman, have shown over the past three decades that people, including experts like physicians, do not exhibit rational expectations, fail to make judgements that are consistent with Bayes' rule, use heuristics that lead them to make systematic blunders, exhibit preference reversals, make different choices depending on the wording of the problem, and suffer from problems of self-control. As a result of these findings, these economists and psychologists recommend what they call "libertarian paternalism", that is an approach that preserves freedom of choice but that authorizes both private and public institutions to steer people in directions that will promote their welfare (Thaler and Sunstein 2003). For this steering a new type of expert is called for, namely the "normative statistician", the expert in rational reasoning with uncertainty.

In planning, decision-making and risk management offices, this new kind of expert is gradually replacing the field expert, that is, the expert with specific knowledge of the field under consideration. It is assumed that this specific filed knowledge can be packed[1] into probabilities, often with the purpose of easy transmission. So, we do not need to evoke a field expert but can better search a database to acquire the relevant probabilities. The problem, however, is that such numbers do not carry information about the context from which they were derived, or the approach that was used to derive them, running the risk of misinterpretation.

As a result, we face a battle for the position of the planner's counsellor between two kinds of experts: 1) The field expert with skilled knowledge of a specific field, inclusive knowledge about usage and application of the appropriate instruments; 2) The "normative statistician" with skilled knowledge of statistical reasoning. This battle is in fact a confrontation between two kinds of rationality. To explore this encounter, the paper will compare both kinds of expertises within the context of decision making in medical practice. The starting point is a classic example of a so-called "base rate fallacy": the Harvard Medical School Test (presented below, in section 2). It appeared that, when a laboratory test result is given, physicians do not take account of the base rate, or pre-test probability, to reach a clinical decision.

A base rate fallacy is considered to be a bias, in the sense of a violation of the axioms of probability and/or a misperception of probabilities. Biases (discussed in section 3) are errors that anyone would want to correct if the matter were brought to his/her attention. A lot of experiments have shown that "reasoning with uncertainty" is tough – even to experts – and that training can be worthwhile. This "normative statistical" perspective on scientific reasoning is compared with another expert perspective on rational decision-making, the so-called Evidence-Based Medicine (EBM) approach. EBM (which will be extensively discussed in section 4) is chosen for exploring different kinds of rationality, because it provides a considerable explicit account of decision making with probabilities. In this approach, a test is only meaningful when the evidence is not clear yet, and is recommended not to apply in extreme cases, as was actually the case in the above Harvard Medical School Test. It will be shown (in section 4) that from this perspective, clinical judgments, including those of the Harvard Medical School Test, are unbiased when tests are used appropriately.

The problem being studied in this paper is what kind of decision-making can be considered as rational, that is, unbiased. Rationality is here roughly defined as correctly applying the rules of logic and those of the probabilistic calculus. It will appear that decision processes in both kinds of expertises are rational and so neither is biased in that sense. If, however, one would compare both approaches with a criterion of biasedness as defined in classical statistics (which is what is done here, in section 5), it appears that *both* approaches are biased.

The distinguishing criterion between both expertises is not their rationality but the way they take the environment into account, or in other words, how they have modelled the context in which the decisions have to be taken. In section 6, two different positions will be compared. One is that a decision, inference, or conclusion is rational when arrived at by correct reasoning insusceptible for any context. The other position, "ecological rationality", is characterized by correct reasoning where one is highly susceptible for the environment in which one takes a decision. Both positions will be discussed for the case of "rational clinical decision making", where one has to decide to ask for a test and subsequently has to interpret its possible outcomes.

2. Interpretations by Physicians of Clinical Laboratory Results

The so-called Harvard Medical School Test, carried out by Casscells, Schoenberger, and Graboys (1978), was a small survey to obtain some idea of how physicians interpret a laboratory result.

> We asked 20 house officers, 20 fourth-year medical students and 20 attending physicians, selected in 67 consecutive hallway encounters at four Harvard Medical School teaching hospitals, the following question: "If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5 per cent, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs?" (Casscells, Schoenberger and Graboys 1978, 999)

Using Bayes' theorem, the "correct" answer should be: 2%.[2] The result of this test was that only 11 of 60 participants gave this answer. The most common answer, given by 27, was 95%. The average of all answers was 55.9%, "a 30-fold overestimation of disease likelihood" (1978, 1000).

Discussing these results, Casscells, Schoenberger and Graboys observe that, despite probabilistic reasoning has been presented in prominent clinical journals for a decade, "in this group of students and physicians, formal decision analysis was almost entirely unknown and even commonsense reasoning about the interpretation of laboratory data was uncommon" (1978, 1000). This problem, however, was considered to be remediable by practical instruction in the theory of test interpretation.

Four years later a similar result was published by David Eddy (1982). He discusses a more specific case of deciding whether to perform a biopsy on a woman who has a breast mass that might be malignant. Specifically, he studied how physicians process information about the results of a mammogram, an X-ray test used to diagnose breast cancer.

The prior probability, Pr(*ca*), "the physician's subjective probability", that the breast mass is malignant is assumed to be 1%. To decide whether to perform a biopsy or not, the physician orders a mammogram and receives a report that in the radiologist's opinion the lesion is malignant. This is new information and the actions taken will depend on the physician's new estimate of the probability that the patient has cancer. This estimate also depends on what the physician will find about the accuracy of mammography. This accuracy is expressed by two figures: sensitivity, or true-positive rate Pr(+ | *ca*), and specificity, or true-negative rate Pr(− | *benign*). They are respectively 79.2% and 90.4%. Applying Bayes' theorem leads to the following estimate of the posterior probability: 7.7%.[3] In an informal sample taken by Eddy, most physicians (approximately 95 out of 100) estimated the posterior probability to be about 75%.

When Eddy asked the "erring" physicians about this, they answered that they assumed that the probability of cancer given that the patient has a positive X-ray, Pr(*ca* | +), was approximately equal to the probability of a positive X-ray in a patient with cancer, Pr(+ | *ca*).

> The latter probability is the one measured in clinical research programs and is very familiar, but it is the former probability that is needed for clinical decision making. It seems that many if not most physicians confuse the two. (Eddy 1982, 254)

According to Eddy, it is not only the physicians who are erring, but a review of the medical literature on mammography reveals a "strong tendency" to equate both probabilities, that is, to equate Pr(*ca* | +) = Pr(+ | *ca*). Generally, erroneous probabilistic reasoning is widespread among practitioners, and according to Eddy, focusing on improving this kind of reasoning will have an important impact on the quality of medical care:

> The probabilistic tools discussed in this chapter have been available for centuries. In the last two decades they have been applied increasingly to medical problems… and the use of systematic methods for managing uncertainty has been growing in medical school curricula, journal articles, and postgraduate education programs. At present, however, the application of these techniques has been sporadic and has not yet filtered down to affect the thinking of most practitioners. As illustrated in this case study, medical problems are complex, and the power of formal probabilistic reasoning provides great opportunities for improving the quality and effectiveness of medical care. (Eddy 1982, 267)

3. Heuristics and Biases

Eddy's article was published in *Judgment under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky. The first chapter of this volume is a reprint of an article by Tversky and Kahneman (1974), with the same title as the book, published in *Science*, eight years earlier.

Tversky and Kahneman (1974, 1982) explain that to assess the probability of an uncertain event, people rely on a limited number of heuristic principles that reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. However, though these "heuristics" are useful, they sometimes lead to "severe and systematic errors", also called "biases". They describe three heuristics with

accompanying biases: "representativeness", "availability", and "adjustment and anchoring". Only the first heuristics is of relevance here, because this is the one that leads to the bias discussed above: "insensitivity to prior probability of outcomes".

One type of probabilistic questions are questions like: "What is the probability that event *B* will generate event *A*?" In answering this question, Tversky and Kahneman (1974, 1982) claim that people rely on the representativeness heuristics, in which probabilities are evaluated by the degree to which *A* is representative of *B*, that is, by the degree to which *A* resembles *B*. One of the biases that go along with this heuristic is the "base-rate fallacy": the neglect of prior probabilities.

When discussing the different heuristics and biases, Tversky and Kahneman (1974, 1982) emphasize that reliance on heuristics and prevalence of biases are not restricted to laymen and "naive subjects", but when they think "intuitively", experienced researchers are prone to the same biases. "Statistical principles are not learned from everyday experience because the relevant instances are not coded appropriately" (1974, 1130; 1982, 18). This lack of an appropriate code also explains why people do not detect the biases in their judgments of probability, when no one brings this to their attention.

4. Evidence-Based Medicine

An increasingly influential movement to rationalize clinical examination is the Evidence-Based Medicine (EBM) approach, which appeared in the early 1990s. This approach was propagated by the so-called Evidence-Based Medicine Working Group,[4] chaired by Gordon Guyatt (EBM 1992), and brought to a broader attention by an editorial of the *ACP Journal Club*, a year earlier (Guyatt 1991). The primary purpose of *ACP* (*American College of Physicians*) *Journal Club*, which originally appeared as a supplement to the *Annals of Internal Medicine*, was "to help make evidence-based medicine more feasible" (Guyatt 1991, A–16). EBM was presented as a "new paradigm for medical practice":

> Evidence-based medicine de-emphasizes intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient grounds for clinical decision making and stresses the examination of evidence from clinical research. Evidence-based medicine requires new skills of the physician, including efficient literature searching and the application of formal rules of evidence evaluating the clinical literature. (EBM 1992, 2420)

This approach resulted in a pocketbook (Sackett et al. 2000, first published in 1997) with a CD and coloured cards in the cover pocket and a book's website (http://hiru.mcmaster.ca/ebm.htm) in which EBM is defined as "the integration of best research evidence with clinical expertise and patient values" (2000, 1).[5]

This book describes the practice of EBM in five steps (Sackett et al. 2000, 3–4):

- Step 1 – converting the need for information (about prevention, diagnosis, prognosis, therapy, causation, etc.) into an answerable question.
- Step 2 – tracking down the best evidence with which to answer that question.
- Step 3 – critically appraising that evidence for its validity (closeness to the truth), impact (size of the effect), and applicability (usefulness in our clinical practice).

- Step 4 – integrating the critical appraisal with our clinical expertise and with our patient's unique biology, values and circumstances.
- Step 5 – evaluating our effectiveness and efficiency in executing steps 1–4 and seeking ways to improve them both for next time.

The use of test results are part of Step 3, which is dealt with in chapters 3–7, and of which chapter 3 "Diagnosis and Screening" is of direct relevance here. The main part of this chapter has been written to help one answering three questions about diagnostic testing, which can also be found on the yellow ochre card 2A:

1. Is this evidence about the accuracy of a diagnostic test valid?
2. Does this (valid) evidence demonstrate an important ability of this test to accurately distinguish patients who do and don't have a specific disorder?
3. Can I apply this valid, important diagnostic test to a specific patient?

The third question is subsequently split up in another set of questions:

I. Is the diagnostic test available, affordable, accurate, and precise in our setting?
II. Can we generate a clinically sensible estimate of our patient's pre-test probability?
    - Are the study patients similar to our own?
    - Is it unlikely that the disease possibilities or probabilities have changed since this evidence was gathered?
III. Will the resulting post-test probabilities affect our management and help our patient?
    - Could it move us across a test–treatment threshold?
    - Would our patient be a willing partner in carrying it out?

In clinical practice, physicians are faced with three choices: to withhold therapy, to order a diagnostic test, or to treat without testing. Therefore they must take into account the reliability, value and risks of both testing and treatment to maximize both diagnostic accuracy and cost effectiveness (Scherokman 1997).

An ideal test should distinguish absolutely between patients who do and who do not have disease. The clinical usefulness of a test is determined by how much it deviates from this ideal. Data on test characteristics are derived from studying the test against a "golden standard test", the test that definitively determines the presence or absence of disease. An example of a "golden standard test" would be biopsy. Patients whom biopsy has shown to have the disease and patients shown not to have the disease are given the diagnostic test in question. To review the accuracy of the test, the results of biopsy and diagnostic test are presented in a two-by-two table (see Table 1).

Two characteristics define the accuracy of a test:

- "Sensitivity" describes the ability of a test to correctly detect disease, $\Pr(+ \mid P) = a/(a + c)$.
- "Specificity" describes the ability of a test to correctly identify absence of disease, $\Pr(- \mid A) = d/(b + d)$.

Table 1 Systematic Review of a Diagnostic Test

		Target disorder	
		present *P*	absent *A*
Diagnostic test result	positive +	*a*	*b*
	negative −	*c*	*d*

Sensitivity and specificity are considered to be stable properties of a test. They do not vary with pre-test probability of disease, also called base rate or prevalence, $\Pr(P)$. In contrast with these test characteristics, the predictive value is not a stable property and varies with the pre-test probability:

- "Positive predictive value": $\Pr(P \mid +) = \dfrac{\Pr(+ \mid P)}{\Pr(+)} \Pr(P) = a / (a + b)$
- "Negative predictive value": $\Pr(A \mid -) = \dfrac{\Pr(- \mid A)}{\Pr(-)} \Pr(A) = d / (c + d)$

Instead of using these "old-fashioned" concepts of sensitivity and specificity, EBM recommends to use the "new-fangled and more powerful" concepts of likelihood ratios to represent the accuracy of a test (Sackett et al. 2000, 72). When dealing with more then one test results, it is easier to use for calculating the post-test probabilities.

- "Likelihood ratio for positive test result": $LR(+) = \Pr(+ \mid P)/\Pr(+ \mid A)$
- "Likelihood ratio for negative test result": $LR(-) = \Pr(- \mid P)/\Pr(- \mid A)$

In general, the likelihood ratio is: $LR(X) = \Pr(X \mid P)/\Pr(X \mid A)$, where X is the random variable indicating a test result and taking values + or −. Then the interpretation of diagnostic test runs as follows:

- Pre-test odds = $\Pr(P)/\Pr(A)$
- Post-test odds = likelihood ratio × pre-test odds = $LR(X) \times \Pr(P)/\Pr(A)$

Tests can be painful and/or risky, so a clinician only asks for a test after a well-considered evaluation of reliability, value and risk. The model for making this rational decision in (Sackett et al. 2000) is based on Pauker and Kassirer (1980). This article describes a model that uses two thresholds to aid physicians in making clinical decisions:

1) a "no treatment/test" threshold, T_t, which is the disease probability at which the expected utility of withholding treatment is the same as that of performing a test;
2) a "test/treatment" threshold, T_{trx}, which is the disease probability at which the expected utility of performing is the same as that of administering treatment.

The decision not to treat, to test, or to treat is determined by pre-test disease probability and both thresholds, see figure 1. The best clinical decision for probabilities below the

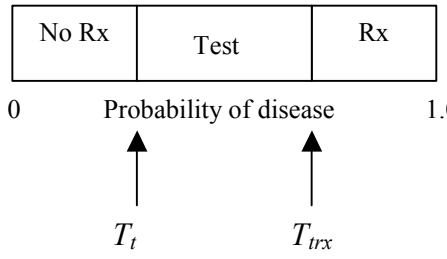


Figure 1 Test-Treatment Thresholds.
Source: Adapted from Pauker and Kassirer 1980: 1111.

"no treatment/test" threshold T_t is to refrain from treatment; for probabilities above the "test/treatment" threshold T_{trx}, the best decision is to administer treatment. When the pre-test disease probability lies between the thresholds, the test result could change the probability of the disease enough to alter the decision, so the best decision would be to administer a test.

So, for clinical decision-making, estimates of disease prevalences are crucial. It is noteworthy to see that Pauker and Kassirer, though referring to Tversky and Kahneman (1974), make, unlike Tversky and Kahneman, a distinction between expert opinions and of those outside the medical domain: "Studies in nonmedical domains show that people have biases and often make inaccurate estimates and that training improves the reliability of such estimates" (Pauker and Kassirer 1980, 1112). The sort of studies Tversky and Kahneman are referring to rely, according to Pauker and Kassirer, on "simple tests in which an actual probability is known (for example, the number of various colored balls in an urn)", whereas in medicine a prevalence "represents a belief or opinion for which no actual or true value exists" (1980, 1112). Moreover, it appears that physicians make probability estimates with "reasonable reliability" (1980, 1112). When published data on probabilities are not specific enough, the "opinions of experts" are needed and used.

In the EBM approach one will find the same position towards expert opinions, that is, needed where data on probabilities are not available, but with caution:

> Clinical experience and the development of clinical instincts (particularly with respect to diagnosis) are a crucial and necessary part of becoming a competent physician. Many aspects of clinical practice cannot, or will not, ever be adequately tested. Clinical experience and its lessons are particularly important in these situations. At the same time, systematic attempts to record observations in a reproducible and unbiased fashion markedly increase the confidence one can have in knowledge about patient prognosis, the value of diagnostic tests, and the efficacy of treatment. In the absence of systematic observations one must be cautious in the interpretation of information derived from clinical experience and intuition, for it may at times be misleading. (EBM 1992, 2421)

From the above-described threshold model test criteria can be inferred. First, according to Scherokman (1997), tests that do not change the probability of disease enough to cross the threshold probability T_{trx} are not useful and should not be ordered. This means that when the pre-test disease probability lies between the thresholds and we

have a positive test result, the post-test disease probability should lie above the test/treatment probability:

$$\Pr(P \mid +) > T_{trx}.$$

This is in fact a weak criterion, because it implies that the disease should (causally) influence the test result:

$$\Pr(+ \mid P) > \Pr(+).^{6} \tag{1}$$

In statistics, an event A is independent of event B if $\Pr(A \mid B) = \Pr(A)$. In probabilistic accounts of causality, it is crudely stated that B causes A if $\Pr(A \mid B) > \Pr(A)$. So, the above requirement (1) obviously excludes tests like flipping a coin.

A stronger test requirement is that it should be "most informative". A test is most informative when its predictive values, $\Pr(P \mid +)$ and $\Pr(A \mid -)$, are optimal. As is said above, these values depend on the pre-test probabilities. It can be shown that both predictive values are optimal when:[7]

$$\Pr(P) = 1 / \left(\sqrt{LR(+)LR(-)} + 1 \right).$$

Usually the test characteristics sensitivity and specificity are about equal, which means that the optimal pre-test probability is about 50%.[8] Generally, it is expected that a test is most informative when the pre-test probability of disease is between 40% and 60% (Scherokman 1997).

These demands on tests with respect to accuracy and applicability give new light on the interpretation by physicians of clinical laboratory results. First, assume that condition for using the test is optimal: $\Pr(P) \approx 0.5$, so $\Pr(A) = 1 - \Pr(P) \approx 0.5$. When sensitivity and specificity are about equal, then

$$\Pr(+) = \Pr(+ \mid P)\Pr(P) + \Pr(+ \mid A)\Pr(A) \approx \Pr(+ \mid P)0.5 + \Pr(- \mid P)0.5 = 0.5$$

So, if physicians assume that a test is used for optimal conditions, there is no question of base rate fallacy, because:

$$\Pr(P \mid +) = \frac{\Pr(+ \mid P)}{\Pr(+)} \Pr(P) \approx \Pr(+ \mid P)$$

Secondly, let us take Eddy's figures: $\Pr(+ \mid P) = 79.2\%$ and $\Pr(- \mid A) = 90.4\%$, and assume that $40\% < \Pr(P) < 60\%$, then $37.44\% < \Pr(+) < 51.36\%$, and so

$$84.6\% < \Pr(P \mid +) < 92.5\%.$$

Most physicians estimated the post-test probability to be about 75%.

And finally, the Harvard Medical School Test figure, $\Pr(+ \mid A) = 5\%$, leads even to higher post-test probabilities, when the prevalence is between 40% and 60%:

$$93\% < \Pr(P \mid +) < 95\%.$$

Recall that most common answer, given by 27 of 60, was 95%.

Physicians are trained not to ask for diagnostic tests when prevalences are too small (or too large). Faced with test results they might have assumed automatically that the test was performed for the right conditions. So, they might have developed a heuristic to read the sensitivity and specificity as predictive values. Seen from this perspective, the physician's high estimates of the post-test probabilities in the case of the Harvard Medical School Test and in Eddy's test are not biased, but show "ecological rationality". (This type of rationality takes account of the environment and will be discussed in section 6.)

5. Statistical Bias

In the literature discussed above, it is assumed that Bayesian reasoning is an unbiased heuristic. In mathematical statistics, however, unbiasedness has a very specific meaning: An estimator, $\hat{\theta}$, is unbiased if and only if it's expected value is equal to the parametric value, θ, it is intended to estimate: $E[\hat{\theta}] = \theta$. A consequence of this specific definition is that an estimator based on Bayesian reasoning is not automatically unbiased. In a widely used standard textbook on statistics, *Introduction to the Theory of Statistics*,[9] one will find the following remarkable observation: "in general a posterior Bayes estimator is not unbiased" (343).[10] So, in an early training in statistics, one is already warned that Bayesian tools and unbiasedness might be incompatible.

Being warned, let us check whether the post-test probability, that is the probability taking account of test results, $\Pr(P \mid X)$, is an unbiased estimator of the pre-test probability, $\Pr(P)$. Let X be the random variable indicating the test result, taking value + or −.

$$E[\Pr(P \mid X)] = E\left[\frac{\Pr(X \mid P)}{\Pr(X)}\right]\Pr(P) = \left[\frac{\Pr(+ \mid P}{\Pr(+)}\Pr(+) + \frac{\Pr(- \mid P)}{\Pr(-)}\Pr(-)\right]\Pr(P) = \Pr(P)$$

So it seems that our worry was unnecessary. Unfortunately, this is not the case. Generally, in rational decision-making (including in EBM), it is highly recommended to use likelihood ratios to estimate the disease odds. When discussing the use of likelihood ratios, Roger Cooke (1991) gives an expression how one can "learn" from observations (adapted from his theorem 6.3, 97):

$$E[LR(X) \mid P] \geq 1, \text{ and equality holds if and only if } \Pr(LR(X) = 1 \mid P) = 1$$

The equality condition can hold only if $\Pr(X \mid P) = \Pr(X \mid A) = \Pr(X)$. A test that would have this latter characteristic is not informative because it is then independent of disease, and should therefore be excluded, see Equation 1.

However, this theorem shows only that one can learn from a test in case the disease is present. It surprisingly happens to be that in case of an absent disease, a test will not "learn" us about the absence of this disease:

$$E[LR(X) \mid A] = \left[\frac{\Pr(+ \mid P)}{\Pr(+ \mid X)} \right] \Pr(+ \mid A) + \frac{\Pr(- \mid P}{\Pr(- \mid A)} \Pr(- \mid A) = 1)$$

This result makes the test biased

$$E[LR(X)] = E[LR(X \mid P] \cdot \Pr(P) + E[LR(X) \mid A] \cdot \Pr(A) > 1$$

So, it appears to be the case that post-test odds are not unbiased estimators for the pre-test odds:

$$E[\Pr(P \mid X) / \Pr(A \mid X)] = E[LR(X)] \Pr(P) / \Pr(P) > \Pr(P) / \Pr(A)$$

The undesired result of this bias is that each time a test result is being taken account of (whatever the result is, positive or negative) the expected disease odds will increase.

6. Ecological rationality

An important critic of Kahneman and Tversky's normative statistical approach is Gerd Gigerenzer:

> If you open a book on judgment and decision making, chances are that you will stumble over the following moral: Good reasoning must adhere to the laws of logic, the calculus of probability, or the maximization of expected utility; if not, there must be a cognitive or motivational flaw. Don't be taken in by this fable. (Gigerenzer 2004, 62)

Gigerenzer describes Kahneman and Tversky's approach as a study of cognitive illusions: its primary aim seems to be to demonstrate that people's judgments do not actually follow the laws of probability or the maximization of expected utility. "The result is a list of deviations from norms, which are interpreted as cognitive fallacies, emphasizing irrationality rather than rationality" (2004, 65).

In Gigerenzer's account of heuristics, the rationality of heuristics is not logical, but ecological. Ecological rationality implies that a heuristic is not good or bad, rational or irrational per se, only relative to an environment. Gigerenzer (2004) mentions twelve examples of phenomena that were interpreted as "cognitive illusions" but which he reevaluated as "reasonable judgments given the environmental structure" (66).

