Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1356-1294

Evidence: philosophy of science meets medicine

John Worrall PhD

Professor of Philosophy of Science, Department of Philosophy, Logic & Scientific Method, London School of Economics, London, UK

Keywords
controlled trial, evidence-based medicine, external validity, logic of evidence, philosophy of science, severe test

Correspondence
Professor John Worrall
Department of Philosophy
Logic & Scientific Method
London School of Economics
London WC2A 2AE
UK
E-mail: J.Worrall@lse.ac.uk

Accepted for publication: 19 November 2009

doi:10.1111/j.1365-2753.2010.01400.x

Abstract

Obviously medicine should be evidence-based. The issues lie in the details: what exactly counts as evidence? Do certain kinds of evidence carry more weight than others? (And if so why?) And how exactly should medicine be based on evidence? When it comes to these details, the evidence-based medicine (EBM) movement has got itself into a mess – or so it will be argued. In order to start to resolve this mess, we need to go 'back to basics'; and that means turning to the philosophy of science. The theory of evidence, or rather the logic of the interrelations between theory and evidence, has always been central to the philosophy of science – sometimes under the alias of the 'theory of confirmation'. When taken together with a little philosophical commonsense, this logic can help us move towards a position on evidence in medicine that is more sophisticated and defensible than anything that EBM has been able so far to supply.

> A wise man proportions his belief to the evidence. (David Hume, *Enquiry Concerning Human Understanding*)
>
> Any belief that the controlled trial is the only way would mean not that the pendulum had swung too far, but that it had come right off the hook. (Austin Bradford Hill, *Reflections on the Controlled Trial*)

Following Hume (and indeed, one would hope, commonsense), it is surely axiomatic that medicine, like any rational pursuit, should be based on evidence. What else should it be based on? Myth? Superstition? The Delphic Oracle?

It isn't a question of *if* medicine is – or should be – evidence-based. The issues lie in the details: *what* exactly counts as evidence? Do certain kinds of evidence carry more weight than others? (And if so why?) And *how* exactly should medicine be based on evidence?

When it comes to these details, the evidence-based medicine (EBM) movement has got itself into a mess – or so it will be argued. In order to start to resolve this mess, we need to go 'back to basics'; and that means turning to the philosophy of science. The theory of evidence, or rather the logic of the interrelations between theory and evidence, has always been central to the philosophy of science – sometimes under the alias of the 'theory of confirmation'. When taken together with a little philosophical commonsense, this logic can help us move towards a position on evidence in medicine that is more sophisticated and defensible than anything that EBM has been able so far to supply.

1. The current situation in evidence-based medicine

Evidence-based medicine was originally understood by many as advocating the very strong view that the *only* evidence worth its name in medicine is that supplied by a properly controlled randomized clinical trial (RCT). Amid reminders of long-lived therapeutic fiascos in the history of medicine (such as blood letting), and of Cochrane's stern warning that this may not be a purely historical phenomenon but instead many of the therapies accepted today might have equally flimsy evidential foundation, EBM seemed to be telling us that we could trust neither so-called clinical expertise, nor historical experience nor 'pathophysiologic rationale'. The only scientific evidence in medicine comes from clinical research and the only clinical research that provides truly reliable evidence for the efficacy of proposed therapies is the RCT. EBMers were soon to insist that this was a misinterpretation, but if so then it was one that they themselves at least sometimes encouraged occasions: for example, the first and second editions of the EBM 'Bible' instructs an evidence-based practitioner facing a therapeutic decision to comb the clinical trials literature for evidence relative to her problem but if she finds an apparently relevant study and 'the study was not randomized we'd suggest that you stop reading it and go on to the next article in your search' [1].

Right from the beginning, of course, more measured voices pointed out that, while the emphasis on evidence has to be correct,

RCTs, whatever their virtues, surely couldn't be the evidential be all and end all [2]. Other forms of trials and investigations surely *may* provide reliable evidence – no doubt EBM-ers could produce a (relatively) few cases (grommets for glue ear was always a favourite example) of contemporary accepted treatments that turned out to be ineffective when subjected to the rigorous scrutiny of an RCT; but, so these more measured voices said, this can't, *pace* Cochrane, be true of many long accepted treatments – such as thyroxine for myxoedema, insulin for diabetic ketoacidosis, vitamin B_{12} for pernicious anaemia, appendicectomy for acute appendicitis, etc. Surely we have extremely solid evidence for the (overall) effectiveness of these and many other treatments even though they long predate the introduction of RCT methodology into medicine?

