Title: Evidence based medicine must be … [‡]

Author: Adam La Caze

School of Pharmacy,

The University of Queensland

Abstract:

Proponents of evidence based medicine (EBM) provide the 'hierarchy of evidence' as a criterion for judging the reliability of therapeutic decisions. EBM's hierarchy places randomised interventional studies (and systematic reviews of such studies) higher in the hierarchy than observational studies, unsystematic clinical experience, and basic science. Recent philosophical work has questioned whether EBM's special emphasis on evidence from randomised interventional studies can be justified. Following the critical literature, and in particular the work of John Worrall, I agree that many of the arguments put forward by advocates of EBM do not justify the ambitious claims that are often made on behalf of randomisation. However, in contrast to the recent philosophical work, I argue that a justification for EBM's hierarchy of evidence can be provided. The hierarchy should be viewed as a hierarchy of comparative internal validity. While this justification is defensible, the claims that EBM's hierarchy substantiates when viewed in this way are considerably more circumscribed than some claims found in the EBM literature.

[‡] This is a preprint of La Caze A. (2009). Evidence Based Medicine must be…. *Journal of Medicine and Philosophy*, 34:509–527. The final published version is available at http://dx.doi.org/10.1093/jmp/jhp034. Please reference published paper.

1 Introduction

Evidence based medicine (EBM) plays a significant role in contemporary medical decision making. EBM's primary claim is that its 'hierarchy of evidence' provides a criterion for judging the reliability of evidence for therapeutic decisions. The higher up the hierarchy, the more reliable the evidence, and the better the justification it gives to therapeutic decisions. The hierarchy divides 'evidence' partly according to methodology, and partly according to proximity to clinical practice.[1] Evidence from randomised trials involving patients, and systematic reviews of such trials, fill the higher tiers of the hierarchy;[2] evidence from non-randomised ('observational') studies occupy the middle tiers of the hierarchy;[3] and, evidence from basic sciences

[1] See, for example, the hierarchy provided in Guyatt and Rennie (2002, 12), and Phillips et al. (2001).

[2] In randomised studies patients are enrolled and allocated to 'treatment' and 'control' groups via a random process. They are then observed over time. At the end of the study the outcomes for each group of patients is compared.

[3] The two main types of observational study are cohort and case-control. Cohort studies find a sample that contains both individuals who are 'exposed' to a treatment (or risk factor) and individuals not exposed. The incidence of the outcome of interest is then compared between these two groups. Case-control studies start with the outcome of interest. A group of individuals who have suffered the outcome, the 'cases', are compared to a group who have not suffered the outcome but are similar in other respects, the 'controls'. The exposure of these groups to the treatment (or risk factor) is then compared. Whereas randomised studies intervene on a sample of patients, observational studies follow participants going about their lives.

such as physiology and pharmacology, and the ‘unsystematic’ evidence of clinical experience fill the lower tiers. EBM’s primary claim is epistemological, however, despite the large literature on EBM, there is very little on the important task of justifying EBM’s epistemological claims. EBM’s starting assumption is that the hierarchy provides increasingly reliable evidence for therapeutic decisions, and it stands in need of a detailed justification.

Proponents of EBM make strong claims on behalf of the evidence hierarchy. Randomised studies are seen to provide especially secure evidence for therapeutic decisions.

> 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)

> 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)

A systematic justification of EBM requires (i) an interpretation of the hierarchy, which describes clearly how it is to be applied in therapeutic decision making, and (ii) a justification for that interpretation, which explains why applying the hierarchy as proposed more reliably informs therapeutic decisions.

This paper focuses on arguments provided by proponents of EBM for the hierarchy of evidence. While there are few arguments provided explicitly for EBM's *hierarchy*, there are a number of arguments that have been provided for informing therapeutic decisions on the basis of evidence from randomised trials. I follow the critical literature, and in particular the work of John Worrall (2002; 2007a; 2007b), in finding that these arguments do not substantiate the very special scientific weight placed in randomised trials by proponents of EBM. But in contrast to Worrall, I argue that a defensible interpretation of EBM's hierarchy can be provided—albeit an interpretation that substantiates much less than what is often claimed by proponents of EBM.

I propose that EBM's hierarchy should be interpreted as a hierarchy of comparative internal validity. 'Internal validity' is the degree to which the results of a study are accurate for the sample of patients included in the study (Fletcher 1996, 12). By 'comparative internal validity' I mean, all other things being equal, studies that utilise the methods higher in EBM's hierarchy, have higher internal validity than studies designed according to the methods lower down the hierarchy. Comparative internal validity is an under-recognised argument for the hierarchy of evidence. While the argument is present in the clinical literature, the claims that it substantiates are considerably more circumscribed than those made by advocates of EBM (and illustrated in quotes such as those provided by Straus, and Collins and MacMahon above). This and the other arguments that have been provided for EBM's hierarchy are examined in Section 2. In section 3, I sketch how EBM's hierarchy can be applied when it is viewed as a hierarchy of comparative internal validity, and illustrate the considerably more limited claims that can be justified on this basis.

2 Arguments for EBM’s hierarchy

A justification for how EBM’s hierarchy better informs therapeutic decisions is not found in the popular guidebooks. Straus et al. (2005, xii) suggest

> Those who wish, and have time for, more detailed discussions of the theoretical and methodological bases for the tactics described here should consult one of the longer textbooks on clinical epidemiology.

One of the textbooks they refer to is their own, Haynes et al. (2006) (then forthcoming). The argument for randomisation is summarised in the chapter written by David Sackett (2006), I quote in full.