In fact, the interpretation of a test result by physicians can be seen as another example of ecological rationality. The (fast and frugal) heuristic is to read the sensitivity of a test as the predictive value when the test result is positive. This is reasonable in an environment of Evidence Based Medicine practice where test results are only asked for when prevalences are not decisive yet, and tests are most informative.

Not taking the environment into account can lead to all kinds of so-called "paradoxes" in statistics. These paradoxes are used to show that when making judgments regarding the likelihood of uncertain events, even mathematically sophisticated

people do not follow the principles of probability theory. According to Kahneman et al. (1982), "this conclusion is hardly surprising because many of the laws of chance are neither intuitively apparent, nor easy to apply" (32). A famous example is the Monty Hall problem.[11] Discussing this problem in the *American Statistician*, Morgan et al. (1991) ended their conclusions with the following question: "How do you expect me to solve a problem that stumped scores of Ph.D.'s [sic] and confused the world's most intelligent person?"! (287). In his Comment, Seymann (1991) separated this question into two (in his view) distinct issues. The first is concerned with clarity of problem definition, and the second is concerned with "why sensible and mathematically well-trained people, given that they agree on what the problem is, still get the wrong solution" (287). To address the latter issue, Seymann gives a few examples (Bertrand's Box Problem, Birthday Problem) well known in statistics, but he refers also to Kahneman, Slovic, and Tversky's edited volume *Judgement Under Uncertainty* (1982), in particular to the Harvard Medical School Test. Interestingly, it is this example of the Harvard Medical School Test which provoked others to respond to Seymann's commentary.

The first comment is by John P. Wendell, who arrived at the same conclusion as section 4, and is thus worthwhile to quote at length:

> This answer of 2% apparently assumes that everyone in the population, whether they have the disease or not, has an equally likely chance of receiving the test and that the false negative rate is zero… Neither of these assumptions is stated or clearly implied in the problem. Stating "you know nothing about the person's symptoms or signs" is not the same as stating that the test has an equal chance of being administered to people in the population, even if that was the intent of the phrase. The medical students and staff that were given this question would know full well that patients having a disease are almost more likely to have a test for their disease administered to them than the general public… The majority response of 95% is consistent with the assumption that persons having the test applied to them have a 50% chance of actually having the disease… Certainly these assumptions are more reasonable than those needed to support the 2% answer. Perhaps this illustration shows not that medically trained people don't understand probability but that some statisticians don't understand medicine. (Wendell 1992, 242)

In a subsequent response, Seymann (1992) stated that to "know nothing about the person's symptoms or signs" is not an instruction to assume random testing, but "a clear instruction to disregard all other information, biases, or prejudices we might have" (1992, 242):

> one must ask where a 50% prior, though perhaps understandable in other circumstances, here results in the fabrication of a new prior and the dismissal of a vital piece of explicit information. (Seymann 1992, 242)

The problem here of course is that what is a "vital piece of explicit information" for a statistician is not necessarily the same as for a physician.

To illustrate this, let us first have a look at one of the fables of Ackoff:

> In a conversation with one of my colleagues I was asked how I would go about determining the probability that the next ball drawn from an urn would be black if I knew the proportion and number of black balls that had previously been drawn. He told me that the urn

> contained only black and white balls. I replied that I would first find out how the urn had been filled. "No", he said, "that is not permissible". "Why?" I asked, "Certainly you have such information". "No, I don't", he replied. "Then how do you know the urn contains only black and white balls?" I asked. "I have it on good authority", he answered. "Then let me talk to that authority", I countered. In disgust he told me to forget about the whole thing because I clearly missed the point. I certainly did. (Ackoff 1974, 89)

The moral of this fable is that the ability to solve a textbook exercise is not equivalent to the ability to solve a real-world problem. Textbook exercises are usually formulated so as to have only one correct answer and one way of reaching it. Real-world problems have neither of these properties. An essential part of problem solving, according to Ackoff, lies in determining what information is relevant and in collecting it.

By discussing six problems in reasoning with probabilities, so-called "teasers", Bar-Hillel and Falk (1982) show that the way we model a problem is strongly dependent on the way the information was obtained.

> The kind of problem in which the conditioning event does turn out to be identical to what is perceived as "the information obtained" can only be found in textbooks. Consider a problem which asks for "the probability of A *given* B". This nonepistemic phrasing sidesteps the question of how the event B came to be known, since the term "give" supplies the conditional event, by definition… Outside the never-never land of textbooks, however, conditioning events are not handed out on silver platters. They have to be inferred, determined, extracted. In other words, real-life problems (or textbook problems purporting to describe real life) need to be *modeled* before they can be solved formally. And for the selection of an appropriate model (i.e., probability space), the way in which information is obtained (i.e. the statistical experiment) is crucial. (Bar-Hillel and Falk 1982, 120–121)

Bar-Hillel and Falk emphasize that a probability space for modelling verbal problems should allow for the representation of the given outcome and the statistical experiment which yields it. They illustrate how different scenarios for obtaining some information yield different solutions. In other words, the way one models a problem is strongly dependent on how the information is obtained. Different ways of obtaining the self-same information can significantly alter the revision of probability contingent upon it. Real-life problems need to be modelled before they can be solved formally. And for the selection of an appropriate model (e.g., probability space), the way in which information is obtained (i.e. the statistical experiment) is crucial.

In the case of The Harvard School Test (Casscells, Schoenberger, and Grayboys 1978) and in the later test by Eddy (1982), it was simply assumed that both questioner and respondent had the same model in mind. However, both were trained differently and therefore had modelled the problem differently.

7. Conclusions

Generally, rational decision-making is conceived as arriving at a decision by a correct application of the rules of logic and statistics. If not, the conclusions are called biased. After an impressive series of experiments and tests carried out the last few decades, the view arose that rationality is tough for all, skilled field experts not excluded. A new type

of planner's counsellor is called for: the normative statistician, the expert in reasoning with uncertainty par excellence.

To unravel this view, the paper has explored a specific practice of clinical decision-making, namely Evidence-Based Medicine. This practice is chosen, because it is very explicit about how to rationalize practice.

One of the key examples of biased expertise is the Harvard Medical School Test, which shows that physicians often commit a base rate fallacy: they confuse the accuracy of a diagnostic test for its predictive value. However, it is shown that for the base rate given in the Harvard Medical School Test it is not rational to ask for a diagnostic test. Moreover, it is shown that for base rates between the test–treatment thresholds, it is an unbiased heuristic to take a test's accuracy as its predictive value.

Most practices of rational decision making prefer the ratios of likelihoods to simple likelihoods because they are easier and more practical to update when new evidence (e.g. a test result) comes in. The term bias has a specific meaning in mathematical statistics. Using this specific interpretation of biasedness, it is shown that paradoxically a rational application of likelihood ratios leads to biased results.

It has also been shown that whether a decision-making process is rational cannot be assessed without taking into account the environment in which the decisions have to be taken. To be more specific, the decision to call for new evidence should be rational too. This decision and the way in which this evidence is obtained are crucial to validate the base rates. Rationality should be model-based, which means that not only the isolated decision-making process should take a Bayesian updating process as its norm, but should also model the acquisition of evidence (priors and tests results) as a rational process. The use of thresholds is an option for that.

Acknowledgements

This paper was presented at the "Biased experts versus plain facts" session of the Annual Meeting of the Society for Social Studies of Science (November 2006), Vancouver, Canada and at the International Congress "The Social Sciences and Democracy: A Philosophy of Science Perspective" (September 2006), Ghent University, Belgium. I am grateful for the comments received from participants at both sessions. I am also grateful for the comments and encouragements from the participants of the project "The Nature of Evidence: How Well Do 'Facts' Travel?" of the Department of Economic History, London School of Economics. In particular I would like to thank Jacqueline Krol for providing her EBM practice-based materials and Jon Adams for his editorial support.

Notes

[1] This concept of 'packaging' is borrowed from Leonelli (forthcoming), where she explores the idea that packaging is needed to make facts travel.

[2] $\Pr(P\mid+)=\dfrac{\Pr(+\mid P)\Pr(P)}{\Pr(+\mid P)\Pr(P)+\Pr(+\mid A)\Pr(A)}\approx 0.001/(0.001+0.05\cdot 0.999)=0.02$, where P: disease is present, A: disease is absent, and +: positive test result.

[3] $$\Pr(ca \mid +) = \frac{\Pr(+ \mid ca)\,\Pr(ca)}{\Pr(+ \mid ca)\Pr(ca) + \Pr(+ \mid benign)\Pr(benign)} =$$

$$0.792 \cdot 0.01 \,/\, (0.792 \cdot 0.01 + 0.096 \cdot 0.99) = 7.7$$

[4] This group consisted of the following members: P. Brill-Edwards, J. Cairns, D. Churchill, D. Cook, A. Detsky, M. Enkin, P. Frid, M. Gerrity, H. Gerstein, J. Gibson, B. Haynes, J. Hirsch, J. Irvine, R. Jaeschke, A. Kerigan, A. Laupacis, V. Lawrence, Mark Levine, Mitchell Levine, J. Menard, V. Moyer, C. Mulrow, P. Links, A. Neville, J. Nishikawa, A. Oxman, A. Panju, D. Sackett, J. Sinclair, and P. Tugwell.

[5] A third edition by S.E. Straus, W.S. Richardson, P. Glasziou, and R.B. Haynes was published in 2005. The 2nd edition is however used for this paper.

[6] If $\Pr(P) < T_{trx}$, and $\frac{\Pr(+ \mid P)}{\Pr(+)} \Pr(P) > T_{trx} \Rightarrow \Pr(+ \mid P) \,/\, \Pr(+) > T_{trx} \;/\; \Pr(P) > 1.$

[7] This can be seen by maximizing $\Pr(P \mid +) \cdot \Pr(A \mid -)$ for $\Pr(P)$.

[8] When $\Pr(+ \mid P) \approx \Pr(- \mid A)$, then also $\Pr(- \mid P) \approx \Pr(+ \mid A)$, and thus $LR(+)LR(-) \approx 1$.

[9] First edition by Alexander Mood was published in 1950. The 2nd edition coauthored by Franklin Graybill appeared in 1963, and the 3rd edition with Duane Boes as third author was published in 1974.

[10] A "posterior Bayes estimator" is defined as $E[Y \mid X]$, where X is a random variable with probability $\Pr(X \mid Y = y)$, and Y a random variable with probability $\Pr(Y)$. A posterior Bayes estimator is an "unbiased" estimator of y when $E[E[Y \mid X] \mid y] = y$. It is shown that a posterior Bayes estimator is unbiased only when this estimator correctly estimates y with probability one. In all other cases the estimator is not unbiased.

[11] This problem raised a good deal of commotion, even among mathematicians, when discussed by vos Savant (1990). She phrased the problem as follows: "Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #1, and the host, who knows what's behind the doors, opens another door, say #3, which has a goat. He says to you, 'Do you want to pick door #2?' Is it to your advantage to switch your choice of doors?" (1990, 13).

References

Ackoff, Russell L. 1974. *Redesigning the future: A systems approach to societal problems.* New York: Wiley.

Bar-Hillel, Maya, and Ruma Falk. 1982. Some teasers concerning conditional probabilities. *Cognition* 11: 109–122.

Casscells, Ward, Arno Schoenberger, and Thomas Grayboys. 1978. Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine* 299 (18): 999–1000.

Cooke, Roger M. 1991. *Experts in uncertainty: Opinion and subjective probability in science.* New York and Oxford: Oxford University Press.

Eddy, David M. 1982. Probabilistic reasoning in clinical medicine: Problems and opportunities. In *Judgment under uncertainty: Heuristics and biases,* edited by Kahneman, Slovic, and Tversky. Cambridge: Cambridge University Press.

Evidence-Based Medicine Working Group. 1992. Evidence-based medicine: A new approach to teaching the practice of medicine. *The Journal of the American Medical Association* 268 (17): 2420–2425.

Gigerenzer, Gerd. 2004. Fast and frugal heuristics: The tools of bounded rationality. In *Blackwell handbook of judgment and decision making,* edited by D. Koehler and N. Harvey. Oxford: Blackwell.

Guyatt, Gordon H. 1991. Evidence-based medicine. *Annals of Internal Medicine* 114 *(ACP Journal Club* supplement): A–16.

Kahneman, Daniel, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Leonelli, Sabina. Forthcoming. Packaging small facts for re-use: Databases in model organism biology. In *How well do facts travel,* edited by P. Howlett and M. S. Morgan. Cambridge: Cambridge University Press.

Mood, Alexander M., Franklin A. Graybill, and Duane C. Boes. 1974. *Introduction to the theory of statistics,* 3rd edition. Tokyo: McGraw-Hill.

Morgan, J. P., N. R. Chaganty, R. C. Dahiya, and M. J. Doviak. 1991. Let's make a deal: The player's dilemma. *The American Statistician* 45 (4): 284–287.

Pauker, Stephen G., and Jerome P. Kassirer. 1980. The threshold approach to clinical decision making. *The New England Journal of Medicine* 302 (20): 1109–1117.

Sackett, D. L., S. E. Straus, W. S. Richardson, W. Rosenberg, and R. B. Haynes. 2000. *Evidence-based medicine: How to practice and teach EBM,* 2nd edition. Edinburgh: Churchill Livingstone.

Scherokman, Barbara. 1997. Selecting and interpreting diagnostic tests. *The Permanente Journal.* Available from http://xnet.kp.org/permanentejournal/fall97pj/tests.html; INTERNET.

Seymann, Richard G. 1991. Comment. *The American Statistician* 45 (4): 287–288.

Seymann, R.G. 1992. Response. *The American Statistician* 46 (3): 242–243.

Thaler, Richard H., and Cass R. Sunstein. 2003. Libertarian paternalism. *The American Economic Review* 93 (2): 175–179.

Tversky, Amos, and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185: 1124–1131. Reprinted as chapter 1 in *Judgment under uncertainty*, ed. D. Koehler and N. Harvey. Cambridge: Cambridge University Press.

vos Savant, M. 1990. Ask Marilyn. *Parade Magazine,* September 9: 13.

Wendell, John P. 1992. Comment. *The American Statistician* 46 (3): 242.

Social Epistemology
Vol. 22, No. 4, October–December 2008, pp. 405–423

Routledge
Taylor & Francis Group

Science, Legitimacy, and "Folk Epistemology" in Medicine and Law: Parallels between Legal Reforms to the Admissibility of Expert Evidence and Evidence-Based Medicine

David Mercer

This paper explores some of the important parallels between recent reforms to legal rules for the admissibility of scientific and expert evidence, exemplified by the US Supreme Court's decision in Daubert v Merrell Dow Pharmaceuticals, Inc. in 1993, and similar calls for reforms to medical practice, that emerged around the same time as part of the Evidence-Based Medicine (EBM) movement. Similarities between the "movements" can be observed in that both emerged from a historical context where the quality of medicine and legal approaches to science were being subjected to growing criticism, and in the ways that proponents of both movements have used appeals to "folk epistemologies" of science to help legitimate their reform aspirations. The term folk epistemology is used to describe the weaving together of formal and informal images of scientific method with normative and pragmatic concerns such as eradicating "junk science", and promoting medical best practice. Perhaps unsurprisingly, given the unfocused breadth of these aspirations the implications of these "reforms" for medical and legal practice have not been straightforward, although they do represent an important new set of rhetorical resources to critique and or legitimate expertise in medical and legal domains. Discussion closes, by noting the growth of calls for these movements to reciprocate in areas where law and medicine intersect, such as medical negligence litigation.

Keywords: Evidence-Based Medicine; Expert Evidence; Daubert; Law and Science; Folk Epistemology

David Mercer is an Associate Professor in STS, School of English Literatures, Language and Philosophy (ELPL) in the Faculty of Arts at the University of Wollongong, Australia. His primary research interests are in public policy in relation to science, expertise and law. Correspondence to: David Mercer, Faculty of Arts, Science Technology and Society Program, University of Wollongong, Wollongong 2522, Australia. Email: david_mercer@uow.edu.au

ISSN 0269–1728 (print)/ISSN 1464–5297 (online)
DOI: 10.1080/02691720802559420

Introduction

This paper analyses some of the important parallels between reforms to legal rules for the admissibility of scientific and expert evidence, exemplified by the US Supreme Court's decision in *Daubert v Merrell Dow Pharmaceuticals, Inc.* (*Daubert*) in 1993, and similar calls for reforms to medical practice that emerged around the same time as part of the Evidence-Based Medicine (EBM) movement. A number of similarities can be drawn between the "movements". First, both emerged at a time when the quality of medicine and legal approaches to science were being subjected to broad criticism. Second, both focused on the quality of expertise as the "the problem" and the adoption of more "scientific" approaches to expertise, which downplayed experience and credentials as ways of authorising expertise, as "the solution". Third, in promoting "scientifically-based expertise", proponents of both movements appealed to "folk epistemologies" of science to help legitimate their reform aspirations. These folk epistemologies wove together formal and informal (quite often inconsistent) models of scientific method with normative and pragmatic concerns. Fourth, whilst the rhetoric used by experts to maintain legitimacy in each domain has become more influenced by concern with "scientific method" and appeals to "mechanical forms of objectivity" (Porter 1995) changes to actual practices have not followed in simple or predictable ways. Perhaps unsurprisingly, given the malleability of the folk epistemologies central to both movements, EBM and *Daubert* have come to mean a variety of different things to different actors in different contexts (Brody, Miller, and Bogdan-Lovis 2005; Timmermans and Mauck 2005). The malleability of these folk epistemologies may have also assisted in the success of EBM and *Daubert* being taken up into policy contexts.

EBM: From "Medical Nemesis" to "Effectiveness and Efficiency"

The ongoing tension between images of medicine as an art and as a science provides an important backdrop to the emergence of EBM. For the better part of the second half of the 20th century in most western nations medical practitioners were extremely successful in balancing these images: appealing to science whilst simultaneously noting the professional traditional sources of authority of medical art, craft and the value of the experience of the clinician (Berg 1995; Porter 1998). The mainstream allopathic medical profession enjoyed, in the post Second World War era, high status and authority, generally beyond other professions and bodies of knowledge. This comfortable position began to come under challenge in the 1970s and 1980s by patient movements (Landzelius 2006), and Marxist, Feminist, and anti-technocracy critiques. This challenge was enhanced by general political and economic calls for the rationing of public services from the late 1970s, which in many countries included significant budgets for medicine. Many of these critiques highlighted the costs and apparent lack of efficacy of standard allopathic medicine and the excessive and largely unaccountable basis for medical authority in the hands of (normally) male doctors/clinicians. Some of these critiques also note the reliance of allopathic medicine on

scientific reductionism, suggesting the need for a broader base for foundations of medicine in alternative therapies (Illich 1976; Ehrenreich 1978; Wright and Treacher 1982). Criticisms also emanated from "within" medical science. In this later genre, and of greatest immediate relevance to the rise of EBM, was the work of Archie Cochrane. In his book "Effectiveness and Efficiency" Cochrane (1972) suggested that many of the tests, procedures and interventions, used in medicine may have been doing more harm than good and may have had no, or limited, evidence for their effectiveness.

The first "formal" formulations of EBM appeared in the early 1990s as part of the efforts of teams lead by Gordon Guyatt and David L. Sackett, linked to the Department of Clinical Epidemiology Canada's McMaster University and the Centre for Evidence-Based Medicine in Oxford. From these initiatives a number of frequently cited formulations of EBM can be found, the most well-known being Sackett's "conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett et. al. 1996, 71). Elaborating on this basic formulation, many EBM proponents critiqued what they believed had been the excessive reliance on the "experience" of the individual clinician, particularly when compared to the superiority of practices based on science, particularly the use of the randomised clinical trial (RCT) (Rudd and Dickenson 2002).

Despite its implicit scientism (as the implications and significance of RCTs are still reliant on expert interpretation) (Richards 1991), EBM's critique also resonated with some more radical critiques of medicine. Well-known feminist writer Ann Oakley suggested EBM could be seen as "fundamentally anti-authoritarian" empowering the lay public by "providing well designed experimental ways of knowing" that check the arrogance of medical professionals (Oakley 2000; Kuhlman 2004). In an environment where critiques of medicine seemed to be multiplying, but in a context where the medical profession and bureaucracies still enjoyed considerable power, some commentators have suggested that the uptake of EBM, as uncomfortable to many medical specialists as may have been, and continues to be (Charlton 1997), was encouraged by EBM's capacity to acknowledge the problems of medicine but at the same time offer a practical pathway for the authority of medicine and medical institutions to be redeemed. As Denny (1999) notes, whilst early EBM claims appeared to strongly downplay the status of the individual clinician, there was still a strong appeal to the authority of medicine: as long as it was medicine working within the new scientific EBM paradigm. EBM would also appear to have enhanced its initial adoption by identifying the source of medical problems in a recognisable category of deviant actors: clinicians engaging in unscientific "medical" practices. Whilst EBM pioneer Sackett has argued in some places that EBM is a tool that *does* value clinical judgment, it is clinical judgment scientifically informed rather than one based on status or experience. In 2000, Sackett wrote a paper titled "The Sins of Expertness and a Proposal for Redemption". He noted that:

> There are still far more experts around than is healthy [for the advancement of clinical science] … adding our prestige to our opinions gives the latter far greater persuasive power than they deserve on scientific grounds alone. (Sackett 2000, 1283)

Daubert: From "Rationality and Ritual" to "Conjectures and Refutations"

Despite the obvious differences between medical policy and legal settings, the cultural environment that helps explain the origins of *Daubert* "reforms" shares important similarities to the emergence of EBM. Around the same time as the emergence of critiques of traditional medicine there were mounting criticisms of law/science interactions. At a more general level of public debate, there were a number of widely publicised cases involving miscarriages of justice. Prominent amongst these were failings of forensic experts in the IRA trials in the UK,[1] and in the Chamberlain case in Australia (Edmond 2002a).[2] There were also serious public questions raised about the adequacy of psychological testimony, particularly syndrome evidence and cases involving "repressed memories". Suggestions that judges should have been more critical and prevented such evidence from being heard, and that expert associations needed to better police the quality of their members were becoming increasingly commonplace across common law jurisdictions (Freckelton 1994). At a more overtly political level many law/science commentators shared a concern with law and science "in general" not just the behaviour of individual scientists and legal institutions. Much of this latter discourse was preoccupied with questions of technological risk. In the US this was manifest in the 1980s and 1990s with the significant growth in the public visibility of toxic tort litigation, of which important examples included Thalidomide, Asbestos, Agent Orange, and Bendectin (Schuck 1986; Sanders 1998). In the UK these matters arose a little more obliquely through judicial and public inquiries investigating topics such as the viability or otherwise, of expanding nuclear power. In this later context, Brian Wynne (1982) suggested that common law legal traditions, when addressing scientific and technical issues, became little more than "rationality rituals". Their dominant technocratic inclinations limited their ability to grapple with the uncertainties and social and political dimensions of science and technology as they systemically excluded considering types of evidence that failed to fit limited legalistic models of rationality. Some other commentators provided complimentary but more politically subtle reading of the technocratic inclinations of courts. Sheila Jasanoff, who dealt mainly with the US legal system, shared Wynne's concerns regarding the parallels between legalism and scientism. She noted that despite their weaknesses, US courts at least offered a vehicle in which new scientific claims about risk could be publicly heard. By doing this they provided a form of informal technology assessment that could indirectly encourage more responsive regulation of technology, and facilitate civic education by exposing the uncertainties inherent in modern techno-scientific practices to the public (Jasanoff 1995; Edmond and Mercer 1996).