Perhaps because they really had never intended the strong claim, or perhaps in response to criticism, EBM-ers started to develop and endorse a much more inclusive and nuanced view. Other types of trial could supply *some* – though invariably less powerful – evidence, 'pathophysiologic rationale' and clinical expertise should be *incorporated* rather than overridden [3]. EBM-ers in fact give precious little advice as to how more exactly this amalgamation of different types of evidence is to be carried out [4]. But it is absolutely clear that, even on this modified view, RCTs retain a very special epistemic role – to the extent that, at least according to some versions, the result of one well-performed RCT 'trumps' any number of observational studies yielding contrary results no matter how many or how large those observational studies may be [5]; and to the extent that according to *all* versions, other evidence should be sought only if RCT evidence is missing, whereupon one is inevitably trading in 'second best' evidence [3].

Evidence-based medicine began in fact to endorse an *evidence hierarchy* – a ranking of evidence from different sources in terms of the 'quality' of that evidence. One such hierarchy can be found at http://www.sign.ac.uk/guidelines. But it is *only* one: a 2002 study identified no less than 40 such systems of grading evidence [6]; while a 2006 survey found 20 more [7]. Some of these are only notationally distinct, and certainly all share one central feature – the pre-eminence given to evidence from RCTs. Results from double blind RCTs are invariably ranked either top of the hierarchy or equal top with (or occasionally second to) the results of meta-analyses of a number of such RCTs. But alongside this agreement there are also significant differences between some of the hierarchies: for example, some rank meta-analyses of RCTs at the very top of the tree, while others omit meta-analyses altogether; and some rank cohort studies ahead of case control studies, some rank them equal and some put case–control studies ahead of cohort [8].

Alongside the development of evidence hierarchies a number of important concessions have recently been made by those who would certainly regard themselves as firmly within the EBM camp. Most notably there has been for the first time an explicit concession that RCTs are unnecessary in the case of 'large' or 'dramatic' effects (together with a not entirely convincing attempt to characterize what a 'large' effect is) [9].

Moreover there seems to have been a swing in the balance in the medical statistical community between classical frequentists and Bayesians. Not so long ago, Bayesians were a largely neglected minority, but increasingly Bayesian techniques are being regarded as at least worthy of attention and classical statistical significance testing is being more and more questioned (at least among experts if not among the rank and file of medical researchers). And Bayesians, at least those of an orthodox stripe, have never seen any direct role for randomization [10].

Reflecting these changes in views about the impact of evidence, one very influential voice has recently argued that (i) the evidential virtues of randomization have been significantly oversold; (ii) those of 'observational studies' often significantly *under*sold; and (iii) that the whole idea of an evidence hierarchy is a major mistake. This voice belongs to Sir Michael Rawlins, head of the UK National Institute for Health and Clinical Excellence and therefore one of the most, perhaps *the* most, influential medical policy maker in the UK [8]. Rawlins would definitely regard himself and his institution as being in favour of applying proper scientific method in medicine and therefore as evidence-based in the general sense; but he is scathing about 'hard line' EBM's advocacy of an evidence hierarchy.

First, such hierarchies are internally unjustified in that they overrate RCTs:

> The notion that evidence can be reliably placed in hierarchies is illusory. Hierarchies place RCTs on an undeserved pedestal for . . . although the technique has advantages it also has significant disadvantages. Observational studies too have defects but they also have merit.

But in Rawlins' view it is not just the internal detail but the whole idea of a hierarchy that is at fault:

> Hierarchies attempt to replace judgement with an oversimplistic, pseudo-quantitative, assessment of the quality of the available evidence. Decision makers have to incorporate judgements, as part of their appraisal of the evidence, in reaching their conclusions.

But this seems to be 'déjà vu all over again': it was precisely the attempt to eliminate clinical judgement with its allegedly very poor history in terms of the therapies it endorsed and to replace it with objective scientific evidence that formed the initial EBM battle cry [3]. Although so far as I know no one has published the response, there will no doubt be EBM-ers who regard Rawlins' view as constituting the abandonment of any evidence-based approach worthy of the name.