> [M]ightn’t a high-quality cohort study be as good as, or even better than, an RCT for determining treatment benefit? Some methodologists have vigorously adopted this view. I disagree with them, for two reasons. First, there are abundant examples of the harm done when clinicians treat patients on the basis of cohort studies. Two recent examples of cohort-based treatment recommendations that failed in RCTs are postmenopausal oestrogen plus progestogen for healthy women and vitamin E for coronary heart disease. (Note my argument here does not apply to determining treatment harm, where observational studies are often the only way to detect a treatment's rare but awful adverse effects.)
>
> My second justification is an unprovable act of faith. It professes that the gold standard for determining the effectiveness of any health intervention is a high-quality systematic review of all relevant, high-quality RCTs. When the other study architectures are measured against this gold standard, they have generated less reliable estimates of effectiveness. For example, Regina Kunz and her colleagues performed a Cochrane Review of randomisation as a protection against selection bias in health care trials. They frequently found a worse prognosis at entry among control patients in nonrandomised studies. Moreover, they documented the over-estimation of treatment effects when the randomisation schedule was not concealed from the

> clinicians who were inviting patients to join RCTs, converting these 'RCTs' into cohort studies.

The candour is refreshing, but the argument is far from compelling. Sackett's first argument is empirical—experience, he suggests, has shown randomised trials to be more reliable. Sackett's 'second justification' is difficult to differentiate from the first. Indeed, it appears to be a repetition of the first argument with some acknowledgement that, as this is an empirical argument, it is not—by Sackett's reckoning—compelling, rather, it is an 'unprovable act of faith'.

I review this and additional arguments that have been provided for the hierarchy by proponents of EBM. Particular focus is given to the distinction made between randomised and non-randomised trials. In doing so I closely follow the arguments provided by Worrall (2002; 2007a; 2007b). Worrall critiques the empirical justification as well as a number of additional arguments for the necessity of randomisation in clinical trials. He comes to the view that randomisation, while often benign, is not *essential*. By 'essential' Worrall means that randomised studies are not 'essential for any truly scientific conclusion to be drawn from trial data' (Worrall 2007a, 452). I review three of the arguments provided for randomisation that Worrall critiques: (i) the empirical justification of EBM's hierarchy, (ii) the view that randomising controls for all confounding factors, known and unknown,[4] and (iii) that randomising uniquely prevents selection bias.[5]

There is much to agree with in Worrall's analysis, in particular, he shows that many of the ambitious claims made by proponents of EBM on behalf of randomisation can not be justified. But, in finding no argument for randomisation to be 'essential' for science, Worrall concludes that no epistemological distinction can be drawn between randomised trials and observational studies.

> The best we can do (as ever) is test our theories against rivals that seem plausible in the light of background knowledge. Once we have eliminated other explanations that we know are possible (by suitable, or *post hoc*, control) we have done as much as we can epistemologically. (Worrall 2007a, 486)

This misses the argument that EBM's hierarchy can be justified as a hierarchy of comparative internal validity. Despite falling short of showing randomised studies are essential for drawing scientific conclusions from data, the argument of comparative internal validity substantiates an epistemological distinction being drawn between randomised studies and observational studies in clinical science.

[4] In statistics, a 'confounding factor' is a third variable, which correlates with two variables that are being investigated for a potential causal relationship. The confounder may either mask a 'true' causal relationship between the two variables under investigation, or make it appear as though a causal relation exists between the variables under investigation when in fact both variables are under the influence of the confounder.

[5] Worrall provides a critique of another argument for randomisation. The argument that randomisation is necessary because it provides the basis for the classical statistical analysis. Assuming for the moment that the EBM hierarchy is not necessarily tied to classical statistical analysis, this argument less relevant to our concerns. Bayesian statistical inference, for instance, does not require randomisation.

Comparative internal validity as an argument for the distinctions made in EBM's hierarchy can be found in the epidemiological literature, but it has been under-emphasised in the philosophical discussions to date. One reason for this is that proponents of EBM have focused on disseminating, advocating and teaching EBM, rather than providing a philosophical justification of the view. The argument for EBM's hierarchy is found elsewhere—notably, in the clinical epidemiological literature. Another reason internal validity has been under-emphasised in the philosophical literature is that the philosophical analyses take the ambitious claims that proponents of EBM have made as a starting point, and search for arguments that could substantiate these claims (Bluhm 2005; Grossman and Mackenzie 2005; Upshur 2005; Worrall 2007b). While this is appropriate, the arguments regarding comparative internal validity are given little attention because they do not substantiate EBM's more ambitious claims. Changing tack, and focussing on arguments that are available in the epidemiological literature highlights the importance of comparative internal validity as a justification of EBM's hierarchy. The claims that can be substantiated on the basis of this justification can then be examined.

While Worrall shows that randomisation is not *essential*, I show the important role randomised interventional studies play in testing certain well-defined therapeutic questions. This discussion may help explain why clinicians are so enamoured by randomised studies, while at the same time avoid the mistake (so often made within the EBM literature) of claiming too much on its behalf.