Another group of more politically conservative commentators viewed the situation quite differently. For them, toxic torts and miscarriages of justice provided graphic examples of a general malaise that had set in to the US and other common law legal systems. Courts were being far too lenient in admitting poor quality expertise into courts and were encouraging the proliferation of "junk science" (Edmond and Mercer 1998a). Notable critics such as Peter Huber wrote polemically of the

legal distortion of "scientific evidence" by greedy lawyers, experts willing to prostitute themselves for a fee, and gullible judges and juries being easy targets for deceit. Huber's work was also supported by key industry lobbies and insurance and pharmaceutical companies who claimed that allowing "junk science" entry into courts was leading to massive financial losses for industry, a crisis in insurance, and suppressing technological and scientific innovation. Huber's rhetoric was far more corrosive than Cochrane and Sackett but his views on the failure of experts share some similarities:

> Maverick scientists shunned by their reputable colleagues have been embraced by lawyers. (1991, 2)

> The legal lips murmured no, no, to seductions scientific, but the eyes and arms said yes. (1991, 13)

In response to the types of concerns noted above, in particular toxic torts, the US Supreme Court reviewed the federal rules for the admissibility for scientific evidence in the 1993 *Daubert* case. In other common law jurisdictions less dramatic, but equally far reaching, reviews of the role of science and experts also began to take place (Woolf 1996; Australian Law Reform Commission 1999; Edmond 2003). *Daubert* generated significant publicity and received 22 *amici curiae* briefs[3] from scientific associations, industry groups, medical and other professional associations who had a stake in the outcome (Edmond and Mercer 1998b). The Supreme Court cited a number of these briefs to generate a set of criteria for judges to consider when making determinations about the admissibility of scientific evidence. The judge was also now expected to play an important "gatekeeping" role: they now needed to consider not just the conclusions generated by expert witnesses but the methodology and reasoning behind them. This new judicial responsibility and pre-occupation with scientific methodology, in particular testing (Edmond and Mercer 2002) supplanted earlier rules for admissibility, such as the "*Frye* test" (set down in the 1923 case *Frye v United States*), which had in a sense deferred to experts regarding the question of evaluating the adequacy of the methodology behind a knowledge claim as long as it came from a generally accepted area of science. The specific model of science proposed by *Daubert* will be discussed in more detail below. There are obvious analogies between EBM's call for medicine to no longer defer to the experience of the eminent clinician and *Daubert's* call for judges to look at the methodology behind a knowledge claim and not simply defer to the eminent expert witness (Tallacchini 2002).

Like the EBM "revolution", the *Daubert* revolution hooked into broader critical debates about problems with law and science, but ultimately focused on "experts" as the problem and "science" as the solution. In a sense both movements can be seen as offering "epistemological fixes" (appeals to ideals of scientific method and mechanical objectivity) to perceptions of broader socio-technical problems surrounding medicine and the interactions of law and science. As might be expected the neatness of the rhetoric of such "epistemological fixes" would be difficult to sustain, as the movements became influential and actual practices and doctrines began to be re-shaped.

EBM as Scientifically-Based Medicine

Having noted the similarities in the broad origins and framing of the two movements, it is useful to move on to consider the significance of appeals to science in EBM and *Daubert* and what sort of "science" this is? A broad definition outlining the scientific pretensions of EBM appeared in a 1995 editorial of the *British Medical Journal* (BMJ):

> [E]vidence-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 decisions; and, finally, performance should be constantly evaluated. (Davidoff et. al. 1995, 1085)

In many of the foundational statements supporting EBM, its proponents were also quite explicit in drawing upon T.S. Kuhn's "Structure of Scientific Revolutions" to emphasise that EBM represented nothing less than a paradigm shift for medicine. Some proponents went as far as suggesting that "the principles of EBM should become standard training of all physicians, and that those physicians who violate its precepts, should ultimately face license suspension" (Sehon and Stanley 2003, 1).

At a foundational level, early EBM's "method rhetoric" (Schuster and Yeo 1986; Mercer 2002a) was strongly based on two simplified philosophical threads: empiricism: in the form of the randomised clinical trial, and a strong faith in the efficacy of biostatistical techniques, i.e. epidemiology. This model of empiricism could be described as simple, in that there is a limited concern with the "theory loading" of observations or possible incommensurability between competing theories. EBM *is* concerned with observer bias in a technical sense regarding, for example good statistical techniques and effective randomisation, but not in a deeper philosophical one: recognising more complex models of epistemology where theory and observation may be intertwined. Such models imply that even as useful and rigorous RCTs might be, they can still also be subject to differing interpretations (Richards 1991). There is also limited evidence of reflection on the possible limits of epidemiology and statistical methodology (Grossman and Mackenzie 2005). It is assumed that evidence "speaks for itself" and that physiological theory, experience and interpretation are untrustworthy. EBM's "method rhetoric" resonates with Theodore Porter's historically reinforced observations that we can expect a strong correspondence between appeals to numbers, routines and classifications ("mechanical" objectivity) as sources for legitimacy of expertise at times of institutional change and uncertainty about the quality of expert judgment (Porter 1995).

From Clinical Trials to Classifications

Whilst many of the features of EBM's simple method rhetoric remain, there is evidence that EBM in practice has become a much more epistemologically variegated entity. One of the first features, a number of commentators have noted, has been the tendency

when EBM principles have been applied to practical contexts, for the generation of hierarchical lists classifying the usefulness or otherwise of different types of scientific/medical practices. In practical terms relying on RCTs alone is too exclusionary and risks discounting too many medical practices and too much existing medical knowledge (Grossman and Mackenzie 2005). Not all lists generated treat various practices the same way. RCTs normally retain their privileged place but there have been a variety of different rankings afforded to other types of study varying between different institutions and contexts. The drift into lists, classifications, and guidelines detracts from the neatness of EBM's original "method rhetoric" by inviting considerations that there might be a variety of different types of evidence: good, bad, and better evidence: not just evidence.

A further set of concerns has surrounded the drift from "bedside EBM"—the analysis of treatments, informing clinicians and providing the patient with room for improved choice, to "regulatory EBM"—the use of EBM and clinical guidelines as tools in the economic management and rationing of medical treatments. Despite foregrounding of science-based rhetoric in EBM, these practical applications of EBM are unsurprising. Management concerns have always been present as a motivation for the development of EBM. From its foundations Cochrane noted not only the problem of medical efficacy, but also the need "for money to be spent wisely". EBM and clinical guidelines have become important tools in the economics of medicine. Payment for treatments whose efficacy doesn't "add up" according to RCTs and meta-analysis may be discouraged, particularly if they are expensive (Caudill and LaRue 2006). The effects on patients, treatments, and political sensitivities to "regulatory EBM" have varied depending on differing political contexts and settings. In countries with longer traditions of socialised medicine and medical technology assessment, such as Norway, Sweden, and Finland, regulatory EBM may offer little change (Dickenson and Vineis, 2002; Vos, Houtepen, and Horstman 2002). In contexts of "managed care", where a third party such as an insurer, can draw upon EBM to influence what medical interventions should be available to a patient, there has been growing debate. In the US lobby groups such as "Citizens' Council on Health Care" (CCHC), which describes itself as "[a] free market resource for designing the future of health care", have embarked on polemical attacks on the spread of EBM into the US health care system. In a report published in 2005 titled evocatively, "How Technocrats are Taking Over the Practice of Medicine: A Wake-up Call to the American People", the CCHC identifies numerous negative implications for patients as a by-product of the introduction of EBM, which include: "… [r]igid standards of care imposed on patients…[r]estrictions on professional freedom and judgment … [r]ationing of health care services … [and the] … [p]oliticization of medicine" (Brase 2005, 2).

EBM as a "Literary" Technology

Another of the more pragmatic aims of EBM pioneers such as Cochrane was to improve the capacity of medical practitioners to gain access to, and manage, medical

information. It was widely noted that there were huge challenges for physicians to keep up with the increasing volume of medical research. To just provide one example, in 1992 alone, more than 6000 articles dealing with adult internal medicine were published in English language clinical journals: to keep abreast of internal medicine alone, a physician would have needed to read 17 articles a day (McQueen 2001). Problems of medical information overload have been one of the factors that have encouraged the development of Clinical Practice Guidelines and EBM has been touted as one of the ways to maintain quality control in their development.

Linked to these concerns, one of the most conspicuous manifestations of EBM in institutional terms has been the establishment across the world of "Cochrane collaborations". These organisations act as clearing houses for EBM certified information, where medical information can be assembled and meta-analysis of studies be carried out: a set of potentially slow and expensive processes. Despite the rhetoric of EBM discounting the experience-based knowledge of the clinician and its promotion of a straightforward epistemology, which justifies medical practices in scientific/empirical terms, a significant part of EBM in practice involves information manipulation, classification, and assisting the development of guidelines. The eminent clinician has not necessarily been simply replaced by the "scientific expert clinician", but by teams of librarians, economists and bureaucrats with skills in informatics as well as basic medical knowledge (Brown and Webster 2004).

EBM linked clinical practice guidelines have been critiqued on a variety of grounds (Cohen, Stavri, and Hersh 2004):

(1) For proliferating, and therefore leading toward re-producing some of the medical information problems supposedly being addressed by EBM.
(2) Being used to guide practice in contexts in which there may be a lack of evidence, or a poor fit between a specific case and a guideline that has been formulated in general terms.
(3) Relying on the synthesis, interpretation and opinion of experts whose identity and possible interests may be buried (black-boxed) in the guideline (Hurwitz 2004; Mulrow and Lohr 2001).
(4) Involving considerable effort and expense to construct.
(5) Given their expense, suffer from political and financial selection biases in terms of the types of therapies and ailments being considered (Rogers 2002).

EBM's "Folk Epistemology"

Despite these and other critiques suggesting that EBM may not be fulfilling its promise as a straightforward "epistemological fix" for problems with medicine, EBM has become the key way of framing questions about medical policy across the world, viewed favourably from editorials of the BMJ to bureaucratic declarations of the World Health Organisation (Guyatt, Cook, and Haynes 2004; Samanta, Samanta, and Gunn 2003).

The flexibility, fluidity, hybridity, and "comprehensiveness", of the types of expertise EBM proponents have come to promote are captured well in a 2004 editorial in the

BMJ. The editorial praises EBM and assesses its first decade and future directions. Some extracts from the editorial are reproduced below:

> Evidence-based medicine… required new skills, including efficient literature searching and the application of formal rules of evidence in evaluating the clinical literature [but is now evolving to an emphasis on] the limitations of using evidence alone to make decisions, and the importance of value and preference judgments that are implicit in every clinical management decision…
>
> The process of producing relevant evidence through high quality research will continue indefinitely, requiring considerable investment by funding agencies all over the world. The process of summarising that evidence is daunting…
>
> Evidence-based medicine's biggest future challenge is one of knowledge translation, ensuring that clinicians base their decision making on the right principles and on current best evidence. [Clinicians will need to use] [c]omputer systems for decision support that can incorporate reminders, directives, and incentives as well as audit and feedback…
>
> [There are increasing demands for the] clinician [to] quickly and accurately ascertain patients' values. (Guyatt, Cook, and Haynes 2004, 990–991)

It would appear for the BMJ, EBM has "become" medicine: it is synonymous with the promotion of clinical research, meta-analysis, and summary, and for clinicians the requirement to have skills in literature searching, the application of formal rules of evidence, clinical judgment, knowledge translation, medical informatics, psychology, and applied ethics.

One of the most significant roles for the "advertised" epistemology of EBM may have been its rhetorical function in helping EBM to absorb criticism, bind together a much looser movement for medical reform than proponents have advertised, and still maintain an authoritative "unitary" image. A factor that may have helped facilitate this has been the way EBM proponents have successfully appropriated the epistemic term "evidence" and seamlessly "woven" it into the pragmatic and normative ideal of "best practice". Whilst the distinctive features lying at the heart of EBM, of critiques of clinical experience and the celebration of the significance of biostatistics, can still be seen as drivers beneath the "surface" of EBM, proponents can absorb critiques of this emphasis by noting that EBM also considers other forms of evidence as long as they still constitute the best way to practice medicine. At a level of popular epistemology: it can appear difficult or even foolish to argue against knowledge based on evidence or medicine based on best practice:

As Sehon and Stanley note:

> [T]here is a tendency for some proponents of EBM to duck these questions [what does "evidence" actually mean in EBM] and avoid this debate by *defining* [italics in original] "evidence-based medicine" such that it includes the best combination of basic science clinical experience, and clinical trials. In so doing, the proponents of EBM come awfully close to simply defining EBM as the best way to practice medicine, whatever that may be. In other words, they respond to the second order conceptual question ("what is EBM") by saying that EBM is whatever approach medicine best answers the normative question ("how ought we to practice medicine?"). The proponents of EBM thereby give the illusion of having answered both sets of questions when in fact they have answered neither". (2002, 2)

EBM's promiscuous folk epistemology may have helped the label become ubiquitous but it is interesting to consider whether important debates about the strengths and weaknesses of RCTs, the value or otherwise of clinical experience, disputed interpretations of the value and quality of clinical guidelines, concerns over where to set the boundaries between definitions of therapeutic effectiveness and cost, have been discouraged by being woven into EBM's fabric.

The *Daubert* Criteria: Science "Fit" for Court

As noted in an earlier section of discussion, the US Supreme Court in *Daubert v Merrell Dow Pharmaceuticals* embarked upon what has been described as a "revolutionary" shift in the admission of scientific expert opinion evidence. *Daubert's* interpretation of the US *Federal Rules of Evidence* 1975 replaced the *Frye* "general acceptance" test for the admissibility of scientific evidence. "General acceptance" had meant that for admission, novel expert opinion evidence needed to conform to methods, principles, and conclusions that had received widespread "acceptance" in the relevant fields. The *Daubert* judgement produced new criteria for the admissibility of scientific evidence. It provided four flexible and non-exhaustive criteria for judges to employ when assessing the validity of purportedly scientific evidence. Criteria to be used in assessing science included: whether the claims can and have been tested (falsificationism); whether the theory or technique has been subjected to peer review and publication; the known or potential rate of error; and, whether there has been "general acceptance" of the claim within a relevant scientific community (Mercer 2002b).

The majority US Supreme Court also emphasised the primacy of Sir Karl Popper's doctrine of testability/falsification in distinguishing science from other forms of inquiry:

> Ordinarily, a key question to be answered in determining whether a theory or technique is scientific knowledge that will assist the trier of fact will be whether it can be (and has been) tested. "Scientific methodology today is based on generating hypotheses and testing them to see if they can be falsified; indeed, this methodology is what distinguishes science from other fields of human inquiry". Green, at 645. See also C. Hempel, *Philosophy of Natural Science,* 49 (1966) ("[T]he statements constituting a scientific explanation must be capable of empirical test"); K. Popper, *Conjectures and refutations: The growth of scientific knowledge,* 37 (5th ed. 1989) ("[T]he criterion of the scientific status of a theory is its falsifiability, or refutability, or testability"). (*Daubert* 1993, 593)

In most (but not all) legal and popular science commentary the *Daubert* criteria were originally treated as an accurate and useful description of science (Goodstein 2000; Foster and Huber 1997; Saks and Faigman 2005). Interestingly at the outset some commentators critical of technocratic sympathies of courts, given past tendencies of deference to the judgment of scientists regarding the traditional *Frye* test, interpreted *Daubert* as potentially liberalising the entry of novel scientific claims (Jasanoff 1995). Like Sackett's rhetoric, courts acting as gatekeepers and considering the methodology behind scientific claims, could be seen as an antidote to expert evidence being admitted to court on the basis of consensus or eminence of experts ahead of the quality of the

"evidence" itself. Other more politically conservative critics praised *Daubert* as vehicle that could limit "unfair" liabilities faced by corporate defendants and the state by inhibiting courts admitting junk scientific claims and reinforcing mainstream well-established science.

Some of the space for these divergent interpretations was created by certain ambiguities in the judgment. The Court noted its criteria should be interpreted flexibly, whilst a minority dissenting opinion questioned the majorities emphasis on Popperian falsification and the wisdom of requiring judges to become amateur scientists (or philosophers of science) to satisfy their gatekeeping role. Further, whilst the "criteria" appeared to be framed as an epistemologically, not pragmatically warranted guide to admissibility, complete with citations to eminent philosophers of science such as Popper and Hempel, the majority also noted that the courts role was not to seek "cosmic understandings" (Daubert 1993, 600). In subsequent cases, the US Supreme Court (in *Joiner* and *Kuhmo*) has confirmed judicial gatekeeping obligations, reinforced ideas that courts should seek scientific truth as their guide, and agreed that the *Daubert* criteria should be extended to the consideration of expert evidence beyond what may have strictly normally be considered as science, such as forms of engineering knowledge (Edmond 2002b).

Science (Un)Fit for Court? *Daubert's* "Exclusionary Ethos"

Earlier suggestions that *Daubert* could become a vehicle to liberalise the admission of expert evidence have not been supported. Surveys of judges, lawyers, and cases in a post-*Daubert* legal environment have reinforced the idea that the *Daubert* criteria tend to have acted as an "exclusionary ethos" and that expert evidence outside mainstream science is now less likely to be considered by courts. Like EBM, in a number of important toxic tort cases, courts have demanded statistically strong epidemiological evidence and testing as preconditions for the entry of expert claims (Edmond and Mercer 2000; Miller and Miller 2005). *Daubert's* demand for quality control extends even further than EBM. The criteria such as peer review and publication, and general acceptance, provide judges with a comprehensive set of resources to justify rejecting the admission of expert testimony. Litigation averse industry lobbies, likely to be defending themselves against liability claims made by plaintiffs who are generally less financially well-equipped to present scientific claims that have been subjected to extensive testing, or are relying on more novel scientific claims yet to have achieved "general acceptance", have been open in praising *Daubert's* impacts. Like EBM, the popularity of *Daubert* (as much as an idea and as, in a sense a social movement) has been profound (Edmond and Mercer 2004b). Even in legal jurisdictions such as Australia, New Zealand, and Canada where *Daubert*-style criteria have not been formally legally adopted, the criteria still frequently warrant favourable mention, often in the context of a judicially sanctioned reliable model of science. Further, in policy arenas there have been calls in the US for the *Daubert* criteria to be used as guide for evaluating proposals for regulation (Mercer 2004; Yearley 2005).

Different *Dauberts*?

As with EBM, criticism of *Daubert* seems to have grown as the criteria have become more popular. First, there have been observations that the criteria may not be being applied consistently across different areas of expert evidence, surveys suggest that state forensic evidence has been far less likely to be rejected on the basis of *Daubert* challenges than plaintiffs' expertise pertinent to pharmaceutical and product liability litigation (Dixon and Gill 2002). This has been the case although much state forensic evidence doesn't easily satisfy *Daubert* types of requirements: various techniques rely on the experience-based knowledge of forensic investigators ahead of testing and generating falsifiable hypotheses, and there has not traditionally been a strong tradition of publication in peer reviewed journals (Cole 2001). Courts would appear to have been both reluctant to dismiss things like state fingerprint evidence whilst simultaneously displaying some uneasiness and inconsistency in responding to claims that such evidence doesn't satisfy the *Daubert* criteria.[4]

Recently, there have been critics who believe that there has been a growing bias against public health regulation in the United States, due to the successful lobbying of the government by various industry groups. The critics have also suggested that *Daubert-based* reforms have been appropriated to help promote this trajectory: as part of a process, to use the phrase of Michaels and Monforton (2005), "manufacturing uncertainty" the setting of unrealistically high standards for regulatory science to justify regulatory inertia. Further critiques have suggested that the *Daubert* criteria are incoherently eclectic in bringing together philosophical views of science that are inconsistent (Haack 2001). For example, most popular versions of Popper's falsification emphasise the limits of other alternative criteria for defining science: falsification is interpreted as superior to simple empiricism, statistical probability and consensus formation (Edmond and Mercer 2002). *Daubert's* inadequacies and eclecticism have arguably been one of the factors that have helped it to be used as an exclusionary tool (Kassirer and Cecil 2002). Depending on the context, judges have shown a willingness to take their gatekeeping functions seriously and expose novel claims to rejection, even if they have been published and peer reviewed journals, if there has been doubts about their general acceptance or the methods according to which claims have been tested (Edmond and Mercer 2004a).

Aside from *Daubert* becoming influential for political reasons (through its adoption as a tool to help address the so called "litigation crisis" (Saks 1992) and problems of junk science) like EBM, *Daubert*'s "success" can also be explained in its appeal to folk epistemology and its epistemological promiscuity. It appears absurd to make arguments against the idea that courts should not consider the best science, and whilst courts may have tended to emphasise testing as a vehicle to help exclude claims, the presence of the *Daubert* criteria help insulate it against challenges that it is promoting any single epistemological view of science. Critics of *Daubert* can find themselves being accused of objecting to the notion that judges should reject *any* poor quality expertise from being admitted to court even though their actual concerns may be that *Daubert* provides a poor model for satisfying these aims (Bernstein 1999). Like EBM, questions

can be raised about how much the *Daubert* criteria have really assisted judges in assessing practical issues such as: what sorts of testing are important? Are there circumstances where testing may not be relevant? What weight should be given to peer review? And, at what point should an expert witness' experience be valued, and when should it be rejected?

Reciprocation and Medical Negligence

Given the commonalities in context and epistemic style between *Daubert* and EBM, it is not surprising that commentary has recently begun to appear on how these approaches can reciprocate in legal and regulatory settings. The most direct possibility for reciprocation is in the area of medical negligence litigation (Hines 2006). In a number of jurisdictions there have been tensions in medical negligence litigation between how much the establishment of standards of care should rely on the views of experienced physicians or on "scientific evidence" such as clinical practice guidelines. In the UK there have been suggestions that with the improved systematic development of clinical guidelines influenced by EBM and the establishment of health auditing bureaucracies such as the National Institute for Clinical Excellence, medical practitioners will find it more difficult to defend themselves against claims of medical negligence by suggesting they have simply followed standards of care set by the medical profession rather than a standard of care backed by an official clinical guideline (Samanta, Samanta, and Gunn 2003). In the US these tensions have also "mapped" on to recent calls for a shift from a "custom-based standard of care" to a "reasonable physician standard" (Jackson 2006). Some commentators have suggested that as more courts follow the latter path they have a further incentive to consider the scientific basis for a "reasonable physician", as "custom" is now only one of the factors to be considered when evaluating a physician's conduct. EBM and *Daubert* have become contenders to help establish models for judging a scientific basis for a standard of care. EBM through its guidance in the construction of clinical practice guidelines (Williams 2004), and *Daubert* in emphasising standards based on testing, peer review, and general acceptance. For some commentators EBM and *Daubert* not only provide part of the solution to improving medical negligence litigation but also provide damning critiques of previous "custom"-based approaches standards of care.

> [C]linical research evidence (evidence-based medicine) proves that customary standards are inefficient. A report from the 1980s showed that only fifteen percent of medical practices were based on clinical trials. Meanwhile, clinical trials have proven that some of the common practices used by physicians are ineffective. Without solid evidence for customary practices, doctors continue to perform inefficient treatment. Applying a *Daubert* analysis resolves many of the weaknesses with the traditional customs standard. It ensures that expert opinion is grounded in scientifically sound principles and methodologies. Published research suggests the finding is methodologically sound because the work has "weathered" peer review. Judges and juries will only hear evidence that the medical community considers real science… Testimonies on the science means physicians cannot be found non-negligent just because they hid behind customary standards. (Hines 2006, 15–16, references omitted)

It is interesting to note that, like many other commentaries on EBM and *Daubert,* the epistemology they are credited with promoting becomes extremely broad. In the quote above, for example, sound medical science would appear to require clinical trials, publication, peer review, and general acceptance, and again there is a strong implication that the experience-based knowledge of experts is unreliable. Whilst it would be obviously valuable for medical negligence cases to use the best quality evidence available, idealised images of science drawing from both EBM and *Daubert* may offer resources that might operate in practice quite differently to how the commentator above envisages (Rosoff 2001).

Despite enthusiastic promotion in much medico-legal literature, US and UK courts have (at least at the time of writing) failed to consistently incorporate *Daubert* and EBM considerations into making medical negligence determinations. Surveys in the US have suggested that, whilst, as noted earlier, the *Daubert* decision has raised the threshold for the admissibility of scientific and medical expertise in product liability and toxic tort cases, the same pattern has not reappeared in medical negligence litigation (Havighurst et. al. 2001; Samanta, Samanta, and Gunn 2003).

