Evidence based medicine can't be . . . *

Adam La Caze
University of Sydney
alacaze@mac.com

Abstract

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 trumps 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.

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 et al., 2006). EBM has also recently gained the attention of philosophers of science (Worrall, 2002, 2007b,a; Bluhm, 2005;

*The final version of this paper is been published in *Social Epistemology*, 22:4,3553–370. DOI: http://dx.doi.org/10.1080/02691720802559438. Please reference published paper.

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 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, p. 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]

[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

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

claims is provided by Rothwell (2007).

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 1980's and 1990's. 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 de-emphasizes intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient grounds for clinical decision making and stresses the examination of ev-

[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 sucessfully convince many in the medical community for the application of clinical epidemiological ideas to be the 'benchmark' when making, or justifying, medical decisions.

> idence 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' judgement 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 focussed 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 labelling 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)

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

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 the table for the hierarchy provided by Guyatt and Rennie (2002, p. 12)[6] (for a more detailed version, see Phillips et al. (2001)):

[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.

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

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 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-contol 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, p. 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 evdience 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 physio-

> logic studies that are related only indirectly—but there is always evidence. (Guyatt and Rennie, 2002, p. 14–5)

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 criticism's of EBM have focussed 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, p. 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 proving the only 'valid', unalloyed, genuinely scientific evidence about the effectiveness of any therapy.

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) 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.

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 et al., 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, p. 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, p. 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

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

> 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, p. 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 judgement 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 judgement 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.

[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 judgements about external validity is increased.

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

[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, p. 983) recognises this point.

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)

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 inform *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 heirarchy, 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 read-

[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.

> ing it and go on to the next article in your search. (Straus et al., 2005, p. 118)

This does not suggest an interpretation of the hierarchy where only certain 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, p. 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 is 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 are 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

[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.

offer. (Collins and MacMahon, 2007, p. 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

[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).

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

[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.

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 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,

[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.

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, can not 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.

References

Armitage, P. 1982. The role of randomization in clinical trials. *Statistics in Medicine*, 1(345-352).

Bluhm, R. 2005. From heirarchy 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 assesment of the effects of treatments on mortality and major morbidity. In Rothwell (2007).

Egger, M. and G. D. Smith. 1995. Misleading meta-analysis. *British Medical Journal*, 310(6982):752–754.

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):629–634.

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):516–534.

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):995–998.

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):924–926.

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

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.

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):1197–1203.

Lindley, D. V. 1982. The role of randomization in inference. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1982:431–446.

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):239–247.

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):333–380.

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):991–992.

Rothwell, P. M. 2005. Treating individuals 2: Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365(9454):176–186. TY - JOUR.

Rothwell, P. M., Editor. 2007. *Treating Individuals: From randomised trials to personalised medicine*. 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*. Elsevier Churchill Livingstone, third edition.

Suppes, P. 1982. Arguments for randomizing. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1982:464–475.

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 J. C. I. Bailar and F. Mosteller, Editors, *Medical Uses of Statistics*, page 480. Boston: NEJM Books.

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

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

Worrall, J. 2007b. Why There's No Cause to Randomize. *British Journal for the Philosophy of Science*, 58(3):451–488.