2.1 The empirical justification of EBM's hierarchy

The empirical justification for EBM's hierarchy, as provided by Sackett in the previous section, cites studies that found randomised trials provide more conservative estimates of treatment effects than non-randomised (observational) studies.[6] Proponents of this argument contend that observational studies provide less conservative estimates of treatment effects because of biases inherent in comparing groups that have not been randomly allocated.[7] The point at issue, however, is whether—and, importantly, how—randomised studies provide more reliable evidence for therapeutic decisions. Let's accept the data Sackett is citing. How does this support the conclusion that the estimates provided by the randomised studies are more reliable? What stops the opposite conclusion: that observational studies are *correct*, or more *likely to be correct*, and randomised studies *under-estimate* treatment effects? The data alone provides no justification for asserting that randomised studies provide the correct estimates. The empirical justification of EBM's hierarchy requires the premiss that randomised studies *are* more reliable in order to make the claim that observational studies over-estimate treatment effects.

Empirical arguments do not provide a justification *for* EBM's hierarchy. They are circular. If you already accept that randomised trials are

[6] See for example Chalmers (1977) and Sacks (1982).

[7] Presumably, one of these biases is also publication bias. If observational studies are more biased, and publication bias was *not* present, then both over-estimation and under-estimation of treatment effects would be expected—not just over-estimation.

the ‘gold-standard’, then the data cited by EBM proponents are grist for your mill. But the data won't compel a sceptic. Both Worrall (2002, S326) and Grossman and Mackenzie (2005, 520) make this point.

While this objection alone is strong enough to sink the empirical argument as a *justification* for EBM’s hierarchy, there are further problems. As Worrall (2007b, 1009–13) notes, recent reports comparing the findings of randomised and observational studies contradict the earlier reports. Benson (2000) and Concato (2000) found estimates from observational studies were consistent with those found in randomised trials in a range of therapeutic areas. Concato (2000) suggest that earlier comparisons of the study methodologies focused on less rigorous observational studies. Therefore, even on its own terms, the empirical data underpinning the argument for the evidence hierarchy is poor.

At the heart of EBM is an epistemological claim: evidence from higher up the hierarchy provides more reliable evidence for therapeutic decisions. As such, EBM’s hierarchy requires a philosophical justification.

2.2 Randomisation controls for *all* confounding factors

Clinical trials are conducted in order to test the effects of an intervention on a defined group of patients; typically trials are set up in order to test whether the intervention causes the beneficial effects suggested by research in basic science, or previous experience. Trials are set up in such a way as to ensure, as much as is possible, that any observed differences between the treatment and control group are due to the effects of the intervention. To achieve this the groups must be as similar as possible. One way to ensure the comparability of the groups is to match the treatment and control groups for all known confounding factors. Prospective cohort studies, for instance, match the experimental groups in this way. Obviously, however, cohort studies are unable to guarantee that the experimental groups are also matched for *unknown* confounders. A common claim in the clinical literature is that randomisation provides this guarantee.

Collins and MacMahon (2007, 23), complain that

> […] non-randomised methods do not provide assurance that all sources of known and unknown bias are adequately controlled, and so cannot exclude the possibility that moderate biases have obscured or inflated any moderate effects, or have falsely indicated a treatment effect when none existed.

According to Collins and MacMahon, randomised methods provide the assurance that all known and unknown confounders are adequately controlled.[8] While there is a sense in which randomisation provides this assurance, it rests on an important ambiguity in the use of the term 'bias' (an ambiguity which will be clarified shortly). Importantly, it is because randomisation is thought to eliminate bias due to confounding factors that it is seen as essential in clinical trials. Being clear on what sense of bias is 'eliminated', shows this claim to be false.[9]

[8] Collins and MacMahon are not the only ones to make this claim. For instance, Kendall et al. (1983) state the following: '[…] by the very nature of the randomisation process, the effects of factors outside the experiment can show no favour to the factors inside it, and our inferences are free from bias.' Worrall provides a number of additional examples (see Worrall (2002, S321–4)).

[9] I am indebted to conversations with Jason Grossman on this point. The comments I make about 'bias' pick up on points made in (Grossman and Mackenzie 2005, 518).

‘Bias’ is a term that is variously applied in the clinical literature. Often it is used informally to refer to any factor that could obscure the ‘true’ results of a trial. (Worrall, for one, uses the term in this way). However, ‘bias’ is used in statistics in a number of more formal ways. In parametric statistics, *statistical* bias refers to the *expectation* of an estimator of a parameter. An estimator is unbiased if its expectation is equal to the true value of the parameter. This is often informally referred to as an estimator's *long run* average.[10] It is *only* in the statistical sense that randomisation eliminates bias due to confounding factors, known and unknown. This is because statistical bias entails consideration of the entire sample space. (Note. In the quote above, Collins and MacMahon can only be referring to statistical bias; otherwise the claim is false).

Consider a population of a trial being randomly allocated to treatment or control. On any given allocation it is possible that the treatment and control group are not equally balanced for a particular confounder. Indeed, as Worrall (2002, S324) has pointed out, given that there could be indefinitely many possible confounding factors, the probability that a confounder is not equally balanced between the groups on any *particular* allocation is high. However, this is not so in the indefinite sequence of trials. If the trial is repeated indefinitely, with a new allocation of the population performed each time, then, in this indefinite sequence of trials, the effect of any unbalanced confounder in a particular allocation will be counteracted by the distribution of that confounder in other allocations. The net effect of all possible confounders

[10] But this is not strictly correct. Rather it is the estimator's average over the sample space; any actual long run (even an infinite long run) won't necessarily equal its average over the sample space.

on an unknown parameter in the indefinite sequence of trials will be zero. Hence, randomisation eliminates statistical bias due to confounding factors.