Conclusion

Despite the significant differences between the domains of law/science and medicine/policy, the *Daubert* and EBM movements share some key similarities. Both emerged from social contexts where the role of science in the law and the efficacy of dominant medical practices were being subjected to a broad range of critiques across a number of social, political and academic domains. *Daubert* and EBM movements, albeit in different rhetorical terms, and at times in contradictory ways, came to focus on particular aspects of these critiques and downplay others. Rather than call directly for institutional or structural changes, in both cases, the quality of expertise, i.e. expert testimony and clinical judgment (based on experience) came to be focused upon as the key problem, and for expertise to become more scientifically (methodologically)-based, as the solution. Supporters and detractors of both movements have made bold claims for their revolutionary nature. How these movements may have altered the actual practices of the experts in question, as noted above, is a more complex issue. Any appeals to base practice on mechanical rules or "scientific method" will encounter challenges in application and both EBM and *Daubert* have promoted images of scientific method that are often inconsistent and unclear.

In this later context, it was suggested that this lack of philosophical consistency and clarity may have so far encouraged the popularity of *Daubert* and EBM amongst policymakers rather than detract from it. Both movements rhetorically favour epistemic positions that elevate forms of "mechanical objectivity", such as ranking the significance of RCTs, biostatistics, and testing, ahead of clinical experience, expert opinion, and interpretation. But in practice, both have still allowed the accommodation of these less epistemologically "correct" forms of knowledge, for pragmatic or political reasons. By weaving together epistemic and normative notions of scientific method and best practice, *Daubert* and EBM would appear to allow policy makers to maintain legitimacy

whilst they enjoy some "disgressionary" slack. For example, in the context of EBM, clinical judgment still appears in the construction of clinical guidelines, and economic considerations are wedded to EBM's epistemology in "regulatory EBM", when it is used as a tool for guiding the rationing of medical funding. Similarly in the context of *Daubert*, the call for judges to consider scientific methodology, especially whether or not an expert's claim is falsifiable and has been tested, co-exists with "softer" more sociological inflected criteria such as publication, peer review and general acceptance. Further, whilst in policy contexts *Daubert* criteria have been used to "chill" toxic tort litigation, which many commentators viewed as a social and political problem, they have not been to used undermine state forensic evidence.

In closing, it should be noted that the flexible meanings and uses of EBM and *Daubert* have not stopped some important changes occurring within the practices of "science in law" and medical policy. Significantly, experts working in these domains are increasingly being confronted with different models of accountability (Timmermans 2005). There is currently a "tension" in the ways expertise is legitimated in these fields. Credentials and experience have become less significant as ways of claiming authority than demonstrations of "epistemological correctness": that experts are following the scientific method. These changes have the potential to encourage new forms of expertise and destabilise the power relationships between different types of experts and expert knowledge. Appeals to more "scientific" statistically-based knowledge may empower researchers and experts with specialised mathematical skills, better able to quantify their claims, but at the same time, in seeming contradiction, lead to empowering hybrid experts, "generalists" with a mixture of scientific bureaucratic and managerial expertise who can claim authority for their practices in terms of their aptitude/experience in translating and applying "scientifically" supported guidelines and standards (Mercer 2004).

Acknowledgements

I would like to thank Jason Grossman, Jerry Ravetz, and Joan Leach for helpful comments on an earlier version of this paper which was presented to the Fourth Queensland Biohumanties Conference: Evidence-Based Medicine at the University of Queensland, Australia, January 9, 2007. Another version of the paper was presented at the Annual 4S Conference in Montreal, Canada, 11 October 2007. Some funding for research for this paper was provided by a grant from the Australian Research Council, "Science litigation and the public accountability of vertically integrated expertise": DP0558176. Thanks also to Elizabeth Silk for invaluable research assistance and critical comments.

Notes

[1] The IRA trials of the Birmingham Six are one of the more prominent examples of a miscarriage of justice in the UK. These involved the conviction and sentencing to life imprisonment of six men for the bombing of two pubs in Birmingham in late 1974. The convictions took

place in 1975, and were upheld through several trials despite allegations against the police for forcing statements, assault against the accused and incorrect/poor scientific evidence. The sentences were finally overturned in 1991 following a third appeal (*Richard McIlkenny, Patrick Hill, William Power, John Walker, Robert Gerard Hunter and Hugh Callaghan* (1991) 93 Cr App R (CA) 287).

[2] The Chamberlain or "dingo baby" cases created significant controversy in Australia during the 1980s as one of the nation's most famous miscarriages of justice. The case involved the accusation that Mrs Chamberlain had killed her young child on a camping trip, which she denied; stating that she believed a dingo (a wild Australian canine) had taken the child. The Chamberlains were convicted of murder although in the appeal *Chamberlain v The Queen [No 2]* (1983–1984) 153 CLR 521, several judges dissented indicating that they had doubts as to the quality of the scientific evidence. The couple were acquitted several years later following the discovery of evidence of the child's clothing in an area frequented by the dogs.

[3] An *amicus curiae* (or "friend of the court") brief is a submission made to the court by a third party who seeks to aid the court by putting forth a view as to the correct interpretation of the law or by providing background information which may be of help to the court. In practice, such submissions are often lobby groups representing the interests of those who feel their interests may be affected by the decision in the case being considered by the court. For a discussion regarding the submission of *amici curiae* briefs in *Daubert* and related cases see Edmond and Mercer (1998b).

[4] It should also be noted that interpretations of the status of fingerprint evidence have also been influenced by the efforts of fingerprint examiners to reshape their practices and knowledge to better satisfy the demands of the *Daubert* criteria and emergence of competing forms DNA evidence which may be better able to satisfy demands for mechanical forms of objectivity. See, for example, Lynch (2004).

References

Australian Law Reform Commission. 1999. Experts: Adversarial (Background Paper 6, Section 3) [cited 8 September 2008]. Available from http://www.austlii.edu.au/other/alrc/publications/bp/6/; INTERNET.

Berg, Marc. 1995. Turning practice into a science: Reconceptualizing postwar medical practice. *Social Studies of Science* 25 (3): 437–476.

Bernstein, David. 1999. Commentary on the politics of jury competence by Gary Edmond and David Mercer. In *Technology and public participation,* edited by Brian Martin, pp. 107–108. Wollongong: Science and Technology Studies, University of Wollongong.

Brase, Twila. 2005. How technocrats are taking over the practice of medicine: A wake up call to the American people. Citizens' Council on Health Care Policy Report, January 2005 [cited 8 September 2008]. Available at http://www.cchonline.org/pdfreport/; INTERNET.

Brody, Howard, Franklin G. Miller, and Elizabeth Bogdan-Lovis. 2005. Evidence-based medicine: Watching out for its friends. *Perspectives in Biology and Medicine* 48 (4): 570–584.

Brown, Nik, and Andrew Webster. 2004. *New medical technologies and society: Reordering life.* Cambridge: Polity Press.

Caudill, David, and Lewis LaRue. 2006. *No magic wand: The idealization of science in law.* Lanham, MD: Rowman and Littlefield Publishers.

Charlton, Bruce, G. 1997. Restoring the balance: evidence-based medicine put in its place. *Journal of Evaluation in Clinical Practice* 3 (2): 87–98.

Cochrane, A. L. 1972. *Effectiveness and efficiency: Random reflections on health services.* London: Nuffield Provincial Hospitals Trust.

Cohen, Aaron. M., P. Z. Stavri, and William R. Hersh. 2004. A categorization and analysis of the criticism of Evidence-Based Medicine. *International Journal of Medical Informatics* 73: 35–43.

Cole, Simon. A. 2001. *Suspect identities: A history of fingerprinting and criminal identification.* Cambridge, MA: Harvard University Press.

Daubert. 1993. *Daubert v Merrell Dow Pharmaceuticals, Inc.,* 509 US 579, 113 S.Ct. 2786, 125L.Ed.2d 469.

Davidoff, F., Brian Haynes, David Sackett, and Richard Smith. 1995. Evidence-based medicine. *British Medical Journal* 310: 1085–1086.

Denny, Keith. 1999. Evidence-Based Medicine and medical authority. *Journal of Medical Humanities* 20 (4): 247–263.

Dickenson, Donna, and Paolo Vineis. 2002. Evidence-Based Medicine and quality of care. *Health Care Analysis* 10: 243–259.

Dixon, Lloyd, and Brian Gill. 2002. Changes in the standards for admitting expert evidence in federal civil cases since the *Daubert* decision. *Psychology, Public Policy, and Law* 8: 251–308.

Edmond, Gary. 2002a. Constructing miscarriages of justice: Misunderstanding scientific evidence in high profile criminal appeals. *Oxford Journal of Legal Studies* 22 (1): 53–89.

———. 2002b. Legal engineering: Contested representations of law, science (and non-science) and society. *Social Studies of Science* 32: 371–412.

———. 2003. After objectivity: Expert evidence and procedural reform. *The Sydney Law Review* 25 (2): 131–163.

Edmond, Gary, and David Mercer. 1996. Manifest destiny: Law and science in America. *Metascience* 10: 40–58.

———. 1998a. Trashing "junk" science. *Stanford Technology Law Review* 3: 1–32.

———. 1998b. Representing the sociology of scientific knowledge and law. *Science Communication* 19 (4): 307–327.

———. 2000. Litigation life: Law-science knowledge construction in (Bendectin) mass toxic tort litigation. *Social Studies of Science* 30 (2): 265–316.

———. 2002. Conjectures and exhumations: Citations of history, philosophy and sociology of science in US federal courts. *Law & Literature* 14 (2): 309–366.

———. 2004a. *Daubert* and the exclusionary ethos: The convergence of corporate and judicial attitudes towards the admissibility of expert evidence in tort litigation. *Law and Policy* 26 (2): 231–257.

———. 2004b. Experts and expertise in legal and regulatory settings. In *Expertise in Regulation and Law,* edited by Gary Edmond. Aldershot: Ashgate Press.

Ehrenreich, John, ed. 1978. *The cultural crisis of modern medicine.* New York: Monthly Review Press.

Foster, Kenneth, and Peter Huber. 1997. *Judging science: Scientific knowledge and the federal courts.* Cambridge, MA: MIT Press.

Freckelton, Ian. 1994. Contemporary comment: When plight makes right – the forensic abuse syndrome. *Criminal Law Journal* 18: 29–49.

Frye. 1923. *Frye v United States,* 293F. 1013 (DC Cir 1923).

Goodstein, David. 2000. How science works. In *Reference Manual on Scientific Evidence,* 2nd edition. Washington D.C. Federal Judicial Center [cited 8 September 2008]. Available from www.fjc.gov/public/pdf.nsf/lookup/sciman00.pdf/$file/sciman00.pdf; INTERNET.

Grossman, Jason, and Fiona Mackenzie. 2005. The randomized controlled trial: Gold standard, or merely standard? *Perspectives in Biology and Medicine* 48 (4): 516–534.

Guyatt, Gordon, Deborah Cook, and Brian Haynes. 2004. Editorial: Evidence based medicine has come a long way. *British Medical Journal* 329: 990–991.

Haack, Susan. 2001. An Epistemologist in the bramble-bush: At the Supreme Court with Mr Joiner. *Journal of Health Politics, Policy & Law*(26): 217–248.

Havighurst, Clarke C., Peter Barton Hutt, Barbara J. McNeil, and Wilhelmine Miller. 2001. Evidence: Its meanings in health care and in law. (Summary of the 10 April 2000 IOM and AHRQ Workshop, "Evidence": Its meanings and uses law medicine and health care). *Journal of Health Politics, Policy and Law* 26 (2): 195–215.

Hines, Nichole. 2006. Why technology provides compelling reasons to apply a *Daubert* analysis to the legal standard of care in medical malpractice cases. *Duke Law and Technology Review* 0018.

Huber, Peter. 1991. *Galileo's revenge: Junk science in the courtroom.* New York: Basic Books.

Hurwitz, Brian. 2004. How does evidence based guidance influence determinations of medical negligence? *British Medical Journal* 329: 1024–1028.

Illich, Ivan. 1976. *Limits to medicine: Medical nemesis, the expropriation of health.* Harmondsworth: Penguin.

Jackson, John Zen, 2006. Judging an inexact science; evidence-based medicine and the standard of care. *New Jersey Law Journal,* 18 September.

Jasanoff, Sheila. 1995. *Science at the bar: Science and technology in American law.* Cambridge, MA: Harvard University Press.

Jasanoff, Sheila. 2002. Science and the statistical victim. *Social Studies of Science* 32: 37–69.

Joiner. 1997. *General Electric Co. v Joiner,* 522 US 136, 139 L.Ed.2d 508.

Kassirer, Jerome P., and Joe S. Cecil. 2002. Inconsistency in evidentiary standards for medical testimony: Disorder in the courts. *Journal of the American Medical Association* 288: 1382–1387.

Kuhlman, Ellen. 2004. Standards, guidelines and evidence based medicine – bringing patients perspectives. Paper presonted at the Society for the Social Studies of Science, 4S EASST Conference, Public Proofs, Science Technology and Democracy, 25–28 August, Paris, France.

Kuhn, T. S. 1962. *The structure of scientific revolutions.* Chicago: Chicago University Press.

Kumho. 1999. *Kumho Tire Co. Ltd. v Carmichael,* 526 US 137, 143 L Ed 2d 238.

Landzelius, Kyra. 2006. Editorial introduction: Patient organization movements and the metamorphoses of patienthood. *Social Science and Medicine* 62 (3): 529–537.

Lynch, Michael. 2004. Science above all else: The inversion of credibility between forensic DNA profiling and fingerprint evidence. In *Expertise in Regulation and Law,* ed. Gary Edmond. Aldershot: Ashgate Press.

McQueen, Mathew J. 2001. Overview of evidence-based medicine: Challenges for evidence-based laboratory medicine. *Clinical Chemistry* 47 (8): 1536–1546.

Mercer, David. 2002a. Scientific method discourses in the construction of "EMF Science". *Social Studies of Science* 32 (2): 205–233.

———. 2002b. The intersection of sociology of scientific knowledge (SSK) and law: Some themes and policy reflections. Law Text Culture 6: 1–22.

———. 2004. Hyper-experts and the vertical integration of expertise in EMF/RF litigation. In *Expertise in Regulation and Law,* edited by Gary Edmond. Aldershot: Ashgate Press.

Michaels, David, and Celeste A. Monforton. 2005. Manufacturing uncertainty: Contested science and the protection of the public's health and environment. *American Journal of Public Health* 95 (S1): S39–S48.

Miller, Donald W., and Clifford G. Miller. 2005. On evidence, medical and legal. *Journal of American Physicians and Surgeons* 10 (3): 70–75.

Mulrow, Cynthia D., and Kathleen N. Lohr. 2001. Proof and policy from medical research evidence. *Journal of Health Politics Policy and Law* 26 (2): 249–266.

Oakley, Anne. 2000. *Experiments in knowing.* New York: The New Press.

Porter, Roy. 1998. *The greatest benefit to mankind: A medical history of humanity.* New York: W. W. Norton and Co.

Porter, Theodore. M. 1995. *Trust in numbers.* Princeton, NJ: Princeton University Press.

Richards, Evelleen. 1991. *Vitamin C and cancer: Medicine or politics?* London: Macmillan.

Rogers, Wendy A. 2002. Is there a tension between doctors' duty of care and evidence-based medicine? *Health Care Analysis* 10 (3): 277–287.

Rosoff, Arnold J. 2001. Evidence-based medicine and the law: The courts confront clinical practice guidelines. *Journal of Health Politics, Policy and Law* 26 (2): 327–368.

Rudd, Ter Meulen, and Donna Dickenson. 2002. Into the hidden world behind evidence-based medicine. *Health Care Analysis* 10 (3): 231–241.

Saks, Michael. J. 1992. Do we really know anything about the behaviour of the tort litigation system – and why not? *University of Pennsylvania Law Review* 140: 1147–1291.

Saks, Michael. J., and David L. Faigman. 2005. Expert evidence after *Daubert*. *Annual Review of Law and Social Science* 1: 105–30.

Sackett, D. L., William M. C. Rosenburg, J. A. Muir Grey, R. Brian Haynes, and W. Scott Richardson. 1996. Evidence based medicine: What it is and what it isn't. *British Medical Journal* 312: 71–72.

Sackett, D. L. 2000. The sins of expertness and a proposal for redemption. *British Medical Journal* 320: 1283.

Samanta, Ash, Jo Samanta, and Michael Gunn. 2003. Legal considerations of clinical guidelines: Will NICE make a difference? *Journal of the Royal Society of Medicine* 96: 133–138.

Sanders, Joseph. 1998. *Bendectin on trial: A study of mass tort litigation.* Ann Arbor: University of Michigan Press.

Schuck, Peter. 1986. *Agent Orange on trial: Mass toxic tort disasters in the courts.* Cambridge, MA: Harvard University Press.

Schuster, John, and Richard Yeo, eds. and comps. 1986. *The politics and rhetoric of scientific method: Historical studies.* Dordrecht: Reidel.

Sehon, Scott. R., and Donald E. Stanley. 2003. A philosophical analysis of the evidence based medicine debate. *BMC Health Services Research* 3: 14 [cited 8 September 2008]. Available from http://www.biomedcentral.com/1472–6963/3/14; INTERNET.

Tallacchini, Mariachiara. 2002. Legalising science. *Health Care Analysis* 10 (3): 329–337.

Timmermans, Stefan. 2005. From autonomy to accountability: The role of clinical practice guidelines professional power. *Perspectives in Biology and Medicine* 48 (4): 490–501.

Timmermans, Stefan, and Aaron Mauck. 2005. The promises and pitfalls of evidence-based medicine. *Health Affairs* 24 (1): 18–28.

Vos, Rein, Rob Houtepen, and Klasien Horstman. 2002. Evidence-based medicine and power shifts in health care systems. *Health Care Analysis* 10 (3): 319–328.

Williams, C. L. 2004. Evidence–based medicine in the law beyond clinical practice guidelines: What effect will EBM have on the standard of care? *Washington and Lee Law Review* 61: 479–532.

Woolf, H 1996. *The right honourable the Lord Woolf, master of the rolls, access to justice—final report* [cited 8 September 2008]. Available from *http://www.dca.gov.uk/civil/final/*index.htm/; INTERNET.

Wright, P. and A. Treacher, eds. and comps. 1982. *The problem of medical knowledge: Examining the social construction of medicine.* Edinburgh: Edinburgh University Press.

Wynne, Brian. 1982. *Rationality and ritual.* Chalfont St Giles, UK: BHS Monographs.

Yearley, Steven. 2005. *Making sense of science: understanding the social study of science.* London: Sage.

Social Epistemology
Vol. 22, No. 4, October–December 2008, pp. 425–431

Routledge
Taylor & Francis Group

The "EBM Movement": Where Did it Come From, Where is it Going, and Why Does it Matter?

Wendy Lipworth, Stacy M. Carter and Ian Kerridge

Evidence-Based Medicine (EBM) has now been part of the dominant medical paradigm for 15 years, and has been frequently debated and progressively modified. One question about EBM that has not yet been considered systematically, and is now particularly timely, is the question of the novelty, or otherwise, of the principles and practices of EBM. We argue that answering this question, and the related question of whether EBM-type principles and practices are unique to medicine, sheds new light on EBM and has practical implications for those involved in all EBM. This is because one's answer to the question (whether explicit or implicit) affects the amount and type of funding and attention received by EBM, the extent to which EBM, and the generation, judgment and use of evidence more generally, can be appropriated by certain groups and questioned by others, and the extent to which truly unique socio-political developments in evidence, and in medicine more generally, are recognized and harnessed.

Keywords: Evidence-Based Medicine; Exceptionalism

Concerns and Debates about Evidence-Based Medicine

The phrase "evidence-based medicine" and the acronym "EBM" have been in common usage since the 1990's, at which time they were introduced as a means of formalizing

Wendy Lipworth is a researcher with a background in clinical medicine, pathology, policy, bioethics and qualitative research. Her current interests lie in the ethics of biomedical publication, and in the ethics of tissue banking research. Stacy Carter's career has included seven years as a speech pathologist in the NSW Hospital system, several years of project management in the NGO health sector, a MPH (Honours) and a PhD in public health. She is now senior lecturer in Qualitative Research in Health at the Centre for Values, Ethics & the Law in Medicine and the School of Public Health at the University of Sydney. With her colleagues she conducts qualitative research mostly focused on risk, health promotion and cancer. Ian Kerridge trained in medicine, philosophy and haematopoietic stem cell transplantation. He is Director and Associate Professor in Bioethics at the Centre for Values, Ethics and the Law in Medicine at the University of Sydney and Staff Haematologist/Bone Marrow Transplant physician at Westmead Hospital, Sydney. Correspondence to: Wendy Lipworth, Centre for Values, Ethics and the Law in Medicine, University of Sydney, Camperdown, NSW 2006, Australia. Email: wendylipworth@yahoo.com.au

ISSN 0269–1728 (print)/ISSN 1464–5297 (online)
DOI: 10.1080/02691720802559446

Archie Cochrane's proposal to privilege evidence over idiosyncratic clinical judgment (Anonymous 1992). During this time, EBM principles and practices have come to have a profound influence on the setting of biomedical research priorities, the generation of public health and clinical practice guidelines and the implementation of these guidelines in practice. At present, all funders and publishers of biomedical research and all policymakers and practitioners of clinical and public health medicine are expected to understand and implement the principles of EBM.

Since its inception, EBM has been the subject of considerable professional and political debate and its principles and practice have evolved over time. There is, for example, an increasing recognition of the importance of integrating clinical expertise and patient values into evidence-based practice (Sackett et al. 2000), of the need to avoid "cookbook" and "defensive" EBM practice (Sackett et al. 1996) and of the need to challenge the traditional hierarchy that privileges randomized trials over all other study designs (Grossman and Mackenzie 2005). Whether these concerns have been addressed adequately is open to debate, with some seeing contemporary EBM as a relatively unproblematic practice and others, particularly those with a more philosophical or sociological bent, seeing EBM as a potentially problematic epistemological and socio-political movement (Goldenberg 2006).

In this paper we will argue that now is the perfect time to address systematically one of the many questions that has been asked about EBM: whether EBM (or, more accurately, its principles and practices) truly represents anything new about medicine, and the related question of whether EBM-type principles and practices are unique to medicine. Put another way, we need to address *systematically*, using a theory of perceptions of novelty and uniqueness, the question of whether EBM is "old hat" or whether EBM truly represents a paradigm shift in biomedical thought and practice (Sackett et al. 1996).

At first glance, this may seem to be a relatively academic question which is perhaps less important than questions about the kind(s) of evidence, and therefore research priorities, that are privileged within EBM (Grossman and Mackenzie 2005; Little 2003; Tonelli 2001) and questions about the ways in which EBM can and should be incorporated into clinical and public health policy and practice (Sackett et al. 1996; Little 2003; Tonelli 2001). But we argue that perceptions of novelty and uniqueness have broad socio-political effects for all of those involved with EBM at the level of research, policy, and practice. These socio-political effects include: the amount and type of funding and attention received by EBM; the extent to which EBM, and the generation, judgment, and use of evidence more generally, can be appropriated by certain groups and questioned by others; and the extent to which truly unique socio-political developments in evidence, and in medicine more generally, are recognized and harnessed.

EBM and Exceptionalism

To understand the link between perceptions of novelty and uniqueness and these broad socio-political implications, it is useful to draw on the sociological notion of "exceptionalism". This provides a means of linking the conceptualization of emerging

phenomena—in particular judgments about their novelty and uniqueness—with the status given the phenomenon, its impacts and the way in which it is controlled.

The notion of "exceptionalism" has been used by political scientists to describe the perception or claim that a particular social or political system (e.g. "America") (Koh 2004) or religion (Lakoff 2004) is unique and has developed in unique ways. Bayer introduced the term into biomedicine when he used it to describe claims that HIV was a unique disease warranting a unique public health response (Bayer 1991). It has since been applied to debates about the uniqueness, or otherwise, of genetic research and testing as compared to other types of medical research and testing (Hodge 2004). In relation to EBM, therefore, an exceptionalist stance would hold that EBM is unique to, and/or novel within medicine, and a non-exceptionalist stance would emphasize the overlap between EBM and other evidence-related practices either outside of medicine, or within medicine prior to the advent of "EBM". Debates about exceptionalism have evolved because it has been recognized that these stances have profound socio-political implications.