In essence then the position that many took EBM to start with – that the only real evidence of therapeutic effectiveness is that provided by the result of an RCT – was nice and crisp but never remotely defensible. Now it is not clear what precise position EBM holds on evidence in medicine. A number of welcome moves away from the initial position have been made but no overall coherent position seems to be on offer. My proposal in the final two sections of this paper is to show that progress can be made towards such a position by returning to 'basics' – that is to the general logic of evidence.

2. Why *control* at all? And why *randomized* controls?

Philosophy of science might seem a surprising source for the resolution of conflicts about evidence in medicine (or elsewhere). Unanimity is *not* what philosophy of science does well; and it might be thought that invoking it, far from resolving the conflicts, would only heighten them. In fact, however, we can get surprisingly far with a single insight that is shared by all serious approaches to confirmation.

Take, for illustration, Popper's version of hypothetico-deductivism on the one hand [11], and Bayesian confirmation theory on the other [10]. In many detailed respects, these two accounts could hardly be further apart, and yet they completely agree that in order for evidence to count at all strongly in favour of a theory it must not only be accounted for by the theory, it must *also* be 'otherwise improbable'. In Popper's system this translates into the claim that a theory is confirmed only by *tests*, and the theory is the more confirmed by a positive test result, the 'more severe' the test is. A test is *not* severe if background knowledge predicts the same result as the theory under test. Similarly the 'Bayes factor' which measures the amount by which a theory's probability is increased when some piece of evidence e is established is the product of two terms, one of which is 1/p(e,B): the more likely the result already was in the light of background knowledge B, the smaller the confirmation and in the limit where the result e already deductively follows from background knowledge, so that p(e,B) = 1, there is no confirmation at all. And clearly if there are alternatives to a theory T that look plausible in the light of background knowledge then those alternatives will have fairly high priors; and if moreover those alternatives entail e then e must itself of course have a prior equal to or greater than any such alternative. Hence again e can supply at best little support to T.

Nor is this basic idea confined to Popperianism and Bayesianism. It is implicit in John Stuart Mill's Logic and takes centre stage in, for example, Deborah Mayo's more recent 'error statistical' approach [12].

This simple and agreed evidential principle supplies the basis for the whole idea of *controlling* a trial. If, in the hackneyed illustration, a bunch of people with colds were given vitamin C and all their colds cleared up within a week, that would lend little if any support to the idea that vitamin C cures colds. Background knowledge already tells us that most colds clear up within a week if left untreated, and so this is not any real test of the vitamin C hypothesis. Instead we need a control group of people with colds who are not given vitamin C. But any old control group is not enough of course, as our basic confirmation theory principle again tells us. Suppose more people recover in the treatment group. If those in that group are younger, fitter, have fewer comorbidities . . . than those in the control group, then a 'positive' result would not really be positive – because background knowledge provides us with many plausible explanations of the better recovery rate to rival the theory that the administration of vitamin C was the cause.

So this simple principle tells us not only that controls are needed, but also that, even with controls, no real evidence may accrue from a better average outcome in treatment group if the two groups are 'inequivalent' – where, that is, there is a difference between the two groups aside from the treatment that can plausibly (in the light of background knowledge) account for the outcome. This already suggests a much greater emphasis on effect size than is standard in the clinical trials literature, where statistical 'significance' is king. (It is this simple insight which underwrites the – only very recent – EBM concession about RCTs being unnecessary for 'dramatic' effects.)

But leave this to one side. It is clear that we could, at least in principle, guarantee by deliberate matching that the two groups were equivalent in respect of any factors that could be argued to be plausible possible 'confounders' in the light of background knowledge: so we could (again in principle – in practice it may be enormously, perhaps dauntingly, complex and time-consuming) make sure that the age distribution, sex distribution, significant co-morbidity distribution, etc. were the same in the treatment and control groups.