Now that the term 'bias' has been disambiguated, the problem for EBM can be made clear. That randomisation eliminates bias due to all known and unknown confounding factors is one of the key arguments EBM employs to justify randomised studies having a special epistemic place in the hierarchy of evidence. Once it is clear what this 'elimination of bias due to confounders' actually amounts to, it is reasonable to question whether it is sufficient to justify the emphasis placed in randomised trials by EBM. Recall the advice of one of the EBM guidebooks: if a study is not randomised, 'bin it'. Bias due to confounders is only eliminated in the indefinite sequence of trials, but (of course) the trials that inform therapeutic decisions are not repeated (at all, let alone indefinitely). In any particular trial allocation a confounder may be distributed unevenly between the experimental groups. While the groups can be examined for the imbalance of any known confounders, this is obviously not possible for unknown factors. Lack of statistical bias provides no assurance with regard to the actual allocation of the trial. In order to substantiate its ambitious claims, EBM needs randomisation to eliminate the informal sense of bias. But, of course, randomisation does not provide this kind of assurance.

While Worrall does not disambiguate 'bias' in this way, he makes a similar point in relation to randomisation.

> The fact is that the subjects have been randomised between control and experimental group only once, and that division either is or is not balanced for the unknown factor at issue. Suppose it is unbalanced, and that this throws the conclusion about the

> efficacy of the treatment off, then it seems to me scant consolation to be told that—although you don't and can't know it,—you were ‘unlucky’, and if the randomisation had been repeated indefinitely you would, in the indefinite long run, have inevitably realised your mistake. (Worrall 2007a, 484)

This, Worrall argues, undermines the view that randomisation is ‘essential’, or *sine qua non*, for clinical trials. Importantly, rejecting these more ambitious claims is consistent with accepting that randomisation plays an important role in the clinical sciences. This role shall be outlined after considering a third argument for randomisation—that randomisation prevents selection bias.

2.3 Randomisation prevents ‘selection bias’

The prevention of ‘selection bias’ is the one ‘cast-iron’ argument for randomisation that Worrall (2007b, 1009) concedes. But while he accepts that the prevention of selection bias provides a reason to randomise, Worrall argues that selection bias (as he defines it) can be avoided through other measures. Worrall’s definition of ‘selection bias’, however, is considerably narrower than the conception of ‘selection bias’ in the epidemiological literature. Under the broader conception, ‘selection bias’ provides a rationale for randomisation that can not be provided by alternative measures.[11]

‘Selection bias’, according to Worrall, is the bias that can occur when trial investigators allocate patients to treatment or control. I will call this ‘investigator-selection bias’. Having the investigators allocate patients to the

[11] While Worrall (2007b, 1008) acknowledges his notion of selection bias is narrower than that used in the medical literature, he does not appear to see the consequences of this difference for his argument.

experimental groups can obscure the analysis in a number of ways.[12] Perhaps the investigators (subconsciously or otherwise) preferentially allocate patients who they judge more likely to respond favourably to the treatment arm. Or, perhaps those patients with recalcitrant illness, or a propensity for side effects, are allocated to the control arm. Allocations such as these have the potential to significantly confound the analysis. Further, if the same investigators that allocated patients to experimental groups are responsible for the subject's treatment and collection of results—that is, the study is at best single blind—then there is a substantial risk of further bias to enter the analysis. For instance, the investigator's knowledge of a participant's treatment allocation may influence how they treat the patient (even a higher level of implicit encouragement may make important differences for some conditions). And, perhaps even more importantly, knowledge of treatment allocation may influence how investigators interpret the patient's response. Clearly, these possible sources of bias undermine our ability to draw the right inference regarding the treatment's efficacy.

Worrall accepts that randomisation prevents investigator-selection bias. But he notes that it does this via two mechanisms: by taking allocation out of the control of the investigators, and by permitting double-blinding.

> Notice however that randomisation as a way of controlling for selection bias is very much a means to an end, rather than an end in itself. The important methodological point is that control of which arm of trial a particular patient ends up in is taken away

[12] Here the informal notion of 'bias' is sufficient. That is, bias is any factor that will obscure the results of the trial from reflecting the real effect of treatment. When statisticians refer to bias, as in selection bias, it can be difficult to discern whether they are referring to bias informally, or 'statistical' parameter bias. Here, at least, the ambiguity does not cause any problem.

> from the experimenters—randomisation (as normally performed) is simply one method of achieving this. (Worrall 2007b, 1009)

Worrall takes this as further support for his conclusion that randomisation is not *essential* for evidence in medicine, but there are problems with this analysis.

Alternative methods for removing investigator-selection bias are only possible for certain study designs. In particular, the design needs to be one in which the allocation of patients into experimental groups *can* be taken out of the hands of investigators—that is, the study needs to be interventional. Whereas randomised studies in EBM's hierarchy are interventional, observational studies are not. While alternative methods can be used to avoid investigator-selection bias in interventional studies, such methods are *not* available in the non-randomised studies that EBM is referring to. In observational studies the choices, and the myriad of other factors, that have caused patients to fall into the 'treatment' (or 'case') and the 'control' groups have already played their part.

As it is used in the clinical literature 'selection bias' occurs when 'comparisons are made between groups of patients that differ in ways, other than the main factors under study, that affect the outcome of the study' (Fletcher 1996, 7–8). This definition incorporates Worrall's investigator-selection bias, but also includes other forms of selection bias. This broader notion of selection bias includes 'patient-selection bias'. This is the bias that can occur when patients 'select' which experimental group they will be a member of. In observational studies this 'selection' is typically anything but explicit. Observational cohort and case-control studies observe patients as they go about their lives, exposing themselves as they do to certain treatments and risk factors, sometimes for identifiable reasons, but often as much due to circumstance. It is this form of selection bias that can be much more difficult to identify and remove from observational studies. For observational studies, taking the allocation of experimental groups out of the hands of the investigators is not possible.