Exceptionalist judgments (i.e. judgments that an emerging phenomenon, such as EBM is unique and/or novel) have both advantages and disadvantages. The perception that something is new, unique and therefore "special" can create a healthy interest in the phenomenon, but it can also lead to a detrimentally exclusive focus on the phenomenon, as well as unnecessary fear and mystique. To say, for example, that genetic testing is a unique and novel kind of medical testing makes genetics seem interesting to funders of research and regulators of genetic testing. This is good for those with genetic diseases, but it also has the potential to disadvantage those with other kinds of diseases, whose concerns are relatively sidelined (Suter 2001), and it can result in an arguably unwarranted fear of genetic testing. EBM exceptionalism might, therefore, result in a healthy interest in the generation and application of evidence in medicine. Indeed, it has been observed recently that:

> The evidence-based medicine (EBM) movement is touted as a new paradigm in medical education and practice, a description that carries with it an enthusiasm for science that has not been seen since logical positivism flourished (circa 1920–1950). (Goldenberg 2006, 2621)

But the resulting "EBM movement" may also draw attention away from other aspects of medicine, and the notion that EBM is new and "special" may make EBM seem unnecessarily mystical and inaccessible to all but the most specialized theorists and practitioners. This could account for the complaint by some that EBM has taken on an inappropriate degree of centrality in medical education and practice, sidelining other concerns, and that EBM has achieved "cult status" and cannot be questioned.[6] (Little 2003)

Exceptionalist judgments can result also in the political "hijacking" of the emerging phenomenon by particular interest groups, since what is new and unique can be more easily appropriated and "owned". This can be good if a sense of ownership motivates action and promotes a sense of responsibility, but the effects are not always positive. Everett complains about the hijacking of the genetics policy agenda by the "genetics

privacy movement" (Everett 2004) and it is conceivable that EBM exceptionalism has resulted in a similar concentration of power in the hands of those who accept and have expertise in the principles and practices of EBM, and has marginalized those who do not (Little 2003).

Non-exceptionalist judgments (i.e. judgments that an emerging phenomenon, such as EBM, is part of the incremental development of a field) also have both advantages and disadvantages. On the positive side, non-exceptionalism allows practitioners to draw on insights from elsewhere rather than "re-inventing the wheel". If, for example, the similarities between genetic testing and other forms of medical testing are recognized, then insights from our understanding of the issues relating to other medical tests, and our regulation of these testing procedures, can be applied to genetics, rather than generating a morass of unnecessary and possibly incomplete genetics-specific regulation and practice. Similarly, it could be very useful to draw on the rich insights and practical strategies of those who, throughout the history of medicine, have reflected on the role of evidence in medical practice (Little 2003). And it could be very useful to recognize the similarities between the evidence-generating and evidence-applying principles of EBM and those of law, engineering and politics. Sophisticated debates about the nature and application of evidence have taken place, for example in relation to expert testimony in law (Sartore and van Doren 2006) and in relation to the political assessment and management of scientific evidence of environmental and technological risks (Jasanoff 2005), and these debates could be applied to similar problems in medicine.

On the other hand, taking a non-exceptionalist stance (particularly if this stance is naïve and ill-considered) can prevent us from focusing on what is truly unique about an emerging phenomenon. To say that genetic testing is simply an extension of other testing procedures is to lose sight of the fact that genetic testing does, arguably, raise novel issues relating to the stability of the information throughout life, the potential for information to be generated about family members, etc. Similarly, to say that EBM is simply the latest iteration of evidence-use in medicine might lead us to lose sight of what is truly new and unique about modern EBM, such as its specific prioritization of some, relatively new, forms of evidence-generation (RCTs and meta-analyses) over others and its grounding in "post-genomic" medicine, with its generation of an unprecedented amount of raw data that needs to be translated (or not) into clinically-relevant evidence.

Given these implications of both exceptionalism and non-exceptionalism, it is crucial to be aware of whether one is taking an exceptionalist or a non-exceptionalist stance. What needs to be avoided is a *naive* position in either direction in which the effects of exceptionalism or non-exceptionalism are hidden, which in turn means that the advantages cannot be maximized and the disadvantages cannot be managed (Lipworth 2005).

The Challenge of Considering Questions About Exceptionalism

In relation to EBM, a critical (i.e. non-naïve) stance on EBM's exceptionalism or non-exceptionalism would require:

1) carefully defining EBM;
2) systematically comparing the principles and practices of EBM to older medical practices and to practices outside of medicine such as law, engineering and politics;
3) deciding, on the basis of the above, whether an exceptionalist, non-exceptionalist or mixed exceptionalist/non-exceptionalist stance should be taken; and
4) harnessing the positive effects and managing the negative effects of the chosen stance.

While the issue of novelty has previously been raised in relation to EBM (Sackett et al. 1996) there is limited evidence that the above steps have been carried out systematically in relation to EBM, and there are a number of reasons why this may be the case.

First, there is no single agreed-upon definition of "EBM". To any individual, the phrase "Evidence-Based Medicine" may refer to one, several or all of: a hierarchy of clinical research methods for generating evidence, a way of evaluating existing clinical research, a method of translating research evidence (and perhaps clinical expertise and patient values) into clinical and public health practice or more abstractly, as a social movement or philosophical construct privileging some forms of knowledge and judgment over others (Ankeny and Mackenzie 2003).

Second, even where a phenomenon such as EBM is clearly defined, questions about uniqueness and novelty are not always easy to resolve. The ongoing question, for example, of whether HIV is in any way a unique phenomenon, warranting a unique public health response, depends upon whether one considers HIV's scale, its affected populations, its mode of spread and its testing procedures to be sufficiently different to those of other infectious diseases. These assessments depend upon one's understanding of HIV infection as well as one's sense of what degree of difference constitutes true novelty and uniqueness. To some, therefore, the formalization of evidence generation and use that has taken place under the banner of EBM is truly a paradigm shift, perhaps even on the scale of a scientific revolution (Anonymous 1992), but to others this may seem to be simply the latest step in the always evolving philosophy and practice of medicine (Little 2003).

A third challenge for those wishing to ask questions about uniqueness and novelty is that such assessments change over time. When HIV first emerged, it was generally believed that infectious diseases had been conquered by medicine and that retroviruses did not cause human disease or cancer, so it is perfectly understandable that HIV would have been considered a unique phenomenon (Gallo and Montagnier 2003, Rosenbrock et al. 2000). But now we have a different understanding of the disease and can, from a biomedical perspective at least, more easily see HIV as just one of many infectious diseases, perhaps with a few unique features. Similarly, EBM, as it emerged, was an unstable and apparently novel phenomenon that could, perfectly understandably be seen as something entirely new, with the nuances of its overlap with older movements in medicine being temporarily obscured.

Practical Implications for Policymakers

Despite these challenges, we argue that, given the conceptual and socio-political implications of naïve exceptionalism and non-exceptionalism, the exercise of asking oneself

what one means by EBM, and whether one sees EBM as being novel and/or unique, is an exercise worth undertaking by anyone involved in "EBM". Indeed, now that EBM is into its third decade, and is well established and relatively stable, this is the perfect time to undertake this exercise.

In order for this process to occur those who set EBM-based research priorities and those who develop EBM-based practice guidelines need to ask themselves what they mean by EBM today and, as a result, whether they wish to take an exceptionalist, non-exceptionalist, or mixed exceptionalist/non-exceptionalist stance on their practice. Once this has been achieved, EBM policies (i.e. the policies driving EBM-based research and practice) and educational materials (i.e. documents used to sell and teach EBM to researchers, policymakers and practitioners) could be reviewed in light of these considerations so that the advantages of the stance are harnessed and the disadvantages are managed.

If an exceptionalist stance is taken, EBM policies and educational materials could emphasize EBM's novelty and uniqueness, harness the interest that is generated by claims of uniqueness and novelty, and incorporate strategies to prevent EBM displacing other concerns in medicine, attaining "cult" status, becoming inaccessible and unquestionable, and being appropriated by only a few powerful groups. If, on the other hand, a non-exceptionalist stance is taken, then EBM policies and educational materials could emphasize its inclusivity and draw explicitly on areas of overlap with other practices both within and outside of medicine. Indeed, a strongly non-exceptionalist stance might even result in the view that the label "EBM" has outlived its usefulness, since it does not represent anything unique or novel, and EBM non-exceptionalists might argue for a return to more foundational, pre "EBM" discussions of knowledge, evidence, judgment and values in medicine. We may, for example, wish to question the privileging of RCTs over other study designs (Grossman and Mackenzie 2005) and, more philosophically, we may wish to question the notion of evidence as "facts" about the world, in light of which scientific beliefs, and medical practices, should stand or fall (Goldenberg 2006).

Whatever stance is taken, we hope that this process of carefully defining modern EBM, and considering it in light of exceptionalism will assist not only EBM policymakers, but also researchers, practitioners, patients, and the general public, in navigating a terrain which is necessarily complex, but can be made less obscure by minimizing conceptual ambiguity and highlighting the socio-political implications of conceptual choices.

References

Ankeny, R. F., and F. Mackenzie. 2003. Commentary. In *A Miles Little Reader. Restoring humane values to medicine,* edited by I. Kerridge, C. Jordens, and E. Sayers, pp. 118–21. Sydney: Desert Pea Press.

Anonymous. 1992. Evidence-based medicine. A new approach to teaching the practice of medicine. Evidence-Based Medicine Working Group. *JAMA* 262: 420–25.

Bayer, R. 1991. Public health policy and the AIDS epidemic. An end to HIV exceptionalism? *New England Journal of Medicine* 324: 1500–4.

Everett, M. 2004. Can you keep a (genetic) secret? The genetic privacy movement. *Journal of Genetic Counseling* 13: 273–91.

Gallo, R., and L. Montagnier. 2003. Retrospective: The discovery of HIV as the cause of AIDS. *New England Journal of Medicine* 349: 2283–5.

Goldenberg, M. J. 2006. On evidence and evidence-based medicine: Lessons from the philosophy of science. *Social Science & Medicine* 62: 2621–32.

Grossman, J., and F. J. Mackenzie. 2005. The randomized controlled trial: Gold standard, or merely standard? *Perspectives in Biology and Medicine* 48: 516–34.

Hodge Jr, J. 2004. Ethical issues concerning genetic testing and screening in public health. *American Journal of Medical Genetics* 125C: 66–70.

Jasanoff, S. 2005. In the democracies of DNA: Ontological uncertainty and political order in three states. *New Genetics and Society* 24: 139–55.

Koh, H. 2004. On America's double standard: The good and bad faces of exceptionalism. *The American Prospect* 15: A16–19.

Lakoff, S. 2004. The reality of Muslim exceptionalism. *Journal of Democracy* 15: 133–9.

Lipworth, W. 2005. Generating a taxonomy of regulatory responses to emerging issues in biomedicine. *Journal of Bioethical Inquiry* 2: 130–41.

Little, M. 2003. "Better than numbers…" A gentle critique of evidence-based medicine. *ANZ Journal of Surgery* 73: 177–82.

Rosenbrock, R., F. Dubois-Arber, M. Moers, P. Pinell, D. Schaeffer, and M. Setbon. 2000. The normalization of AIDS in Western European countries. *Social Science & Medicine* 50: 1607–1629.

Sackett D. L., W. M. Rosenberg, J. A. Gray, R. B. Haynes, and W. S. Richardson. 1996. Evidence based medicine: What it is and what it isn't. *British Medical Journal* 312: 71–2.

Sackett, D. L., S. E. Straus, W. S. Richardson, W. Rosenberg, and R. B. Haynes. 2000. *How to practice and teach EBM.* New York: Churchill Livingstone.

Sartore, J. T., and R. van Doren. 2006. Daubert opinion requires judges to screen scientific evidence. *Pediatrics* 118: 2192–4.

Suter, S. 2001. The allure and peril of genetics exceptionalism: Do we need special genetics legislation? *Washington University Law Quarterly* 79: 669–748.

Tonelli, M. R. 2001. The limits of evidence-based medicine. *Respiratory Care* 46: 1435–40.

Social Epistemology
Vol. 22, No. 4, October–December 2008, pp. 433–460

Routledge
Taylor & Francis Group

Minimum Message Length and Statistically Consistent Invariant (Objective?) Bayesian Probabilistic Inference—From (Medical) "Evidence"

David L. Dowe

"Evidence" in the form of data collected and analysis thereof is fundamental to medicine, health and science. In this paper, we discuss the "evidence-based" aspect of evidence-based medicine in terms of statistical inference, acknowledging that this latter field of statistical inference often also goes by various near-synonymous names—such as inductive inference (amongst philosophers), econometrics (amongst economists), machine learning (amongst computer scientists) and, in more recent times, data mining (in some circles).

Three central issues to this discussion of "evidence-based" are (i) whether or not the statistical analysis can and/or should be objective and/or whether or not (subjective) prior knowledge can and/or should be incorporated, (ii) whether or not the analysis should be invariant to the framing of the problem (e.g. does it matter whether we analyse the ratio of proportions of morbidity to non-morbidity rather than simply the proportion of morbidity?), and (iii) whether or not, as we get more and more data, our analysis should be able to converge arbitrarily closely to the process which is generating our observed data.

For many problems of data analysis, it would appear that desiderata (ii) and (iii) above require us to invoke at least some form of subjective (Bayesian) prior knowledge. This sits uncomfortably with the understandable but perhaps impossible desire of many medical publications that at least all the statistical hypothesis testing has to be classical non-Bayesian—i.e. it is not permitted to use any (subjective) prior knowledge.

Keywords: Minimum Message Length; MML; Inference; Bayesianism; Statistical Invariance; Statistical Consistency; Evidence; Evidence-Based Medicine

David L. Dowe is Associate Professor of Computer Science at Monash University in suburban Melbourne. Correspondence to: Clayton School of Information Technology, Monash University, Clayton, Vic. 3800, Australia. Email: david.dowe@infotech.monash.edu.au.

ISSN 0269–1728 (print)/ISSN 1464–5297 (online)
DOI: 10.1080/02691720802576291

Introduction

Data is collected in medical and other scientific studies to provide "evidence" in support of or against a variety of hypotheses. Ultimately, we hope that collection and analysis of such data evidence in turn both enables us to accurately infer any underlying process from which the data is generated and also to accurately predict as yet unmeasured outcomes.

We will examine here several desirable properties—or desiderata—for a statistical inference technique in analysing medical and other data. We will address the issue of whether or not all of these desiderata can be simultaneously satisfied and when some sort of trade-off might be necessary.

One property that we want from our statistical inference technique is that of **statistical consistency**. Informally, this says that as the amount of data collected increases, we converge closer and closer and arbitrarily close to whatever underlying process can be said to be generating the data. Single and multiple latent factor analysis are but a few examples of problems for which frequently-used modelling tools are statistically inconsistent.

Another property which we want from our statistical inference tool is the ability to make probabilistic models and accurately quantify noise. Diagnoses such as "yes"/"no" or "presence"/"absence" of some condition are less useful than models which give a probability of a diagnosis. Rather than respond with "no", a system returning a probability of (say) 10% of some condition enables the medical experts to decide upon possible treatment and further tests; and certainly there is much more difference between (say) 10% and 45% than there is between two (less informative) responses of "no".

Another property which we presumably also want from our statistical inference tool is that of **statistical invariance**—namely, that the inferred value is independent of the framing of the problem. By "*framing*", I don't particularly mean linguistic framing but rather a statistical or (statistically) parametric framing. (For example, variations of Bertrand's paradox say that in a cube of side-length between 1 and 2, the side-length has probability 1/2 of being less than 1.5 but the volume has probability 1/2 of being less than 4.5, which—paradoxically?—is not 1.5^3.) To elaborate, if we know a skin lesion or tumour to be circular, then statistical invariance would require that the estimated area is equal to π times the square of the estimated radius. Whether we parameterise in terms of radius or area, we get the same answer.

Perhaps the single main other issue to mention in the use of "evidence" is the difference between *inference* and *prediction*. Inference is the use of one—ideally the "best"—theory to model the observed data and find a pattern within it. Prediction is concerned with forecasting as yet unseen data. Unless the currently observed data has one outstanding single best theory, prediction is often best done by combining more than one theory.

Desiderata in (Probabilistic) Inference and (Probabilistic) Prediction

Data can be both time-consuming and expensive to collect and obtain. It is often useful to know the accuracy to which the data was measured (Wallace and Dowe 1993, 1–3,

1994, 38, secs 2 and 2.1, 2000, sec. 2, 74, col. 2; Dowe, Allison et al. 1996, sec. 2; Kissane, Bloch, Dowe et al. 1996, 651; Comley and Dowe 2003, sec. 9, 2005, sec. 11.3.3, 270; Fitzgibbon, Dowe and Vahid 2004, eqn (19); Wallace 2005, secs 3.1.1 and 3.3; Dowe, Gardner and Oppy 2007; Dowe 2008, sec. 0.2.4)—as, after all, no-one knows their height or weight to infinitely many decimal places (if such a notion were even to make sense). It is also important to make good use of the data—whether we are making some sort of (probabilistic) inference, doing some sort of (probabilistic) prediction or perhaps (Dowe 2008, sec. 0.2.5) doing some kind of hypothesis test.

When doing (probabilistic) inference to some hypothesis, H, from (observed) data, D, we look at some possible desiderata, or properties that we might desire in our inference technique(s).

Statistical Invariance

Many problems can be phrased in several equivalent ways. Informally, *statistical invariance* says that we infer the same answer no matter how we phrase (or parameterise) the problem. Let us give several examples to clarify this point:

1. if p is the proportion of the population with a certain condition (or illness, diagnosis, prognosis, etc.) and q is the relative "odds ratio" proportion of those thus affected divided by those unaffected, then $q = p/(1-p)$ and $p = q/(1+q)$;
2. if r and A are the radius and area of a circle respectively through which an epidemic has spread (or, alternatively, of a surface lesion), then $A = \pi r^2$ and $r = \sqrt{A/\pi}$;
3. if a cube (maybe call it C) has side length l, face area A and volume V, then $l = A^{1/2} = V^{1/3}$, $A = l^2 = V^{2/3}$ and $V = l^3 = A^{3/2}$;
4. if a vector in the plane (such as direction and strength of a magnetic field) has direction θ and distance (or strength), κ (in polar co-ordinates) and can also be thought of (in Cartesian co-ordinates) as (x, y), then $(x, y) = (\kappa \cos\theta, \kappa \sin\theta)$ and[1] $(\kappa, \theta) = (\sqrt{x^2+y^2}, \tan^{-1}(y/x))$.

In the language of statistical inference, $\hat{\theta}$ denotes the estimated value of θ. The hat (or circumflex), ˆ, denotes an estimated value. Recall that, informally, statistical invariance says that we get the same answer no matter how we phrase (or parameterise) the problem. So, for example, with item 1 above, statistical invariance of an estimator would give us that $\hat{q} = \hat{p}/(1-\hat{p})$ and equivalently $\hat{p} = \hat{q}/(1+\hat{q})$. Not all problems have a "natural" parameterisation (or framing), so if we don't have statistical invariance then we have to get a different estimate for each re-parameterisation (or re-framing), potentially leading to awkward situations where for some cube (as in item 3) we might perhaps rather curiously estimate poorly matching values such as (e.g.) $\hat{l} = 0.98$, $\hat{A} = 1.03$ and $\hat{V} = 0.97$.

Notice also that many notions of "error", like bias and squared error, are not invariant to re-parameterisation. However, the notion of Kullback–Leibler divergence (or Kullback–Leibler distance) from the next section is one measure which is invariant under re-parameterisation.

Kullback–Leibler Divergence (or Kullback–Leibler Distance)

The Kullback–Leibler divergence is a measure of the difference between two probability distributions. It has the property of being invariant to re-parameterisation. The divergence is typically not symmetrical, though, meaning that the distance from distribution f to distribution g is not necessarily the same as the distance from distribution g to distribution f. This lack of symmetry is why some prefer the term "*divergence*" to "*distance*".

If f and g are both discrete distributions, with probabilities $f_1, \ldots, f_N$ and $g_1, \ldots, g_N$ for an N-state multinomial distribution (such as a two-sided coin with $N = 2$, or a six-sided dice with $N = 6$), then the Kullback–Leibler distance from f to g is defined as $KL(f,g) = \Delta(g\|f) = \sum_{i=1}^{N} f_i \log(f_i / g_i)$.

If f and g are both continuous-valued distributions, then we replace the summation by an integral and the Kullback–Leibler distance from f to g is defined as $KL(f, g) = \Delta(g||f) = \int f \log(f/g)$.

The Kullback–Leibler distance between two Bayesian networks (or graphical models) f and g can be defined as in Tan and Dowe (2006, sec. 4.2) or Dowe (2008, sec 0.2.5) and the Kullback–Leibler distance between two mixture models can be defined similarly. And there is no problem having a hybrid of both discrete- and continuous-valued variables.

Statistical Consistency

Informally, statistical consistency says that, as we get more and more data, we converge more and more closely—and, ultimately, arbitrarily closely—to the true underlying model. More formally, if θ is a parameter value, N is a sample size and $\hat{\theta}$ is a parameter estimate from a sample of size N, then *statistical consistency* says that

$$\forall \theta \, \forall \in > 0 \, \exists N_0 \, \forall N \geq N_0 \; Pr(|\theta - \hat{\theta}| < \in) > 1 - \in .$$

In other words, as we get more and more data, then with arbitrarily large probability we can converge arbitrarily closely to any true underlying model. Given our intuition that more and more data should enable us to infer more and more accurately, and given how expensive and time-consuming it can be to collect data, statistical consistency—that more data will ultimately take us to the correct answer—seems like one of the very least things we should seek in an inference method.

The notion of statistical consistency raises at least four other issues. First, it raises the issue of *efficiency* (Wallace 2005, sec. 3.4.5; Dowe, Gardner and Oppy 2007, sec. 8; Dowe 2008, sec. 0.2.5, especially footnote 162), the idea of not just converging on the true model (consistency) but of converging on the true value as quickly—or as efficiently—as possible. A second issue, perhaps subtly different to (asymptotic) efficiency, is that of—loosely speaking—performing well on small sample sizes. Statistical consistency guarantees asymptotic convergence, and (asymptotic) efficiency guarantees performing as well as possible on large (asymptotic) sample sizes. Efficiency and excellent small-sample performance are surely related, but surely also not identical. Third, it raises the issue of consistency when the model is misspecified (or, equivalently, when the true

model is not in the class of models being considered by the estimators) (Grünwald and Langford 2007; Dowe 2008, sec. 0.2.5). Given that many, if not perhaps most, inference problems have to contend with misspecification (e.g. Normal distributions are often used to model heights and other variables that can't take negative values), misspecification and inference methods which might or might not be susceptible to its vagaries (Grünwald and Langford 2007; Dowe 2008, sec. 0.2.5) should be paid greater attention. Fourth (and last), there is the issue of methods which are statistically consistent for (easier) problems where the number of parameters remains fixed but which do or don't remain statistically consistent for (harder) problems where the number of parameters increases as the amount of data increases (to the point where the amount of data per parameter is always bounded above, as in section "Amount of Data per Parameter Bounded Above"). It is known that some inference methods (such as Maximum Likelihood and AIC from sections "Maximum Likelihood" and "Akaike's Information Criterion (AIC) and Penalised (Maximum) Likelihood" often become statistically inconsistent in such cases (Neyman and Scott 1948), while at least one other method (MML from section "Minimum Message Length") appears to remain statistically consistent (Dowe and Wallace 1997; Wallace 2005, secs 4.2–4.5; Dowe, Gardner and Oppy 2007, secs 6.1 and 8; Dowe 2008, sec. 0.2.5).

Probabilistic Inference—vs. Mere Non-probabilistic Classification

Many inference problems are in ([supervised or] extrinsic) classification and—especially within the machine learning community—these are often regarded as problems of right vs. wrong. Probabilities are often neglected, whereas for certain ([moderately] serious) medical conditions the threshold for further investigation or treatment might well be other than 50%. If a patient presenting with chest pains were deemed to "only" have a 40% probability of heart attack or even deemed to "only" have a probability of 15% of heart attack, it would be somewhere along the lines of irresponsible, negligent and legally challenging to classify this as a "no" and not give treatment. This remains true whether the patient was presenting in person or telephoning a service such as *Nurse On Call* for a provisional symptom-based assessment over the phone. Even in DNA microarray classification, it is more prudent and probably also safer to report probabilities. While "right"/"wrong" is a fairly easy and seemingly natural scoring system to use, it is not invariant to re-framing of questions. As an example, consider a four-class problem which can be divided in three reasonable different ways into two two-class problems. The "right"/"wrong" score will depend upon the relevant division. However, probabilistic inference with log-loss scoring (Dowe and Krusel 1993, 4, Table 3; Dowe et al. 1998, sec. 3; Needham and Dowe 2001, Figs 3–5; Tan and Dowe 2002, sec. 4, 2004, sec. 3.1, 2006, secs 4.2–4.3; Kornienko, Dowe and Albrecht 2002, Table 2; Comley and Dowe 2003, sec. 9, 2005, sec. 11.4.2; Tan and Dowe 2003, sec. 5.1; Kornienko, Albrecht and Dowe 2005a, Tables 2–3, 2005b; Tan, Dowe and Dix 2007, sec. 4.3; Dowe 2008, sec. 0.2.5, especially footnote 175 [and 176]) not only scores probabilities, but it also has the desirable feature that the optimal long-term strategy is to give the true probabilities (if known).

If you assign probabilities $\{p_i : i = 1, \ldots, N\}$ for events $\{e_i : i = 1, \ldots, N\}$ such that $p_i \geq 0$ and $\sum_{i=1}^{N} p_i = 1$, then for some constant c (however chosen), if event e_j is the event that actually happened, then log-loss (or "*probabilistic bit-cost*") scoring awards a score of $c + \log p_j$. This scoring system has been used for Australian Football League (AFL) matches since early 1995 (Dowe, Farr et al. 1996; Dowe et al. 1998, sec. 3; Dowe 2008, sec. 0.2.5). With a probability of p on one team (and $1 - p$ on the other—in a match between two teams), using the constant $c = 1$, this competition at www.csse.monash.edu.au/~footy gives scores of $1 + \log_2 p$ if you're right, and $1 + \log_2 (1 - p)$ if you're wrong.

Log-loss scoring is invariant under re-framing of the problem (Dowe 2008, sec. 0.2.5, especially footnote 175 [and 176]), and appears—rather importantly—to enjoy the property of being unique in this respect.

And just as log-loss scoring appears to be unique in its invariance under re-framing of the problem, so, too, in some sense the (analogous) Kullback–Leibler divergence from section "Kullback–Leibler Divergence (or Kullback–Leibler Distance)" seems to also be unique in retaining invariance under re-framing of a problem. Both $KL(f, g) = \Delta(g||f)$ and $KL(g, f) = \Delta(f||g)$ are invariant to the level of detail of re-framing of the problem and appear to be unique in having this property—although, clearly, any linear combination $\alpha KL(f, g) + (1 - \alpha)KL(g, f)$ (with $0 \leq \alpha \leq 1$) will also share this invariance. (Interestingly, the difference between the approaches in Dowe (2008, sec. 0.2.2, footnotes 64 and 65) mentioned in section "Properties of MML (and Approximations)" largely comes down to the difference between $KL(f, g)$ and $KL(g, f)$. This said, for those interested in the finer detail of current state-of-the-art MML approximations, it seems opportune here to re-visit an issue from (Dowe, 2008, sec. 0.2.2, footnote 65). Upon reflection, (Dowe, 2008, sec. 0.2.2, footnote 64, eq (3)) should be *further* from Maximum Likelihood than the method from (Dowe, 2008, sec. 0.2.2, footnote 65). I wrap up this note by idly speculating about the merits of returning to (Dowe, 2008, sec. 0.2.2, footnotes 64 and 65) with a hybrid method involving $\alpha KL(\theta^*, \theta) + (1 - \alpha)KL(\theta, \theta^*)$ with $\alpha = 1/2$.)

And just as log-loss scoring retains the above uniqueness in its invariance under re-framing of the problem when we add (or subtract) the entropy of the prior (or a multiple thereof) (Dowe, 2008, footnote 176), again, so, too, the Kullback–Leibler divergence—or even any linear combination $\alpha KL(f, g) + (1 - \alpha)KL(g, f)$ (with $0 \leq \alpha \leq 1$)—retains its invariance when we add (or subtract) the entropy of the prior (or a multiple thereof).

Bayesianism vs. Non-Bayesianism

Much metaphorical "blood" has been spilt on the issue of whether or not prior beliefs should be incorporated into analysing data. Bayesians are those who contend that any prior knowledge should be used. While this seems fairly clear (to me), it opens some cans of worms. One issue is exactly how should we quantify our prior beliefs? Another issue is to ask what two different people with different prior beliefs—or, more

extremely, (as expert witnesses) in opposing sides of a class action, malpractice suit or other legal battle—should do to reconcile the fact that their different prior beliefs will give rise to different answers.

Some classical (non-Bayesian) statisticians have suggested quite wrongly that Bayesian methods are not statistically invariant. While it is true that some Bayesian methods are not statistically invariant, some most certainly are (Wallace and Boulton 1975).

Some Bayesian statisticians are almost self-conscious about the presence of a prior probability distribution representing prior beliefs, and try in a variety of ways to make such a term as objective as possible. Taking the log-likelihood from section "Maximum Likelihood" the Fisher information is the determinant of the matrix of the expected second partial derivatives of the log-likelihood (Wallace 2005, sec. 5.1). One of many attempts to be as objective as possible while still being Bayesian is to use the observation of Jeffreys that the Fisher information has the same mathematical form as a prior (Jeffreys 1946), and to use it as a prior. Jeffreys himself never advocated this (Wallace 2005, sec. 1.15.3), it seems rather odd that our *prior* beliefs should depend upon the observed data, and this (so-called) "Jeffreys prior" frequently either has an infinite integral or other failings (Wallace and Dowe 1999a, sec. 5, 277, col. 2, 1999b, sec. 2.3; Comley and Dowe 2005, sec. 11.4.3, 273; Wallace 2005, sec. 10.2.1; Dowe 2008, footnote 75).

It would be fair to say that the community is still a long way from being unified in the best way to analyse data. But it must be pointed out that not only can Bayesian methods be statistically invariant (Wallace and Boulton 1975; Wallace 2005) but that, furthermore, it has been conjectured (Dowe et al. 1998, 93; Edwards and Dowe 1998, sec. 5.3; Wallace and Dowe 1999a, 282, 2000, sec. 5; Comley and Dowe 2005, sec. 11.3.1, 269; Dowe, Gardner and Oppy 2007, sec. 8; Dowe 2008, sec. 0.2.5) that only Bayesian methods can give both statistical invariance and statistical consistency on the harder problems (with the amount of data per parameter bounded above) in sections "Statistical Consistency" and "Amount of Data per Parameter Bounded Above".

We now look at (probabilistic) prediction in the next section and then, in section "Some Methods of Inference: Maximum Likelihood, AIC, (Bayesian) MAP, etc." we look at some classical (non-Bayesian) and some Bayesian approaches to inference, including (in section "Bayes's Theorem and Bayesianism") a discussion of Bayes's theorem.

(Probabilistic) Prediction

The distinction between *inference* and *prediction* is that inference is concerned with finding the single best theory while prediction is concerned with finding the most probable future data—see also Wallace and Dowe (1999a, sec. 8) and Wallace (2005, sec. 10.1.2). Classical non-Bayesians seem either to conflate these two notions or to regard the single best inference as necessarily being the best predictor. In the Bayesian approach, theories have a prior probability (distribution) before the data is seen and then a posterior probability (distribution) after the data is seen.

The optimal Bayesian predictor combines all theories available, weighting them according to their respective posterior probabilities. Classical non-Bayesian inference tends to over-fit and err on the side of under-estimating any spread in the data. The single best (Bayesian) inference tends to give the best estimate of spread in the data. The best (Bayesian) predictor makes a weighted Bayesian combination of theories, as in section "Prediction". This results in a slightly conservative over-estimate of the spread in the data, due to the combination of diverse theories (Wallace 2005, sec. 4.9).

And, of course, the quality of any probabilistic predictions can be measured using the log-loss ("*probabilistic bit cost*") scoring method from section "Probabilistic Inference—vs. Mere Non-probabilistic Classification".

Some Methods of Inference: Maximum Likelihood, AIC, (Bayesian) MAP, etc.

Given data, D, how do we (best) choose which hypothesis, H, to infer? Recalling the discussion of Bayesianism from section "Bayesianism vs. Non-Bayesianism", we look at several approaches below. We consider classical (non-Bayesian) approaches in sections "Maximum Likelihood", "Akaike's Information Criterion (AIC) and Penalised (Maximum) Likelihood" and "Other: Other Classical, Other Bayesian, etc.", and we consider Bayesian approaches in sections "Bayes's Theorem and Bayesianism", "Maximum A Posteriori (MAP)", "Other: Other Classical, Other Bayesian, etc." and "Minimum Message Length (MML)".

Maximum Likelihood

Maximum Likelihood says that, given data D, we should choose the hypothesis, H, for which the likelihood $Pr(D|H)$ is maximised. Given the monotonicity of the likelihood function, Maximum Likelihood is equivalent to minimising $-\log Pr(D|H)$.

This classical approach to inference is statistically invariant—and a hand-waving argument for this is that stretching the likelihood function in and out sideways will not affect the maximum height or any height. But Maximum Likelihood tends to over-fit (especially on small sample sizes) (Wallace and Dowe 1993), "finding" non-existent patterns in random noise. One simple case in point is, where even in the case of the Gaussian distribution, the Maximum Likelihood estimator of the variance has to be corrected and multiplied by $\frac{N}{N-1}$ for sample size, N (Dowe, Gardner and Oppy 2007, sec. 6.1.1). Another simple case in point is the *bus number problem* (Dowe 2008, footnote 116, 535–536), where we arrive in a new town with θ buses numbered consecutively from 1 to θ. If we see only one bus and observe its number, x_{obs}, then Maximum Likelihood tells us to estimate θ as x_{obs}. This will typically be a silly under-estimate.

At least two or three more issues arise with Maximum Likelihood.

One issue is how do we choose between models of increasing complexity and increasingly good fit—e.g. constant, linear, quadratic, cubic, …? Maximum Likelihood advocates an unambiguous approach when all is parameterised (e.g. we know that the function is linear with Gaussian noise), but when models are nested it doesn't give a way of avoiding the most complicated model.

A second issue is that Maximum Likelihood chooses the hypothesis to make the already observed data as likely as possible. But the data has already been observed—so, philosophically, choosing the hypothesis to make the already observed data as (retrospectively) probable as possible seems to be stating the problem back to front. Shouldn't we instead find some way of choosing H so as to maximise $Pr(H|D)$? Plenty of Bayesians might consider this to be self-evident or at worst close to conclusive (Berger and Wolpert 1988; Bernardo and Smith 1994) (but classical likelihood-based reasoning and its advocates do live on (Glymour 1981; Forster and Sober 1994), as per section "Other: Other Classical, Other Bayesian, etc.").

A third issue, which is mentioned in section "Statistical Consistency", is that Maximum Likelihood is known to be statistically inconsistent for a wide range of problems where the amount of data per parameter is bounded above (Neyman and Scott 1948; Wallace and Freeman 1992; Wallace 1995; Wallace and Dowe 2000, sec. 5; Dowe, Gardner and Oppy 2007, secs 6.1 and 8; Dowe 2008, sec. 0.2.5).

Akaike's Information Criterion (AIC)—see next section—is one attempt to address the first issue. The second issue is the contentious "Bayesianism vs. non-Bayesianism" issue of section "Bayesianism vs. non-Bayesianism". If we think that maximising $Pr(H|D)$ makes more sense than maximising $Pr(D|H)$, then it makes sense to explore Bayesian approaches—such as Maximum A Posteriori (MAP) and Minimum Message Length (MML) from sections "Maximum A Posteriori (MAP)" and "Minimum Message Length (MML)" respectively, both of which use Bayes's theorem (from section "Bayes's Theorem and Bayesianism").

Akaike's Information Criterion (AIC) and Penalised (Maximum) Likelihood

Where Maximum Likelihood advocates minimising $-\log Pr(D|H)$, the Akaike Information Criterion (AIC) advocates minimising $2.(-\log Pr(D|H) + k)$, or equivalently $(-\log Pr(D|H) + k)$, where k is the number of free parameters (Akaike 1970, 1973). (It is worth mentioning that there is a substantial literature on AIC where the penalty, k, has been changed to, e.g. $\frac{3}{2}k$ and a variety of other constants multiplied by k. This will typically not change things overly much.) So, in the special case when the model class is known, the relevant variables have already been selected and we only need to do parameter estimation (e.g. we are fitting a univariate cubic polynomial with Gaussian noise, $y = (\sum_{i=0}^{3} a_i x^i) + N(0,\sigma^2)$ as per section "Problems with Increasing Numbers of Parameters" for some (a_0, a_1, a_2, a_3, σ^2) to be inferred), AIC reduces to Maximum Likelihood. The fact that AIC is the likelihood function with a penalty term (namely, k) means that AIC can be regarded as a form of *penalised likelihood*.

For the wide range of problems where the amount of data per parameter is bounded above (Neyman and Scott 1948; Wallace and Freeman 1992; Wallace 1995; Wallace and Dowe 2000, sec. 5; Dowe, Gardner and Oppy 2007, secs 6.1 and 8; Dowe 2008, sec. 0.2.5) and there is no variable selection, AIC reduces to Maximum Likelihood and suffers the same problems of statistical inconsistency. For a comparison of AIC and the

Bayesian MML approach (from section "Minimum Message Length (MML)"), see Wallace and Dowe (1999a, sec. 9) and Dowe, Gardner and Oppy (2007).

Bayes's Theorem and Bayesianism

Following discussions such as that in section "Bayesianism vs. non-Bayesianism" let us explore Bayesianism—the notion that we should look at $Pr(H|D)$ rather than at $Pr(D|H)$.

The Bayesian approach takes into account our prior beliefs over the space of possible hypotheses. We will write the prior probability of H as Pr(H). This is the probability distribution over the space of hypotheses prior to—or *before*—seeing any data. We can combine the prior, $Pr(H)$ and the (statistical) likelihood function, $Pr(D|H)$, to calculate the posterior distribution—which is the probability of hypotheses *after* seeing the data. The relationship between the prior ($Pr(H)$), the likelihood ($Pr(D|H)$) and the posterior ($Pr(H|D)$) can be shown using Bayes's theorem, which can also be thought of in terms of a Venn diagram. Repeated application of Bayes's theorem thus gives

$$Pr(H)\cdot Pr(D\,|\,H) = Pr(H\,\&\,D) = Pr(D\,\&\,H) = Pr(D)\cdot Pr(H\,|\,D). \tag{1}$$

So, this now gives the posterior probability of H given D as

$$\begin{aligned}\text{posterior}\,(H\,|\,D) &= \Pr(H\,|\,D) = \frac{\Pr(H)\cdot\Pr(D\,|\,H)}{\Pr(D)} = \frac{1}{\Pr(D)}(\Pr(H)\cdot\Pr(D\,|\,H)) \\ &= \frac{\text{prior}(H)\cdot\text{likelihood}(D\,|\,H)}{\text{marginal}(D)},\end{aligned} \tag{2}$$

where the marginal probability of D, $Pr(D)$ or marginal(D), is the prior probability that D is the data-set generated. Informally, depending upon whether H is a discrete space over which we sum or a continuous space over which we integrate, we can write $Pr(D) = \sum_H Pr(H)Pr(D\,|\,H)$ or $Pr(D) = \int_H Pr(H)Pr(D\,|\,H)\mathrm{d}H$. Discrete spaces include cases such as all attributes are categorical (e.g. drinker or non-drinker, smoker or non-smoker, male or female, etc.) and continuous spaces include attributes such as height or weight. Of course, hybrid spaces with both discrete (categorical) and continuous attributes exist, and the above formula for Pr(D) is modified to sum over the discrete attributes and integrate over the continuous attributes. The term *attribute* sometimes goes by alternative names including (e.g.) *dimension*, *feature*, *field* and *variable*. Note that all the hypotheses are used to calculate $Pr(D)$ but that they are all summed or integrated out—and, as such, $Pr(D)$ is independent of any individual hypothesis or rival hypotheses being considered for inference. So, from Equation 2, maximising $Pr(H|D)$ is equivalent to maximising $Pr(H)\cdot Pr(D|H)$.

The Bayesian interested in doing inference is quite probably going to be interested in choosing H to maximise $Pr(H|D)$—or, equivalently, to maximise $Pr(H)\cdot Pr(D|H)$. But issues arise here and, if we are not careful and principled, we might be

left open to a criticism from a classical statistician along the lines of "Classical approaches based on likelihood and penalised likelihood are invariant under re-parameterisation, but maximising the Bayesian posterior usually isn't". One issue here will be when we are dealing exclusively with discrete (categorical) attributes—and so are dealing with probabilities—and when instead at least one of our attributes is continuous and so we are dealing not with probabilities per se but with *probability densities*.

Maximum A Posteriori (MAP)

As its name suggests, the Bayesian method of Maximum A Posteriori (or MAP) maximises the posterior probability (or density), $Pr(H|D)$, or equivalently, the prior multiplied by likelihood. When all attributes are discrete (categorical), this is statistically invariant under re-parameterisation. However, when at least one of the attributes is continuous, then both $Pr(H)$ and $Pr(H|D)$ are *densities*. As an example, if the hypothesis, H, concerns a height, then $Pr(H)$ and $Pr(H|D)$ must be measured in units of 1/length, or length^{-1}, in order that the integral of the prior along the height axis gives a probability of 1. In other words, if we're multiplying something by a height (in cm) and the answer is 1, then that something must be in cm^{-1}. This gives us some insight into why MAP is generally not statistically invariant. The Bayesian prior on a length will look quite different to the prior on its square, an area—and, indeed, their maxima and minima, etc. will generally be different. The statistical likelihood ($Pr(D|H)$) is invariant but the prior ($Pr(H)$) isn't, so $Pr(H) \cdot Pr(D|H)$ and the posterior also won't be invariant in general, and therefore the maximum of the posterior—namely, the MAP estimate—also won't be invariant (Dowe, Oliver and Wallace 1996; Wallace and Dowe 1999b, secs 1.2–1.3, 1999c, sec 2, col. 1, 2000, secs 2 and 6.1; Comley and Dowe 2005, sec. 11.3.1; Dowe 2008, sec. 0.2.3). The similarity between MAP and Maximum Likelihood means that MAP inherits the statistical inconsistency results of Maximum Likelihood described in section "Maximum Likelihood" for problems where the amount of data per parameter is bounded above. Even when all attributes are discrete (and so issues of density do not arise), even then MAP can inherit the statistical inconsistency tendencies of Maximum Likelihood for problems where the amount of data per parameter is bounded above (Dowe 2008, footnote 158).

The good news is that if we re-visit MAP very carefully and make sure that our posterior is a *probability* and not a *density*, then we arrive at something which is statistically invariant. If we take some more care (when required), then we also get statistical consistency for the hard problems where the amount of data per parameter is bounded above. This approach is Minimum Message Length (MML) (Wallace and Boulton 1968, 1975; Wallace 2005), which we will discuss in section "Minimum Message Length (MML)". But, first, we all too briefly gloss over some of the many other approaches to inference in the next section and then—in the remainder of section "Some Methods of Inference: Maximum Likelihood, AIC, (Bayesian) MAP, etc."—touch on other issues pertaining to inference.

Other: Other Classical, Other Bayesian, etc.

In this section, we attempt to mention some of the myriad of alternative estimation techniques used in the literature and not yet discussed above. One can only do one's best with such an impossible task, but it is worth re-emphasising the point from section "Bayesianism vs. Non-Bayesianism" that the community remains far from unified in how best to do inference. The classical (non-Bayesian) community is far from unified. And, whether or not MML is "*the*" way to do inference, the Bayesian community currently remains a long way from unified.

Schwarz's (1978) Bayesian Information Criterion (BIC) is independent from and coincidentally equivalent to the 1978 version of Minimum Description Length (MDL) (Rissanen 1978) (which, in turn, shares much in common with MML, as per Wallace and Dowe (1999a, secs 6.2 and 7, 1999b), Wallace (2005, sec. 10.2), Comley and Dowe (2005, sec. 11.4.3) and Dowe (2008, sec. 0.2.2), although MML pre-dates MDL by a decade (Wallace and Dowe 1999a, sec. 1, 271, col. 1; Comley and Dowe 2005, sec. 11.1; Dowe 2008, secs 0.2.2–0.2.4)). Recall from section "Akaike's Information Criterion (AIC) and Penalised (Maximum) Likelihood" that AIC was a penalised likelihood of the form $-\log Pr(D|H) + k$ where k is the number of free parameters. BIC advocates minimising $-\log Pr(D|H) + \frac{k}{2} \log N$, where N is the sample size of the data. For sufficiently large N (indeed, once $N \geq 8 > e^2$ and $\log N > 2$), we see that the BIC penalty of $\frac{k}{2} \log N$ becomes greater than the AIC penalty of k. So, for larger sample sizes, BIC tends to give a larger (and, we contend, more appropriate) penalty than AIC.

The Vapnik–Chervonenkis dimension, Structural Risk Minimisation (SRM) and Support Vector Machine (SVM) approach (Vapnik 1995) is a (classical or) non-Bayesian approach which came from the machine learning community and is only slowly working its way through statistics and econometrics. That said, there have been efforts to do this in a Bayesian way and also in a (Bayesian) MML way (Vapnik 1995, sec. 4.6; Tan and Dowe 2004; Dowe 2007, 2008, sec. 0.2.2), including explicitly modelling (Dowe 2008, footnote 53, fourth way, 527–528) the distribution of *all* the variables, including the input variables.

The minimum expected Kullback–Leibler distance (MEKLD, or MEKL, or minEKL) estimator (Dowe et al. 1998; Wallace 2005, secs 4.7–4.9; Dowe, Gardner and Oppy 2007, sec. 6.1.4) is a Bayesian estimator which uses the notion of Kullback–Leibler distance (from section "Kullback–Leibler Divergence (or Kullback–Leibler Distance)") to attempt to optimise the (average) log-loss probabilistic score from section "Probabilistic Inference—vs. Mere Non-probabilistic Classification" on future, as yet unseen, data. It does this by taking the Bayesian posterior distribution $Pr(H|D)$ over hypotheses H to then get a distribution $f(y|D)$ on "expected" future data, y. Having such a probability distribution on the "expected" future data, it then seeks a hypothesis H which - in average expectation - optimises the "expected" log-loss penalty. The purpose of minEKL is to be the best possible (Bayesian) predictor within the parameter space. It is perhaps curious that this was the original motivation behind AIC (Akaike 1970, 1973), although Akaike tried to do this without use of a Bayesian prior. A comparison between MEKLD and AIC is given in Dowe, Gardner and Oppy (2007, sec. 6.1.4).

We now say something about hypothesis testing (as a form of inference), experimental design and prediction in the next three sections respectively, and then talk about Bayesian Minimum Message Length (MML) inference in section "Minimum Message Length (MML)".

Hypothesis Testing

Recall from section "Maximum Likelihood" that Maximum Likelihood tries to choose a hypothesis, H, to make the already observed data, D, as retrospectively likely as possible. Classical hypothesis tests do the same curious thing, trying to say how probable the observed data would be if the actual hypothesis were true—rather than how probable the hypothesis is given the data. As such, classical hypothesis tests—like maximum likelihood—often neglect how complicated or even tightly-peaked the hypothesis is (Dowe 2008, sec. 1 and footnotes 57 and 58). In fairness, the classical hypothesis test tries to objectively side-step any use of Bayesian priors, although they often (inadvertently?) include a prior which can be slightly curious (Dowe 2008, sec. 0.2.5).

Experimental Design, Data Collection Protocol and Likelihood Principle

This (brief) section is partly to mention that any experiment should be designed "randomly" to collect as much information as possible (Dowe 2008, sec. 0.2.7, 544). Whatever the data collection protocol, the statistical *likelihood principle* says, roughly, that the likelihood function $Pr(D|H)$ is all that we need to know about the data (Berger and Wolpert 1988; Grossman forthcoming). As such, Maximum Likelihood will always honour the likelihood principle. In changing from a Binomial protocol (when we sample a fixed number of times) to a Negative Binomial protocol (when we sample until a fixed number of successes), MML (from section "Minimum Message Length (MML)") gives at worst a minor violation of the likelihood principle (Wallace 2005, sec. 5.8) (although I am not convinced that this constitutes a valid criticism of MML). But, as discussed in Wallace and Dowe (1999b, sec. 2.3.5) and Wallace (2005, sec. 10.2.2), some inference methods—even those doing all they can to avoid using a Bayesian prior—can be in substantial violation of the likelihood principle.