Hence, the avoidance of certain types of selection bias (investigator-, and patient-selection bias) is an important argument for randomised trials in the clinical sciences. Or, more correctly, an important argument for prospective interventional studies. It is not randomisation *per se* that permits the avoidance of selection bias, but that these studies are prospective and interventional. Clearly, there are corollaries to Worrall's argument: *randomisation* is not essential—other methods could be used to take the allocation of subjects out of the control of investigators. But the advantage of interventional studies—that allocation of participants *can* be taken out of the hands of investigators—is important, and it is missed on Worrall's analysis due to his narrow conception of selection bias. Indeed, the important distinction in EBM's methodological hierarchy is not randomised versus non-randomised, but randomised interventional versus non-interventional. Whether randomisation, or an alternative method for reducing selection bias, is used in prospective interventional studies is certainly worthy of debate. But this debate is orthogonal to the questions raised by EBM and its methodological hierarchy once the interventional/non-interventional distinction is made clear.

Prospective interventional randomised studies have a benefit over observational studies in that, when properly done, they rule out a certain type of selection bias (patient-selection bias, as I have labelled it here). Observational studies cannot rule out this bias. Of course, this benefit does not make randomised interventional studies infallible, nor does it mean that these studies are 'essential for any truly scientific conclusion to be drawn from trial data'. Both randomised trials and observational studies employ a vast range of methods to provide results that are as reliable as possible. There are many potential biases, including other forms of selection bias, that can occur in randomised interventional studies. And there is also a host of methods to assist isolating, and reducing the influence of selection bias in observational studies—including assuring that the control and treatment group are well matched according to background knowledge.

Worrall is right to call proponents of EBM on their over-ambitious claims with regard to randomisation. But it is entirely consistent, while rejecting the over-ambitious claims of EBM, to support arguments for randomised interventional studies that appeal to somewhat more modest benefits. This opens the way for a justification of randomised trials, and EBM's hierarchy, that is both present in the clinical literature, and defensible.

2.4 All other things being equal, randomised trials have higher internal validity compared to alternative methods

The reduction of selection bias that is achieved through randomised interventional studies is the key argument found in the clinical literature for marking an epistemic distinction between randomised interventional studies and observational studies. Any difference between the groups under comparison, other than treatment, that can influence the outcome of the study is a potential source of selection bias. Compared with randomised interventional studies, observational studies are more prone to selection bias—even when the experimental groups of a cohort study are well matched according to background knowledge, and *post hoc* adjustment has been conducted. Collins and MacMahon (2007, 16) provide a summary of the argument.

> As discussed, randomisation minimises systematic errors (i.e. biases) in the estimates of treatment effects, allowing any moderate effects that exist to be detected unbiasedly in studies of appropriately large size. By contrast, observational studies—such as cohort studies and case-control studies—involve comparisons of outcome among patients who have been exposed to the treatment of interest, typically as part of their medical care, with outcome among others who were not exposed (or comparisons between those with different amounts of exposure). The reasons why certain patients received a particular treatment while others did not are often difficult to account for fully, and, largely as a consequence, observational studies are more prone to bias than are randomised trials.

This is the justification that is provided by Collins and MacMahon for EBM's distinction between randomised trials and observational studies; and it is seen *repeatedly* in the clinical literature and epidemiological textbooks. It is important to be clear as to what this is an argument for. The argument is one of comparative internal validity: all other things being equal, compared to alternative methods, randomised trials have higher internal validity. What kinds of claims does this argument substantiate for EBM?

Internal validity asks the question: How likely are the results of the study to be true for the participants involved? By contrast, ‘external validity’ refers to how well the findings of a study can be generalised to hold in patients not directly involved in the study. In clinical medicine, there is a focus on informing therapeutic decisions, and therefore an important balance needs to be achieved between internal and external validity. Clearly, it is important for the results of any study to be an accurate reflection of what has occurred in the sample population, but internal validity is not sufficient for reliably informing therapeutic decisions. Therapeutic decisions need not only the results of clinical studies to be accurate for the sample population, but also for the patients who will be treated with the intervention. (At the very least, a principled way of justifying how the results of a trial apply to patients presenting at the clinic is required.)

A useful distinction made in relation to drug treatments is that of ‘efficacy’ versus ‘effectiveness’. ‘Efficacy’ refers to the effects of the drug under experimental conditions. ‘Effectiveness’, on the other hand, refers to the effects of the drug in typical patients under routine care. Prior to the drug reaching the market you want to ensure the drug is efficacious. Once on the market it is the effectiveness of the drug in the patients who will be treated that is paramount.

Improvements in internal validity are achieved through two mechanisms: (i) by placing the participants of the study under ‘experimental’ conditions, to reduce, as much as possible, some of the many things that could influence the participants’ progress other than the treatment under investigation, and (ii) by excluding participants who will complicate the analysis, that is, by ensuring the experiment is conducted on a relatively homogenous group of patients (as a result patients in clinical trials are often younger, and suffering fewer comorbid illnesses). Both of these mechanisms for improving internal validity, however, reduce external validity. Every imposed experimental condition removes the participant from their normal environment, making it difficult to infer the effect of the treatment in 'routine practice'. And, because once the drug is on the market, clinicians will want to treat the full range of patients who suffer the condition, the narrow inclusion criteria of many trials raises the difficult question of whether patients excluded from participating in the trial will respond in the same manner as those included. How to best ensure the external validity of clinical research is recognised, by proponents and critics alike, as one of the most important issues for EBM to address (Black 1996; Rothwell 2005; Upsur 2005).

Concerns about internal validity are concerns about trial methodology. To say a trial has high internal validity is to say that the trial employed methods to prevent a range of errors in inference that are known to happen when observing data on the effect of a therapy in a population. Randomising experimental groups in a prospective interventional study is one such method. So too is conducting an interventional study rather than an observational study when wanting to establish the efficacy of a treatment. EBM's hierarchy organises the study designs commonly used in clinical research according to internal validity.

There are many methodological techniques that are important to the internal validity of a clinical trial. While many of these methods are not included in the presentation of the hierarchy, it is clear they are very important to proponents of EBM. Examples include: adequate concealment of patient allocation to experimental groups (that is, maintenance of double blinding); proper treatment and analysis of ‘drop outs’ (trial participants who leave the study); the proper use and interpretation of statistical tests; and many many more. A cursory glance at any clinical epidemiological textbook, including the one written by the leading proponents of EBM, Haynes et al. (2006), is enough to confirm this; they are full of such methodological concerns.

Comparative internal validity provides a justification for EBM’s entire hierarchy, not just the distinction between randomised interventional studies and observational studies. For instance, observational cohort studies are placed higher on EBM’s hierarchy than case-control studies because cohort studies have higher internal validity (see the hierarchy provided by Phillips et al. (2008)). Case-control studies rely on investigators to define a control group that have not suffered the outcome under investigation. In order to ensure the control group is appropriately comparable to the ‘cases’ a number of assumptions about exposure to the risk factor under investigation are required. Prospective cohort studies, by contrast, can follow a more natural group of patients, some exposed to the risk factor under investigation, and others not exposed. Investigators do not have to construct a control group as they do in case-control studies.[13] This means there is more opportunity for error in case-control studies, and hence, these studies possess lower internal validity.

[13] There are variants of the case-control design that overcome this problem, for instance: nested case-control studies. Nested case-control studies use cohort studies to define the control group.

There are a number of important points to make regarding comparative internal validity as an argument for EBM's hierarchy of evidence. First, the judgements of internal validity incorporated into EBM's hierarchy are specific to the clinical sciences. The methods to improve internal validity are canonical. Mostly, they have been collected through the experience of observing and testing the effects of drugs. Each method attempts to rule out, or minimise, a particular kind of error. Randomisation, blinding, prospective trials, intention-to-treat analyses, and so on, are all methods for preventing specific erroneous inferences in therapeutic trials. While these methods have been built up through experience, this does not mean that a philosophical account referring to a logic of evidence can not be provided. But any such account will be specific to the clinical sciences.

While there is no one factor that makes the clinical sciences unique, the particular confluence of factors that make up the clinical sciences *is* unique. For starters any account needs to recognise the high degree of unexplained inter-patient variability in response to therapy; theory does not adequately predict response in real-life patients. (Arguably, this is what makes statistical approaches so important—and controversial—in the clinical sciences.) In addition, a range of practical considerations of particular importance to the clinical sciences needs to be recognised. For instance, the ability to conduct randomised trials (something impossible in many contexts), and the importance of health, and an appropriate conservatism toward risk when dealing with matters of health. Each of these factors (and many more besides) impinge not only on the kinds of error that can occur, but also the kinds of error that matter, and thus on the kind of methods that have been developed to avoid these errors. Identifying EBM's hierarchy with the internal validity of therapeutic trials not only provides a justification for the hierarchy, but also emphasises the problems of extending the hierarchy beyond this context.

Once outside of the context of a large drug trial other sources of error may become more important. Or, the method developed to reduce the error in the drug trial may no longer work. Randomisation, as Grossman and Mackenzie (2005, 527–8) show, is a case in point. Simply changing the context of an interventional study from testing a drug to testing, say, a social intervention in schools, can be enough to change randomisation from increasing internal validity to decreasing it. Large drug trials have many participants, and hence many 'units of analysis' that can be randomised. In such a situation, randomisation is a convenient way to both take treatment allocation out of the hands of investigators, and ensure the experimental groups are roughly equally balanced for known confounding factors. A trial of a social intervention in ten schools, with five schools receiving the intervention and five receiving control, might have as many pupils involved as the drug trial has participants, but because the intervention is school-wide it has far fewer units of analysis. With such a small effective sample size randomisation is much less likely to ensure the experimental groups are roughly equally balanced compared to if the investigators deliberately balanced the groups according to which factors are considered important. Providing investigators can justify their allocation, and ensure selection bias has been minimised, deliberate matching will result in higher internal validity in the test of a social intervention in schools than randomisation. Similar accounts can be given for the other methods of improving internal validity in

EBM’s hierarchy of evidence; care needs to be taken to ensure the methods that are being employed are pertinent to the case at hand.

It is also essential to be clear what the ‘all other things being equal’ part of this argument means for EBM’s hierarchy. The guidebooks advise that if a study is not randomised, ‘bin it’. Viewing the hierarchy as a hierarchy of internal validity, however, requires considerably more nuanced judgements. The hierarchy only provides increasing internal validity when *all other things are equal*. Study designs higher up the hierarchy rule out more possible sources of error. But at each level of the hierarchy there are many sources of error, and many methods that can, and should, be applied to ensure the results of the study are as reliable as possible. It is only when all the additional measures that improve internal validity have been taken that there is an assurance that a randomised interventional study possesses higher internal validity than a prospective cohort study. As should be obvious, there is no assurance that a randomised interventional study that has ruled out patient-selection bias, but left another source of error unchecked, will be any more reliable than a carefully conducted cohort study, which has employed every method possible to ensure its results are valid. The quality of evidence that a study provides always requires a judgement of whether all reasonable sources of error have either been ruled out or accounted for. Study designs higher up EBM’s hierarchy are able to employ more methods to reduce potential sources of error, but this by no means ensures the quality of evidence that is provided by any particular study using a design listed high in the hierarchy.