Having said above something one could interpret as meaning that we wish to design our experiment to have the maximum expected information gain, over the next 16 lines or so I'd now like to change tack and put forward something of a paradox—or the making thereof. Consider experiments (or tests) T_1, T_2, …, T_s, …, T_i, … such that, for some $n > 0$, experiment T_i has probability $2^{-2^{-(n+i)}}$ of yielding 0 information. The experiments can be independent of one another, and perhaps are such that with probability $1-2^{-2^{-(n+i)}}$, experiment T_i yields $2^i/(1-2^{-2^{-(n+i)}})$ bits, thus yielding an expected information gain from experiment T_i of $(1-2^{-2^{-(n+i)}}) \times (2^i/(1-2^{-2^{-(n+i)}})) = 2^i$ bits. Letting $s \geq 1$, the probability that, starting

with experiment T_s, all the experiments T_s, T_{s+1}, … yield 0 information is $2^{-2^{-(n+s)}} \times 2^{-2^{-(n+s+1)}} \times \ldots = 2^{-(2^{-(n+s)}+2^{-(n+s+1)}+\ldots)} = 2^{-(2^{-(n+s-1)})}$. On the other hand, the expected information gain from starting with experiment T_s is $2^s + 2^{s+1} + \ldots = \infty$. Paradoxically, for a finite number (say j) of experiments, T_s, …, T_{s+j-1}, the larger the value of s that we start with, the larger our expected information gain but the greater the probability that all the experiments from T_s to T_{s+j-1} and forever will all yield 0 information. With $n > 0$, this probability $2^{-(2^{-(n+s-1)})}$ of getting no information rapidly approaches 1 for increasing s. We can extend the paradox by averaging the information gains of T_s, T_{s+1}, …, T_u and then letting u tend to infinity. As u gets larger, the expected average information gain (divided by (u-s+1)) tends to infinity (on the one hand) but—(on the other hand) curiously—the probability that the average information gain (divided by (u–s+1)) is arbitrarily close to 0 becomes arbitrarily close to 1.

And, last, while not about design *per se* but rather about protocol, an issue about the reporting and collection of results (rather than directly about the collection of data) is the unfortunate trend of not reporting negative results (Dowe 2008, sec. 0.2.5) and of only reporting positive results. There seems to be some sort of widespread—but fortunately not universal—implicit result-reporting protocol in some communities whereby negative results only get to be published when following on in response to a reported positive result. This can easily become data censoring of a primitive kind and can give rise to all sorts of bias. (Although it is not in a medical area, this view is presumably shared by the recently formed *Journal of Interesting Negative Results in Natural Language Processing*.)

Prediction

Recall the distinction(s) between inference and prediction in section "(Probabilistic) Prediction" (Wallace and Dowe 1999a, sec. 8; Wallace 2005, sec. 10.1.2). As in the AIC approach of Akaike (1970, 1973), the classical approach to prediction seems to be to find the single best theory—and use that for both inference and prediction. One could argue that it makes more sense intuitively—even in the classical (non-Bayesian) approach, such as Akaike's—to combine several theories which perform similarly (Wallace and Dowe 1999c, sec. 4). Empirical results would certainly suggest (Dowe, Gardner and Oppy 2007) that this is true in the case of AIC.

The Bayesian approach to prediction consists of taking every available theory in the parameter space and weighting it according to its posterior probability, and then using this to get a predictive distribution over expected future data. The resultant distribution will not necessarily be in the original parameter space—e.g. if our distribution is known to be $N(0, \sigma^2)$ and we consider all such possible distributions weighted by the posterior density of σ^2 (or, equivalently, of σ), the result will be an infinite mixture of Normal distributions.

Note that where the predictive distribution is not in the parameter space, the best fit to the predictive distribution from within the parameter space turns out to be the

MEKLD estimator (Dowe et al. 1998; Wallace 2005, secs 4.7–4.9; Dowe, Gardner and Oppy 2007, sec. 6.1.4) from section "Other: Other Classical, Other Bayesian, etc.". This makes sense, because MEKLD gives the best expected log-loss score (amongst hypotheses within the parameter space) and also because (recalling section "Probabilistic Inference—vs. Mere Non-probabilistic Classification") log-loss scoring rewards the true probabilities (if known) and appears to be unique in doing so (Dowe 2008, sec. 0.2.5).

Where there is one outstandingly good theory, then prediction and inference come very much to the same thing (Dowe 2008, sec. 0.3.1). However, they can vary when there is no outstandingly good theory, whereupon it is a good idea for predictive purposes to make a weighted combination of good inferences (Wallace and Dowe 1999a, sec. 8, 1999c, sec. 4), ideally—where feasible—weighting over the entire posterior.

Minimum Message Length (MML)

From Equation 2 and section "Bayes's Theorem and Bayesianism", we have that choosing H to maximise $Pr(H|D)$ is equivalent to choosing H to maximise $Pr(H) \cdot Pr(D|H)$. By the monotonicity of the logarithm function, this is equivalent to minimising $-\log(Pr(H) \cdot Pr(D|H)) = -\log Pr(H) - \log Pr(D|H)$. Simply changing notation, we can equivalently write

$$\begin{aligned} \arg\max_H Pr(H \mid D) &= \arg\max_H Pr(H) \cdot Pr(D \mid H) \\ &= \arg\min_H -\log Pr(H) - \log Pr(D \mid H). \end{aligned} \tag{3}$$

All data-sets—or at least all the ones I've used and/or heard of—are finite. This is partly so because, as mentioned at the start of section "Desiderata in (Probabilistic) Inference and (Probabilistic) Prediction", all heights and weights, etc. are measured to finite accuracy and finitely many decimal places (Wallace and Dowe 1993, 1–3, 1994, 38, secs 2 and 2.1, 2000, sec. 2, 74, col. 2; Dowe, Allison et al. 1996, sec. 2; Kissane, Bloch, Dowe et al. 1996, 651; Comley and Dowe 2003, sec. 9; Fitzgibbon, Dowe and Vahid 2004, eqn (19); Comley and Dowe 2005, sec. 11.3.3, 270; Wallace 2005, secs 3.1.1 and 3.3; Dowe, Gardner and Oppy 2007; Dowe 2008, sec. 0.2.4).

Given that heights, weights and other measurements are measured and recorded to finite accuracy, a fact often neglected by statisticians is that the likelihood, $Pr(D|H)$, can be viewed as a *probability* rather than as a *density* (of zero point mass). Let us clarify with an example. The probability of measuring a height of (say) 1.84 m is the probability that the height is between (say) 1.835 and 1.845 m. So, $Pr(D|H)$ is a probability (as measured), even if (in some sort of theory) perhaps a density. In the case that our parameter space is a continuum—such as the mean and variance, μ and σ^2, of heights—if we suitably quantise this down into at most countably many permissible (or usable) estimates, then $Pr(H)$ will also correspond to a probability and not a density (and we can even argue that the standard deviation, σ should be bounded below by a multiple of the measurement accuracy (Wallace and Dowe 1994, sec. 2.1; Dowe, Allison et al. 1996, sec. 2; Kissane, Bloch, Dowe et al. 1996, 651;

Comley and Dowe 2003, sec. 9; Comley and Dowe 2005, sec. 11.3.3; Dowe 2008, sec. 0.2.4)). As alluded to at the end of section "Maximum A Posteriori (MAP)", this will give us $Pr(H)$, $Pr(D|H)$ and (consequently) $Pr(H|D)$ all as probabilities—and *not* as densities.

Information Theory, Compression and MML

It should be clear to the reader from section "Maximum A Posteriori (MAP)" and/or from Dowe (2008, footnote 158 and sec. 0.2.3) that MML is, in general, different from MAP.

Let us now mention a result from information theory Shannon (1948) that an event e_i of probability p_i can be encoded in a prefix code by a code-word of length l_i where $-\log p_i \approx l_i < (-\log p_i) + 1$. This can be achieved with a Huffman code, which successively joins the two least probable events together and iterates. For details of code construction, see Wallace (2005, chap. 2, especially sec. 2.1), including the example in Wallace (2005, sec. 2.1.4 and fig. 2.5). As a simple example, if we have $\{p_1, p_2, p_3, p_4, p_5\} = \{1/2, 1/4, 1/16, 1/16, 1/8\}$ and we are considering binary codes, then a Huffman code would first join e_3 and e_4 (call this $e_{3,4}$) to give a probability of $1/16 + 1/16 = 1/8$, then next join $e_{3,4}$ and e_5 (call this $e_{3,4,5}$) to give a probability of $1/8 + 1/8 = 1/4$, then join e_2 to $e_{3,4,5}$ (call this $e_{2,3,4,5}$) to give a probability of $1/4 + 1/4 = 1/2$, and then finally join e_1 and $e_{2,3,4,5}$, resulting in a probability of $1/2 + 1/2 = 1$, whereupon it would stop. With "up" branches being given the bit 0 and "down" branches being given the bit 1, this would result in a binary code tree as in Figure 1 where e_1, e_2, e_3, e_4 and e_5 respectively have the code-words 0, 10, 1100, 1101 and 111. In this friendly example, we note that $l_i = -\log_2 p_i$ in each case.

Given that an event of probability p_i can be represented by a code-word of length l_i, looking at Equation 3, we see that $-\log Pr(H)$ can be regarded as the length of a message for encoding a hypothesis and $-\log Pr(D|H)$ can be regarded as the length of a message for encoding the data given this hypothesis. So, maximising the posterior *probability*, $Pr(H|D)$, is equivalent to minimising the length of a two-part

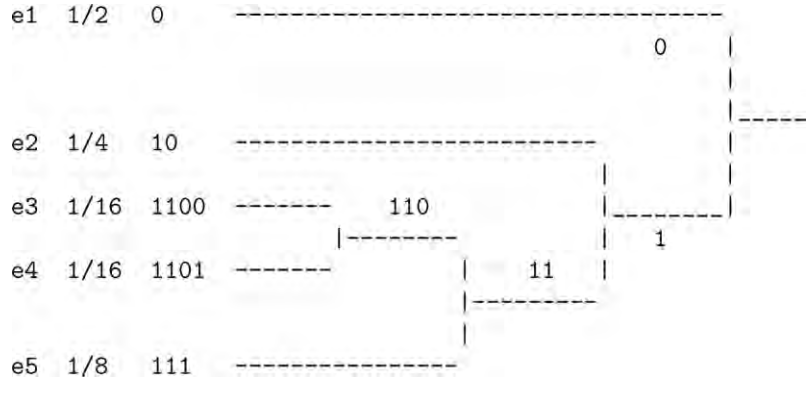


Figure 1 A simple Huffman code tree.

message, $-\log Pr(H)$ $-\log Pr(D|H)$, for jointly encoding the hypothesis and the observed data given this hypothesis. Hence the name minimum message length (MML). Given that MML is maximising a *probability* and *not* a density, and given the benefits of this (as per section "Properties of MML (and Approximations)"), MML can be thought of as MAP done properly (Wallace and Dowe 1999b, secs 1.2–1.3, 1999c, sec. 2, col. 1, 2000, secs 2 and 6.1; Comley and Dowe 2005, sec. 11.3.1; Dowe, Gardner and Oppy 2007, sec. 5.1, coding prior; Dowe 2008, footnote 158).

Philosophers wanting to know more about MML might wish to read Wallace (2005), Dowe, Gardner and Oppy (2007) and Dowe and Oppy (2001). Some of the many other articles of interest include Wallace and Boulton (1968), Comley and Dowe (2003, 2005), Wallace and Dowe (1999a) and Dowe (2008).

Ockham's Razor and MML

We recall the idea from Ockham's razor—or a common interpretation thereof—that if two theories fit the data equally well then one should prefer the simpler. Given MML's desire to quantitatively find (relatively) simple theories that fit the data (relatively) well, one can regard MML as being not only a quantitative version of Ockham's razor, but perhaps also a generalisation. Where Ockham's razor only seems to tell us which theory to prefer when both fit the data equally well, MML gives us a quantitative trade-off between simplicity and goodness of fit.

Figure 2 gives an example of two rival hypotheses, H_1 and H_2, for the data. We see that the encoding of Data given H_2 is shorter than that of Data given H_1, meaning that H_2 fits the data with a better log-likelihood than did H_1. However, we also see that the code length for H_1 is far shorter than that of H_2, meaning that H_1 is far more probable a priori (or simpler) than H_2. In this example, the shorter two-part message length is the explanation involving H_1, and so it would be the preferred MML inference. For further comments on MML and Ockham's razor, see, e.g. Needham and Dowe (2001), Comley and Dowe (2005, sec. 11.4.3) and Dowe (2008, footnotes 18 and 182).

Turing Machines, Algorithmic Information Theory and MML

A *Turing machine* (TM) (Wallace 2005, sec. 2.2.1) is an abstract mathematical model of a computer program. It can be written in a language from a certain alphabet of symbols (such as 1 and (blank) " "). We assume that Turing machines have a read/write

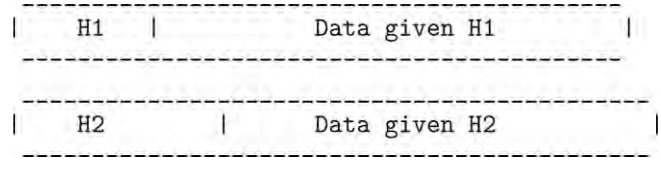


Figure 2 Two-part message lengths for two rival hypotheses for some Data.

head on an infinitely long tape. A Turing machine in a given state (with the read/write head) reading a certain symbol either moves to the left (L) or to the right (R) or stays where it is and writes a specified symbol. The instruction set for a Turing machine can be written as:

$$f: States \times Symbols \rightarrow States \times (\{L, R\} \cup Symbols)$$

Without loss of generality we can assume that the alphabet is the binary alphabet {0,1}, whereupon the instruction set for a Turing machine can be written as:

$$f: States \times Symbols \rightarrow States \times (\{L, R\} \cup Symbols).$$

Any known computer program can be represented by a Turing Machine. *Universal* Turing Machines (UTMs) are like compilers and can be made to emulate *any* Turing Machine (TM).

An example of a Turing machine would be a program which, for some a_0 and a_1, when given any input x, calculates (or outputs) $a_0 + a_1 x$. In this case, x would input in binary (base 2), and the output would be the binary representation of $a_0 + a_1 x$.

A *Universal Turing machine* (UTM) (Wallace, 2005, sec. 2.2.5) is a Turing machine which can simulate any other Turing machine. So, if U is a UTM and M is a TM, then there is some input c_M such that for any string s, $U(c_M s) = M(s)$ and the output from U when given the input $c_M s$ is identical to the output from M when given input s. In any other words, given any TM M, there is an emulation program (or code) c_M so that once U is input c_M it forever after behaves as though it were M.

The notion of *algorithmic information theory* (or *Kolmogorov complexity*) (Solomonoff 1964; Kolmogorov 1965; Chaitin 1966) of a string x is the length of the shortest input l_x to a UTM U such that $U(l_x) = x$—i.e. U will output x if given input l_x. Informally, if the length of l_x is the same as (or larger than) the length of x, then we can say that x is random in some sense. Similarly, if the length of l_x is much less than the length of x, then we can say that x is non-random.

Of these works from the mid-1960s, Kolmogorov (1965) and Chaitin (1966) study this important new concept while Solomonoff (1964) is also interested in using it for prediction. This work from the mid-1960s was very shortly before the first appearance of MML (Wallace and Boulton 1968). And just as MML can be regarded as a quantitative version of Ockham's razor (or as a generalisation thereof) as per section "Ockham's Razor and MML" and Figure 2, MML can also be regarded as (two-part) Kolmogorov complexity (Wallace and Dowe 1999a; Comley and Dowe 2005; sec. 11.4.3; Wallace 2005, chap. 2; Dowe 2008, secs 0.2.2, 0.2.7 and 0.3.1). Here, the first ("hypothesis") part of the message tells the Turing machine what hypothesis program is to be emulated but no output is written yet. The bit string in the second ("data given hypothesis") part of the message then causes the emulation program to output the original data string.

Note the analogy between MML as information theory in section "Information Theory, Compression and MML" and MML as *algorithmic* information theory in this section. And, very relatedly, note that the Kolmogorov complexity is dependent on the choice of UTM, and that this is a *Bayesian* choice (Wallace and Dowe 1999a, secs 2.4 and 7; Comley and Dowe 2005, sec 11.3.2; Dowe 2008, footnote 133).

Properties of MML (and Approximations)

The approach obtained from strictly minimising the message length as above is called Strict Minimum Message Length—or Strict MML, or SMML (Wallace and Boulton 1975; Wallace and Freeman 1987; Wallace and Dowe 1999a, sec. 6.1; Wallace 2005, chap. 3; Dowe, Gardner and Oppy 2007, sec. 5; Dowe 2008, footnotes 12, 153, 158 and 196 and sec. 0.2.2). Despite the many desirable properties of SMML (Wallace 2005, sec. 3.4), it can be computationally intractable even for relatively simple problems (Wallace 2005, sec. 3.2.9). (Historically, Strict MML (Wallace and Boulton 1975) came 7 years after MML (Wallace and Boulton 1968).) In practice, we often use a variety of applications, and these are also known to share many of the desirable properties of SMML.

Given data, D, the MMLD (or I_{1D}) approximation (Wallace 2005, secs 4.10 and 4.12.2; Dowe 2008, sec. 0.2.2) seeks a region R which minimises

$$-\log\left(\int_R h(\vec{\theta})\mathrm{d}\theta\right) \quad -\frac{\int_R h(\vec{\theta})\cdot\log f(\vec{D}\,|\,\vec{\theta})\mathrm{d}\theta}{\int_R h(\vec{\theta})\mathrm{d}\theta}. \tag{4}$$

The length of the first part is the negative log of the probability mass inside the region, R. The length of the second part is the (prior-weighted) average over the region R of the log-likelihood of the data, D.

An earlier approximation similar in motivation which actually inspired MMLD is the Wallace-Freeman approximation (Wallace and Dowe 1999a, sec. 6.1.2; Wallace 2005, chap. 5),

$$-log\left(h(\vec{\theta})\cdot\frac{1}{\sqrt{\kappa_d^d\ \mathrm{Fisher}\,(\vec{\theta})}}\right) \quad -log\, f(\vec{x}\,|\,\vec{\theta})+\frac{d}{2}, \tag{5}$$

which was first published in the statistics literature (Wallace and Freeman 1987).

The term $1/\sqrt{\kappa_d^d\ \mathrm{Fisher}\,(\vec{\theta})}$ gives a measure of uncertainty or quantisation in hypothesis space, where d is the number of continuous-valued parameters, κ_d is a constant (Fitzgibbon, Dowe and Vahid 2004, 441; Wallace 2005, table 3.4) between $1/12$ and $1/(2\pi e)$ and the Fisher information, $\mathrm{Fisher}\,(\vec{\theta})$, is as described in section "Bayesianism vs. Non-Bayesianism". (More specifically, in maths-speak, κ_d corresponds to the geometry of the optimally tesselating—or tiling—Voronoi region in d dimensions. In plainspeak, circles are compact but don't tile because they leave gaps, squares tile the plane, but hexagons tile optimally. $\kappa_2 = 5/(36\sqrt{3})$ corresponds to the geometry of a hexagon.) The term $d/2$ is the round-off in the second part of the message due to the uncertainty in the parameter estimate.

Perhaps the first thing to mention about Strict MML is its generality (Wallace 2005, sec. 3.4.3), that it is always defined—as likewise is Kolmogorov complexity. Strict MML, Wallace–Freeman and MMLD are all statistically invariant (Wallace 2005), as also are the estimators from Dowe (2008, sec. 0.2.2, footnotes 64 and 65) alluded to near the end of the section "Probabilistic Inference—vs. Mere Non-probabilistic Classification". Various theoretical results exist about the statistical consistency and efficiency of Strict MML (Wallace and Freeman 1987, 241; Barron and Cover 1991; Wallace 1996, 2005, sec. 3.4.5; Dowe, Gardner and Oppy 2007, sec. 5.3.4), and specific examples demonstrate the statistical consistency of Wallace–Freeman (Dowe and Wallace 1997) and similar approximations (Wallace and Freeman 1987; Wallace 1995). Many papers (e.g. Wallace and Freeman 1992; Wallace and Dowe 1993, 1999a, sec. 9, 1999b; Wallace 1995; Dowe, Oliver and Wallace 1996; Fitzgibbon, Dowe and Vahid 2004; Tan and Dowe 2002, 2003, 2004; Dowe, Gardner and Oppy 2007) attest to excellent small-sample performances of the Wallace–Freeman (or similar) approximation.

Another, possibly prophetic, thing to mention is that Strict MML first appeared in 1975 (Wallace and Boulton 1975) and the approximation from Equation 5 with the lattice constants (κ_d) first appeared in 1987 (Wallace and Freeman 1987), where κ_2 corresponds to the hexagon. When the trinomial (or 3-state multinomial) distribution—which has $d = 2$, as the parameters are p_1 and p_2 (because $p_3 = 1 - p_1 - p_2$)—was first done using Strict MML well over a decade later, with a uniform prior and $N = 60$ data-points, the partition (of the triangle) for the trinomial distribution turned out to contain an absolute abundance of hexagons (Wallace 2005, fig. 3.1, 166).

Problems with Increasing Numbers of Parameters

Consider the univariate polynomial regression problem of Dowe, Gardner and Oppy (2007, sec. 6.2) and Dowe (2008, sec. 0.2.3). Given data $(x, y)_{j=1,\ldots,N}$, we seek d, a_0, …, a_d, σ^2 such that $y = (\sum_{i=0}^{d} a_i x^i) + N(0,\sigma^2)$. This is a problem of nested models (or sub-families) (Dowe, Gardner and Oppy, 2007, sec. 7.1), in that (e.g.) every quadratic is also a cubic.

Studies (Wallace 1997; Dowe 2008, ref. 281; Dowe, Gardner and Oppy 2007, sec. 6.2.1; Dowe 2008, ref. 281) show that the classical Maximum Likelihood and AIC methods from sections "Maximum Likelihood" and "Akaike's Information Criterion (AIC) and Penalised (Maximum) Likelihood" over-fit, over-estimating the model order and (as in the section "Maximum Likelihood") under-estimating the variance, σ^2. MML gets the model order correct more often, sometimes under-estimating it (Dowe, 2008, footnote 153) and certainly getting a smaller squared error in prediction.

A different problem with nested models is that of econometric autoregressive time series. Models with terms from only the recent past are a special case of models including all of these terms and terms from the more distant past. Studies (Fitzgibbon, Dowe and Vahid 2004; Dowe, Gardner and Oppy 2007, sec. 6.2.2) similarly show the classical Maximum Likelihood and AIC methods over-fitting, and MML managing to give better predictions using a lower model order.

Amount of Data per Parameter Bounded Above

In the classic Neyman–Scott problem (Neyman and Scott 1948; Dowe and Wallace 1997; Wallace 2005, secs 4.2–4.5; Dowe, Gardner and Oppy 2007, sec. 6.1; Dowe 2008, secs 0.2.5 and 0.2.3), we measure N people's heights J times each (say $J = 2$) and then infer

1. the heights $\mu_1, \ldots, \mu_N$ of each of the N people,
2. the accuracy (σ) of the measuring instrument.

We have JN measurements from which we need to estimate $N + 1$ parameters. $JN/(N + 1) \leq J$, so the amount of data per parameter is bounded above (by J), the notion of which we flagged in section "Statistical Consistency"

$$\hat{\sigma}^2_{\text{Maximum Likelihood}} \rightarrow \frac{J-1}{J}\sigma^2,$$

and so for fixed J as $N \rightarrow \infty$ we have that Maximum Likelihood is statistically inconsistent—under-estimating σ and "finding" patterns that aren't there. As alluded to in the section "Properties of MML (and Approximations)", MML remains statistically consistent for the Neyman–Scott problem (Dowe and Wallace 1997).