Furthermore, the increased potential for high internal validity of randomised studies does not ensure *all* the results of a well-conducted randomised trial are reliable. Indeed, the statistical techniques employed in analysing randomised trials provide optimal warrant only to the primary hypothesis under test. Subgroup analyses and findings on secondary endpoints are less reliable—at least to the extent that these outcomes are underpowered. This is important to keep in mind when considering the primary hypotheses that most clinical trials are set up to test. Primary hypothesis tests in clinical trials almost invariably relate to the *benefits* of a therapy; questions related to a drugs safety are typically relegated to secondary endpoints. While there is no theoretical reason why randomised trials can't be set up to test the safety of a drug as a primary hypothesis, the practical constraints are considerable. First, there is the ethical question of whether it is appropriate to conduct a trial when the primary purpose of the trial is to detect adverse effects. Second, even if such trials are considered ethical, the problem of whether patients would consent to be included in such a trial remains. And third, even if these issues can be overcome, there is a certain degree of inertia in the system. The medical fraternity (as much as it acts as a single unit), the regulatory authorities who are responsible for outlining what research needs to be completed before a therapy will be permitted onto the market, and the pharmaceutical industry which funds the vast majority of large drug trials, currently hold that trials that test primary 'benefit' hypotheses are sufficient to 'prove' the drug is ready for the market. The proportion of trials that incorporate safety endpoints into the primary hypothesis under test is unlikely to increase while this view is pervasive.

Of course, concern about the reliability of analysis of secondary endpoints and subgroups is not new. Meta-analysis is the standard reply to the problem of providing reliable estimates for secondary endpoints and subgroup analyses. When a number of relevantly similar randomised trials have been conducted they can be combined in order to conduct a meta-analysis. Such meta-analyses will provide more reliable results for these secondary outcomes. EBM recognises this by placing meta-analyses and other systematic reviews of a number of randomised interventional trials at the pinnacle of the hierarchy. But this option is only available when multiple trials have been conducted, and will only improve the reliability of the estimates if the trials are similar enough in the relevant respects; needless to say this is not always the case.[14]

These considerations substantially restrict the claims that can be made by EBM's hierarchy on the basis of considerations of internal validity. The importance of restricting the claims of EBM to therapeutic questions and the results of primary hypothesis tests and meta-analyses is emphasised. This avoids much of the rhetoric sometimes employed by proponents of EBM. On this view EBM's hierarchy of evidence does not provide general epistemological rules, nor are randomised interventional studies infallible, nor do randomised studies carry *very special* scientific weight when compared to observational studies. The comparative benefits of randomised interventional studies can be put simply: randomised studies can employ more methods that improve internal validity than observational studies. The quality of evidence that any particular study provides is a separate judgement taking much more into consideration.

[14] See for example Egger and Smith (1995), Egger et al. (1997), and Smith and Egger (1998).

3 Evidence based medicine must be

I have argued that EBM's hierarchy of evidence is best interpreted as a hierarchy of comparative internal validity. Furthermore, I have shown that identifying EBM's hierarchy with comparative increases in internal validity emphasises that EBM's claims should be limited to the results of primary hypothesis tests and meta-analyses. In the introduction, I suggested that any justification of EBM's hierarchy of evidence needs to provide (i) a clear interpretation of the hierarchy (that is, how it should be applied), and (ii) a justification for why applying the hierarchy in this way more reliably informs therapeutic decisions. I now examine how well identifying EBM's hierarchy with comparative increases in internal validity fulfils these requirements.

Applying the hierarchy viewed as a hierarchy of internal validity is considerably more complicated than the simple advice provided by the EBM guidebooks on applying the hierarchy. When interpreting EBM's hierarchy as a hierarchy of internal validity the type of question under consideration is vital. Say, *Drug X* has passed the early stages of clinical testing. The drug appears safe, and has a promising pharmacological profile. If the question is whether *Drug X* is *efficacious* in a defined population of patients, then EBM's hierarchy provides clear advice. The most accurate method to measure the efficacy of *Drug X*, all other things being equal, is via a well-conducted randomised controlled trial. While this methodology is fallible, it rules out more potential sources of error than methodologies lower down EBM's hierarchy.

Things become more complicated if the question changes to whether *Drug X* possesses a particular side effect. Note, this question just asks whether *Drug X* possesses the side effect, not whether it possesses the side effect in any particular population; that is, what is being considered is the side effect equivalent of 'efficacy'. It is still possible to suggest the 'best' methodology to address this question is a well-conducted randomised trial, but for this to be so, some hefty assumptions are required. First, the side effect has to be suspected in order to set up a trial to test the hypothesis that the side effect exists. (Clearly, this is not always so in practice). And second, you have to be able to conduct the randomised trial. While there is no theoretical impediment to this, the practical constraints discussed in the previous section mean that such a trial is less likely to be conducted.