What makes the Neyman–Scott problem difficult is that, even though the amount of data is increasing unboundedly, the amount of data *per parameter* is bounded above. This is sufficient to preserve the small sample bias from section "Maximum Likelihood". This is somewhat awful for Maximum Likelihood and Akaike's Information Criterion (AIC).

Other examples of the amount of data being bounded above include

- latent factor analysis—single (Wallace and Freeman 1992; Edwards and Dowe 1998) and multiple (Wallace 1995, 2005, sec. 6.9; Dowe, Gardner and Oppy 2007, sec. 6.1.3; Dowe 2008, sec. 0.2.3), and
- fully-parameterised mixture modelling (Wallace and Dowe 2000, sec. 4.3; Wallace 2005, sec. 6.8; Dowe, Gardner and Oppy 2007, sec. 6.1.3; Dowe 2008, sec. 0.2.5).

These problems are more commonplace than one might at first realise. The factors from latent factor analysis correspond to notions like I.Q. or octane rating. More specifically, if we get N people to sit J aptitude tests or if we test N petrols on J engines, then what we wish to infer are statistical factors—such as I.Q. and octane rating. These I.Q.s (for each of the N people in turn) and the octane ratings (for each of the N petrols in turn) are known as the factor scores. But we also need to estimate the factor loads, or the load vector. This basically tells us how important, relevant or otherwise—and, if relevant, how difficult/easy—each aptitude test or engine test is. Both Maximum Likelihood and AIC again struggle in such cases, with Akaike (1987) himself adopting a Bayesian prior—actually, a "prior" which changes as the sample size changes (Akaike 1987, sec. 5, 325; Dowe, Gardner and Oppy 2007, sec. 6.1.3 and footnote 22)—for latent factor analysis. Empirical studies (Wallace and Freeman 1992; Wallace 1995) again show MML outperforming these methods—even when they have been helped

out with Bayesian priors—and doing so with simpler models. For these types of problems (with data per parameter bounded above), classical methods often appeal to Bayesianism for help (Wallace 2005, sec. 4.5).

By acknowledging *uncertainty* (or quantising) when doing parameter estimation, MML is statistically consistent on all of these problems. MML is about *inference*, seeking the *truth* (Dowe 2008, secs 0.2.4 and 0.2.6). (Indeed, Steven L. Gardner would like to relate MML to the notion in philosophy of *approximate truth*.) It gives a statistically invariant—and statistically consistent—Bayesian method of point estimation. It gives general consistency results where classical non-Bayesian approaches are known to break down. It is also efficient, working well on all real inference problems currently known to the author.

The above evidence and experience has led to the following two conjectures.

Conjecture 1 (Dowe et al. 1998, 93; Edwards and Dowe 1998, sec. 5.3; Wallace and Dowe 1999a, 282, 2000, sec. 5; Comley and Dowe 2005, sec. 11.3.1, 269) Only MML and very closely-related Bayesian methods are in general both statistically consistent and invariant.

Conjecture 2 (*Back-up Conjecture*) (Dowe, Gardner and Oppy 2007, sec. 8; Dowe 2008, sec. 0.2.5) If there are any such non-Bayesian methods, they will be far less efficient than MML.

Before proceeding to a final discussion and conclusion, it seems appropriate to first mention some medical and humanities applications of MML.

Medical, Biological and Other Applications of MML

Some Medical-related Applications of MML

The second application ever of MML to real-world data was in a classification of depression (Pilowsky, Levine and Boulton 1969). Studies classifying and clustering grieving families include (Kissane, Bloch, Dowe et al. 1996; Kissane, Bloch, Onghena et al. 1996), with a classification of sub-groups within autism given in Prior et al. (1998) and a classification of distress syndromes in Clarke et al. (2003). Another study of a diagnostic nature was McKenzie et al. (1993).

A fairly routine application of some MML clustering software (Edgoose, Allison and Dowe 1998, sec. 6; Dowe, Allison et al. 1996, sec. 5, 253; Wallace 1998, sec. 4.2; Dowe 2008, footnote 85) gave that proteins apparently fold with the Helices (and Extendeds) forming first and then the "Other" turn classes forming subsequently to accommodate these structures. Some further applications of MML clustering are cited in Wallace and Dowe (1994) (and Dowe 2008).

DNA microarray data (Tan, Dowe and Dix 2007) can be studied by MML (Dowe 2008, sec. 0.2.7, footnote 196), and MML image analysis (Wallace 1998; Visser and Dowe 2007) is also ripe for medical applications.

And, although it doesn't just apply to medical data, if we take a noise-free unstructured, unnormalised database of sufficient size and then apply MML Bayesian nets

(Comley and Dowe 2003, 2005; Dowe 2008, sec. 0.2.5) (from section "Kullback–Leibler Divergence (or Kullback–Leibler Distance)") to this, we get the elegant result that the MML Bayesian net inference will result in a normalised database (Dowe 2008, sec. 0.2.6, footnote 187). If there is sufficient data, this will be a full normalisation.

Some Applications of MML in the Humanities

Many applications of MML to real-world data-sets and a variety of subject areas exist—see, e.g. Wallace (2005), Dowe (2008) and elsewhere. For the curious reader, I'd like to give some admittedly all too brief pointers to reading on MML in philosophy and humanities. These include

- MML and an argument that—contrary to widely-held views in physics, philosophy and many fields—entropy is *not* time's arrow (Wallace 2005, chap. 8; Dowe 2008, sec. 0.2.5),
- MML, existence of "miracles" (Dowe 2008, sec. 0.2.7), cosmological arguments and "Intelligent Design" (I.D.),
- MML and linguistics—inferring "dead" languages (Ooi and Dowe 2005; Dowe 2008, sec. 0.2.4),
- MML, Kolmogorov complexity (Wallace and Dowe 1999a), measures of "intelligence" (Dowe and Hajek 1998; Hernandez-Orallo 2000; Legg and Hutter 2007; Dowe 2008, sec. 0.2.5) and a possible variation on the (so-called) Lucas–Penrose argument in the philosophy of mind that humans are (supposedly) more intelligent than machines can be (Dowe 2008, footnotes 70–71 and sec. 0.2.3),
- MML and the Efficient Markets Hypothesis, in which appeals to the relationship between MML and Kolmogorov complexity (as per section "Turing Machines, Algorithmic Information Theory and MML") tell us that markets are *not* provably efficient (Dowe and Korb 1996; Wallace 2005, sec. 9.1, 387; Dowe 2008, sec. 0.2.5), and
- varying the *elusive model paradox* (Dowe 2008, footnote 211) so that each bit (0 or 1) in a sequence of bits is to be the bit which was *not* predicted to be the (most probable) next bit in the sequence. (Recalling sections "(Probabilistic) Prediction" and "Prediction", we can do this by inferring a model from the past bits—as per the original elusive model paradox—or by combining several models and predicting.) For mathematicians and computer scientists, this gives us a new non-computable number. In terms of game theory, psychology, sociology and making of "thriller" movies with lots of plot "twists and turns", this relates to leaving a trail so that those trying to follow you will (almost) always—or at least as often as possible—be surprised by your next step.

Discussion and Conclusion

When seeking to draw conclusions from medical (or other) data evidence, sometimes the problem is particularly friendly—there are few parameters to be estimated, there is

an abundance of data and there is relatively little noise in the data. Inference here should be sound, the best predictor should be close to the derived inference, and the reported power of hypothesis tests should not be unreasonable.

But in the not uncommon case that the amount of data per parameter is limited, great care should be taken. Not only do the classical approaches to inference start to differ, but the ones in common usage at the time of writing tend to over-estimate the relevance of the explanatory variables and under-estimate the noise. Answers obtained by classical methods will typically improve when replaced by the Bayesian MML approach, and this seems to hold regardless of sample size. Here, care should be taken in the choice of a Bayesian prior, lest one be accused of fudging one's results.

The seeming objectivity of the classical approach versus the seeming more reliable results from the Bayesian MML approach leaves us with something of a quandary. As a *first recommendation*, at the very least, the data analyst should be aware of these issues and should ideally at least mention something along these general lines when publishing. As a *second recommendation*, if one is using a classical approach, then—where possible—the data should also be analysed in a Bayesian (MML) way alongside whatever classical analysis is chosen. If one is understandably concerned about (e.g.) how one's peers might regard a Bayesian analysis, one can repeat the study with a different set of Bayesian priors. One can then report discrepancies and—one hopes—similarities amongst different approaches. As a *third recommendation*, probabilistic predictions should be made and—given the apparent uniqueness result(s) (from the section "Probabilistic Inference—vs. Mere Non-probabilistic Classification" and Dowe (2008, footnote 175))—should probably be scored with log-loss.

Acknowledgements

I thank Chris Wallace (1933–2004) (Dowe 2008) for giving me a proper appreciation of Bayesianism and for introducing me to—and training me in—Minimum Message Length (Wallace and Boulton 1968; Wallace 2005). I thank Dr Jason Grossman for alerting me to the (subtle) distinction in the section "Statistical Consistency" between statistical consistency (or even efficiency) and small-sample performance, and also for discussion motivating the section "Experimental Design, Data Collection Protocol and Likelihood Principle". As in Dowe (2008, footnote 116), I am grateful to Claire Leslie for telling me about the *bus number problem* from the section "Maximum Likelihood". And I thank those who anonymously reviewed the paper for useful guidance in helping me make the paper clearer and better.

Notes

[1] If $\tan^{-1}$ ranges from $-\pi/2$ to $\pi/2$, then we take the negative square root for $x < 0$. We can write this more properly as $(\kappa, \theta) = (\text{sign}(x) \cdot \sqrt{x^2 + y^2},\ \tan^{-1}(y/x))$.

References

Akaike, H. 1970. Statistical prediction information. *Annals of the Institute of Statistical Mathematics* 22: 203–17.

———. 1973. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd international symposium on information theory,* edited by B.N. Petrov and F. Csaki, pp. 267–81. Budapest: Akademiai Kiado.

———. 1987. Factor Analysis and AIC. *Psychometrika* 52 (3): 317–332.

Barron, A. R., and T. M. Cover. 1991. Minimum complexity density estimation. *IEEE Transactions on Information Theory* 37: 1034–54.

Berger, J. O., and R. L. Wolpert. 1988. *The likelihood principle,* 2nd edition, Institute of Mathematical Statistics monograph series. California, USA: Hayward.

Bernardo, J. M., and A. F. M. Smith. 1994. *Bayesian theory.* New York: Wiley.

Chaitin, G. J. 1966. On the length of programs for computing finite sequences. *Journal of the Association for Computing Machinery* 13: 547–69.

Clarke, D. M., G. C. Smith, D. L. Dowe, and D. P. McKenzie. 2003. An empirically-derived taxonomy of common distress syndromes in the medically ill. *Journal of Psychosomatic Research* 54: 323–30.

Comley, Joshua W., and David L. Dowe. 2003. General Bayesian networks and asymmetric languages. Paper presented at Proceedings of the Hawaii International Conference on Statistics and Related Fields, 5–8 June.

———. 2005. Minimum message length and generalized Bayesian nets with asymmetric languages. Chap. 11 in *Advances in minimum description length: Theory and applications (MDL handbook),* edited by P. Grünwald, M. A. Pitt, and I. J. Myung, pp. 265–94. Cambridge, MA: MIT Press.

Dowe, D. L. 2007. Discussion following "Hedging predictions in machine learning, A. Gammerman and V. Vovk". *Computer Journal* 2 (50): 167–8.

———. 2008. Foreword re C. S. Wallace. *Computer Journal* 51 (5): 523–560.

Dowe, D. L., L. Allison, T. I. Dix, L. Hunter, C. S. Wallace, and T. Edgoose. 1996. Circular clustering of protein dihedral angles by minimum message length. In *Pacific symposium on biocomputing '96,* edited by L. Hunter and T. Klein, pp. 242–55. Singapore: World Scientific.

Dowe, D. L., R. A. Baxter, J. J. Oliver, and C. S. Wallace. 1998. Point estimation using the Kullback–Leibler loss function and MML. In *Proceedings of the 2nd Pacific-Asia conference on research and development in knowledge discovery and data mining (PAKDD-98),* Volume 1394 of *LNAI,* edited by X. Wu, Ramamohanarao Kotagiri, and K. Korb, pp. 87–95. Berlin: Springer.

Dowe, D. L., G. E. Farr, A. J. Hurst, and K. L. Lentin. 1996. Information-theoretic football tipping. Paper presented at the 3rd Conference on Maths and Computers in Sport, pp. 233–41. [See also Technical Report TR 96/297, Dept. Computer Science, Monash University, Australia 3168, Dec 1996.]

Dowe, D. L., S. Gardner, and G. R. Oppy. 2007. Bayes not bust! Why simplicity is no problem for Bayesians. *British Journal for the Philosophy of Science* 58 (4): 709–54.

Dowe, D. L., and A. R. Hajek. 1998. A non-behavioural, computational extension to the Turing test. Paper presented at the International Conference on Computational Intelligence & Multimedia Applications (ICCIMA'98), Gippsland, Australia, February, pp. 101–6.

Dowe, D. L., and K. B. Korb. 1996. Conceptual difficulties with the efficient market hypothesis: Towards a naturalized economics. Paper presented at the Proceedings on Information, Statistics and Induction in Science (ISIS), pp. 212–23. [See also Technical Report TR 94/215, Dept. Computer Science, Monash University, Australia 3168, 1994.]

Dowe, D. L., and N. Krusel. 1993. *A decision tree model of bushfire activity.* Technical report TR 93/190, Dept. of Computer Science, Monash University, Clayton, Vic. 3800, Australia, September.

Dowe, D. L., J. J. Oliver, and C. S. Wallace. 1996. MML estimation of the parameters of the spherical Fisher distribution. In *Algorithmic learning theory, 7th international workshop, ALT '96,*

Sydney, Australia, October 1996, proceedings, Volume 1160 of *Lecture notes in artificial intelligence,* edited by S. Arikawa and A. Sharma, pp. 213–227. Berlin: Springer.

Dowe, D. L., and G. R. Oppy. 2001. Universal Bayesian inference? *Behavioral and Brain Sciences (BBS)* 24 (4): 662–3.

Dowe, D. L., and C. S. Wallace. 1997. Resolving the Neyman–Scott problem by Minimum Message Length. In *Proceedings of computing science and statistics – 28th symposium on the interface,* Volume 28, edited by L. Billard and N. I. Fisher, pp. 614–18. Interface Foundation of North America.

Edgoose, T., L. Allison, and D. L. Dowe. 1998. An MML classification of protein structure that knows about angles and sequence. In *Pacific symposium on biocomputing '98,* edited by R. B. Altman, A. K. Dunker, L. Hunter, and T. Klein, pp. 585–96. Singapore: World Scientific.

Edwards, R. T., and D. L. Dowe. 1998. Single factor analysis in MML mixture modelling. In *Proceedings of the 2nd Pacific-Asia conference on research and development in knowledge discovery and data mining (PAKDD-98),* Volume 1394 of *Lecture notes in artificial intelligence (LNAI),* edited by Xindong Wu, Ramamohanarao Kotagiri, and Kevin B. Korb, pp. 96–109. Berlin: Springer.

Fitzgibbon, L. J., D. L. Dowe, and F. Vahid. 2004. Minimum message length autoregressive model order selection. Paper presented at the Proceedings of the International Conference on Intelligent Sensors and Information Processing, Chennai, India, January, pp. 439–44. IEEE (IEEE Press).

Forster, M., and E. Sober. 1994. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science* 45: 1–35.

Glymour, C. 1981. Why I am not a Bayesian. *Theory and Evidence,* edited by C. Glymour and D. Stalker, pp. 63–93. Princeton: Princeton University Press.

Grossman, J. Forthcoming. The likelihood principle. In *Handbook for philosophy of science,* Volume 7*, Philosophy of statistics.* New York: Elsevier.

Grünwald, Peter D., and John Langford. 2007. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning* 66 (31): 119–149.

Hernández-Orallo, José. 2000. Beyond the Turing test. *Journal of Logic, Language and Information* 9 (4): 447–66.

Jeffreys, H. 1946. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A* 186: 453–4.

Kissane, D. W., S. Bloch, D. L. Dowe, R. D. Snyder, P. Onghena, D. P. McKenzie, and C. S. Wallace. 1996. The Melbourne family grief study, I: Perceptions of family functioning in bereavement. *American Journal of Psychiatry* 153: 650–8.

Kissane, D. W., S. Bloch, P. Onghena, D. P. McKenzie, R. D. Snyder, and D. L. Dowe. 1996. The Melbourne family grief study, II: Psychosocial morbidity and grief in bereaved families. *American Journal of Psychiatry* 153: 659–66.

Kolmogorov, A. N. 1965. Three approaches to the quantitative definition of information. *Problems of Information Transmission* 1: 4–7.

Kornienko, L., D. W. Albrecht, and D. L. Dowe. 2005a. A preliminary MML linear classifier using principal components for multiple classes. In *Proceedings of the 18th Australian joint conference on artificial intelligence (AI'2005),* Volume 3809 of *Lecture notes in artificial intelligence (LNAI),* Sydney, Australia, edited by S. Zhang, and Ray Jarvis, pp. 922–6. Berlin: Springer.

———. 2005b. *A preliminary MML linear classifier using principal components for multiple classes.* Technical report CS 2005/179, School of Computer Sci. & Softw. Eng., Monash Univ., Melb., Australia.

Kornienko, Lara, David L. Dowe, and David W. Albrecht. 2002. Message length formulation of support vector machines for binary classification – A preliminary scheme. In *Proceedings of the 15th Australian Joint Conference on Artificial Intelligence,* Volume 2557 of *Lecture notes in artificial intelligence (LNAI),* edited by B. McKay, and J. K. Slaney, pp. 119–130. Berlin: Springer-Verlag.

Legg, S., and M. Hutter. 2007. Universal intelligence: A definition of machine intelligence. *Minds and Machines* 17 (4): 391–444.

McKenzie, D. P., P. D. McGorry, C. S. Wallace, L. H. Low, D. L. Copolov, and B. S. Singh. 1993. Constructing a minimal diagnostic decision tree. *Methods in Information in Medicine* 32: 161–6.

Needham, S. L., and D. L. Dowe. 2001. Message length as an effective Ockham's razor in decision tree induction. Paper presented at the 8th International Workshop on Artificial Intelligence and Statistics (AI+STATS 2001), pp. 253–60.

Neyman, J., and E. L. Scott. 1948. Consistent estimates based on partially consistent observations. *Econometrika* 16: 1–32.

Ooi, J. N., and D. L. Dowe. 2005. Inferring phylogenetic graphs of natural languages using minimum message length. Paper presented at CAEPIA 2005 (11th Conference of the Spanish Association for Artificial Intelligence), Volume 1, pp. 143–52.

Pilowsky, I., S. Levine, and D.M. Boulton. 1969. The classification of depression by numerical taxonomy. *British Journal of Psychiatry* 115: 937–45.

Prior, M., R. Eisenmajer, S. Leekam, L. Wing, J. Gould, B. Ong, and D. L. Dowe. 1998. Are there subgroups within the autistic spectrum? A cluster analysis of a group of children with autistic spectrum disorders. *Journal of Child Psycholology and Psychiatry* 39 (6): 893–902.

Rissanen, J. J. 1978. Modeling by shortest data description. *Automatica* 14: 465–71.

Schwarz, G. 1978. Estimating dimension of a model. *Annals of Statistics* 6: 461–4.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27: 379–423 and 623–56.

Solomonoff, R. J. 1964. A formal theory of inductive inference. *Information and Control* 7: 1–22, 224–54.

Tan, P. J., and D. L. Dowe. 2002. MML inference of decision graphs with multi-way joins. In *Proceedings of the 15th Australian Joint Conference on Artificial Intelligence,* Volume 2557 of *Lecture notes in artificial intelligence (LNAI),* edited by R. McKay and J. Slaney, pp. 131–42. Berlin: Springer Verlag.

———. 2003. MML inference of decision graphs with multi-way joins and dynamic attributes. In *Proceedings of the 16th Australian Joint Conference on Artificial Intelligence,* Volume 2903 of Lecture Notes in Artificial Intelligence (*LNAI*), edited by T. D. Gedeon, and L. Chun Che Fung, pp. 269–81. Berlin: Springer.

———. 2004. MML inference of oblique decision trees. In *Proceedings of the 17th Australian Joint Conference on Artificial Intelligence,* Volume 3339 of *Lecture Notes in Artificial Intelligence (LNAI),* edited by G. I. Webb, and Xinghuo Yu, pp. 1082–8. Berlin: Springer.

———. 2006. Decision forests with oblique decision trees. In *Proceedings of the 5th Mexican international conference on artificial intelligence,* Volume 4293 of *Lecture Notes in Artificial Intelligence (LNAI),* edited by A. F. Gelbukh, and C. A. Reyes García, pp. 593–603. Berlin: Springer.

Tan, P. J., D. L. Dowe, and T. I. Dix. 2007. Building classification models from microarray data with tree-based classification algorithms. In *Proceedings of the 20th Australian Joint Conference on Artificial Intelligence,* Volume 4830 of *Lecture Notes in Artificial Intelligence (LNAI),* edited by M. A. Orgun, and J. Thornton, pp. 589–98. Berlin: Springer.

Vapnik, V. N. 1995. *The nature of statistical learning theory.* Berlin: Springer.

Visser, Gerhard, and D. L. Dowe. 2007. Minimum message length clustering of spatially-correlated data with varying inter-class penalties. In *Proceedings of the 6th IEEE international conference on computer and information science (ICIS) 2007,* pp. 17–22. Piscataway, NJ: IEEE Press.

Wallace, C. S. 1995. *Multiple factor analysis by MML estimation.* Technical report CS TR 95/218, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, Clayton, Melbourne, Australia.

———. 1996. False oracles and SMML estimators. In *Proceedings of the Information, Statistics and Induction in Science (ISIS) Conference,* edited by D. L. Dowe, K. B. Korb, and J. J Oliver,

pp. 304–316. Singapore: World Scientific. [Was previously Tech Rept 89/128, Dept. Comp. Sci., Monash Univ., Australia, June 1989.]

———. 1997. *On the selection of the order of a polynomial model.* Technical report, Royal Holloway College, England, UK. Chris released this in 1997 (from Royal Holloway) in the belief that it would become a Royal Holloway Tech Rept dated 1997, but it is not clear that it was ever released there. Soft copy certainly does exist, though. Perhaps see www.csse.monash.edu.au/~dld/CSWallacePublications; INTERNET.

———. 1998. Intrinsic classification of spatially correlated data. *Computer Journal* 41 (8): 602–611.

———. 2005. *Statistical and inductive inference by minimum message length.* Information Science and Statistics series. Berlin: Springer Verlag.

Wallace, C. S., and D. M. Boulton. 1968. An information measure for classification. *Computer Journal* 11 (2): 185–94.

———. 1975. An invariant Bayes method for point estimation. *Classification Society Bulletin* 3 (3): 11–34.

Wallace, C. S., and D. L. Dowe. 1993. *MML estimation of the von Mises concentration parameter.* Technical Report 93/193, Dept. of Computer Science, Monash University, Clayton 3168, Australia, December.

———. 1994. Intrinsic classification by MML – the Snob program. In *Proceedings of the 7th Australian Joint Conference on Artificial Intelligence,* edited by C. Zhang, J. Debenham and D. Lukose pp. 37–44. Singapore: World Scientific.

———. 1999a. Minimum message length and Kolmogorov complexity. *Computer Journal* 42 (4): 270–283.

———. 1999b. Refinements of MDL and MML coding. *Computer Journal* 42 (4): 330–7.

———. 1999c. Rejoinder. *Computer Journal* 42 (4): 345–7.

———. 2000. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing* 10 (1): 73–83.

Wallace, C. S., and P. R. Freeman. 1987. Estimation and inference by compact coding. *Journal of the Royal Statistical Society series B* 49 (3): 240–52. See also Discussion on pp. 252–65.

———. 1992. Single-factor analysis by minimum message length estimation. *Journal of the Royal Statistical Society B* 54 (1): 195–209.