Relying on a randomised trial set up to test a primary 'benefit' hypothesis to detect adverse effects creates a number of problems. Not only because safety outcomes are relegated to secondary endpoints, but also because a trial set up to test a benefit hypothesis will select patients who have less complex medical histories than the population of patients with will eventually receive the drug. While selecting patients with only the condition under investigation makes sense when testing a drug's efficacy, this distorts both the detection and estimation of side effects more likely to occur in patients with multiple pathology. Randomised trials are the 'best' way of confirming the possible side effect of *Drug X* in only the most theoretical sense—the randomised trials typically conducted in practice will *not* provide the most reliable evidence relating to the drug's side effects.[15] Judging whether the side effect of *Drug X* exists will instead rely on: the serendipitous findings of randomised trials set up to test benefit hypotheses; the accumulation of randomised trial evidence in meta-analyses; or, evidence from lower down EBM's hierarchy—or, on each of these if they are available.

When the standpoint of the question is changed to that of a clinician regarding a marketed drug, the same questions can be addressed. But here the limitations on the claims substantiated by EBM's hierarchy are even more important. The clinician's most pressing question is whether the drug's benefits will outweigh its potential for harm in the patient presenting to the clinic. Interpreting EBM's hierarchy as one of comparative internal validity makes it explicit that reference to the hierarchy provides only *partial* assistance in answering this question. All things being equal, estimations of the drug's efficacy are more likely to be accurate in well-conducted randomised studies. Questions regarding the drug's potential for harm are more difficult. Because safety endpoints are likely to be secondary endpoints (and under-powered) in randomised trials, the information these trials provide can be less reliable. Evidence from observational studies may be more helpful in this case, but here a judgement on the potential for any selection bias needs to be made. Crucially, EBM's hierarchy (interpreted as a hierarchy of increasing internal validity) provides some assistance, but it does not answer the *effectiveness* question asked by the clinician. These are questions of

[15] This assumes the incidence of the side effect is lower than the degree of benefit afforded by the drug—as one would hope. If this is not so, and the ill effects of the treatment are as large as the benefits, then the randomised trials conducted to confirm the benefits will be large enough to reliably detect the harms.

external validity, and they are additional to the questions addressed by EBM's hierarchy.

4 Conclusion

EBM's hierarchy is best understood as a hierarchy of comparative internal validity. The constraints this justification places on the EBM's hierarchy are too often under-appreciated (or at least under-emphasised) in the clinical literature discussing EBM.

First, the increased validity is only substantiated in a small subset of highly specific questions; its strongest claims are made for questions relating to a drug's efficacy. Even extending the question to the same drug's side effect profile raises considerable challenges.

Second, the justification is comparative, not absolute. In the situations in which it is obtained the greater opportunity for internal validity does not ensure infallibility. And the incremental benefits of being able to employ additional methods for ruling out certain types of error will only be realised if other appropriate methods to rule out or reduce sources of error are utilised. Far from placing very special weight in randomised trials, this justification emphasises the need for careful judgement on all sources of error as well as on the methods utilised to reduce them.

Third, this justification makes clear that high internal validity is not sufficient for reliably informing therapeutic decisions. Interpreted as a hierarchy of comparative internal validity, EBM's hierarchy of evidence provides a framework for developing arguments about evidence and the application of evidence to therapeutic decisions. The details of the question and the case at hand always matter. Interpreting EBM's hierarchy as one of comparative internal validity makes the challenge of external validity explicit. Being clear as to what questions are well answered by EBM's hierarchy assist framing the additional questions that need to be addressed to inform the therapeutic decision.

References

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

Black, N. 1996. Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal*, 312(7040):1215–1218.

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.

Chalmers, T. C., R. J. Matta, H. Smith, Jr., and A. M. Kunzler, 1977. Evidence favouring the use of anticoagulants in the hospital phase of acute myocardial infarction. *New England Journal of Medcine*, 297:1091–7.

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

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

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

Egger, M., Smith, G. D., Schneider, M., and Minder, C. 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.

Fletcher, R. H., Fletcher, S. W., and Wagner, E. H. 1996. *Clinical Epidemiology: The Essentials*. 3rd ed. Lippincott Williams and Wilkins: Baltimore.

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

Haynes, R. B., Sackett, D. L., Guyatt, G. H., and Tugwell, P. 2006. *Clinical Epidemiology: How to Do Clinical Practice Research,* 3rd ed. Lippincott Williams and Wilkins.

Kendall, M., Stuart, A., and Ord, J. K. 1983. *The Advanced Theory of Statistics*, volume 3, 4th ed. Charles Griffin and Company Limited.

Phillips, B., et al. 2001. Oxford Centre for Evidence-Based Medicine Levels of Evidence (May 2001).URL: http://www.cebm.net/?o=1023 Accessed July 2008.

Rothwell, P. M. 2005. External validity of randomised controlled trials: "To whom do the results of this trial apply?" *Lancet*, 365(9453):82–93.

Sackett, D. L. 2006. The principles behind the tactics of performing therapeutic trials. In Haynes, R. B., et al. (editors). *Clinical Epidemiology: How to Do Clinical Practice Research,* 3rd ed. Lippincott Williams and Wilkins, chapter Six, pages 173–243.

Sacks, H., Chalmers, T. C., and Smith, Jr., H. 1982. Randomized versus historical controls for clinical trials. *American Journal of Medicine*, 72(2):233–240.

Smith, G. and Egger, M. 1998. Incommunicable knowledge? Interpreting and applying the results of clinical trials and meta-analyses. *Journal of Clinical Epidemiology*, 51(4):289–95.

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

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.

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

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

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