However, two problems still seem inescapable. *First*, who says what background knowledge consists of and what possible confounders it renders 'plausible'? Isn't this allowing judgement to play a role in what ought to be an entirely objective process of evaluating evidence? *Second*, whatever we make of 'background knowledge', it is inevitably incomplete – even when we have matched our trial groups as closely as possible with respect to so-called '*known* confounders' (really factors that plausibly might play a role in outcome according to background knowledge) it is always possible, trivially, that the two groups are unmatched with respect to some factor that background knowledge as it stands gives us no reason to think might play a role in outcome but which in fact does. This is the problem of so-called 'unknown confounders' (again really *unsuspected* confounders).

It is then of course very easy to see from this perspective the chief appeal of RCTs: they, or so their defenders often allege, control all at once for all possible confounders – known and unknown; and hence at a stroke *both* do away with any reliance on judgement about background knowledge *and* eliminate the worry about its incompleteness.

Before analysing this central argument for RCTs, let's reflect for a moment on the first problem with the idea of deliberately matching the trial groups – its reliance on 'judgement' in the form of what background knowledge does or does not make plausible. It is important, I think, to realize that acknowledging such a role for judgement would not at all mean that we have stepped outside of regular scientific method. Even in physics, it is well recognized that scientists always face what is sometimes called the 'Duhem Problem': the need to invoke auxiliary assumptions in any test of a theory [13]. Suppose that Newton's theory of gravity is being tested by the celebrated Cavendish torsion experiment. In order to deduce a definite outcome we need to assume that the total force acting on the two bodies is, at least to a very good approximation, their gravitational interaction (matched by the torsion in the wire holding the moveable body). Other possible forces need therefore to be eliminated or at least sharply minimized. Background knowledge tells us what sorts of other forces there might be – electric and magnetic forces, for example, and how to ensure that they are eliminated. And hence background knowledge centrally informs the test. Nor is this judgement in any sense mystical or unanalysable. Not in physics, and not in medicine either. In physics, we know if bodies are charged then they will be subject to electric and magnetic forces alongside the gravitational force; and in medicine, we know that treatments that work well in the young have proved harmful in the elderly, that treatments that have no major side-effects in males may have such side-effects in females, that treatments for condition C may be fine in those patients for which C′ is not also present, deadly for those with that co-morbidity. In both cases what background knowledge gives us good reason to believe is important, a matter of 'judgement' if you like, but certainly not a matter of *unanalysable* judgement.

But now let's focus on the major 'gold standard' argument for RCTs – that by randomizing you bypass background knowledge and any possible judgement altogether *and* solve the problem of possible 'unknown confounders' by ensuring the comparability of

the experimental and control groups. Thus if you log on to the UK Cochrane Centre web site you will be told by its Director Mike Clarke that 'In a randomized trial, the *only* difference between the two groups being compared is that of most interest: the intervention under investigation' [14].

Even Michael Rawlins [8], aware as he is of the limitations of RCTs, accepts that they do have this one great advantage:

> The greatest strength of an RCT is that the allocation of the treatments is random so that the groups being compared are similar for baseline factors.

It is not clear to me that the 'similarity' claim (let alone the identity claim implicit in Mike Clarke's remark) makes any sense once we are talking about *all possible* extraneous factors that might play a role. But in any event, the claim, assuming it makes sense at all, is simply false (as everyone including Clarke and Rawlins *really* knows).

No one really believes that, given a *particular* random division, the groups are bound to be equal in all other respects and hence any difference in the outcome is automatically attributable to the difference in treatment. That is, no one (not even Mike Clarke, as he indicates later in his article [14]) really believes that having randomized is a *sufficient* condition for establishing that any observed effect must be due to the treatment and to the treatment alone. In any *particular* randomized division, it is of course entirely possible that some factor is unbalanced between the two groups.

This is in fact quietly conceded by orthodox RCT methodology. At least in those trials (the majority) where no attempt has been made to match with respect to known prognostic factors, trialists are recommended to look at the particular division into control and experimental groups that randomization has given them and check for 'baseline imbalances'. That is, trialists should check that the two groups are not in fact unbalanced with respect to some factor that background knowledge tells us might play a causal role. If such baseline imbalances are found then the recommendation – clearly for practical rather than epistemic reasons – is to re-randomize in the hope that this time no baseline imbalances will occur.

But if it is admitted, as of course it must be, that an imbalance in 'known' factors is possible, then it must equally be acknowledged that there may be an imbalance in 'unknown' confounders, factors which do in fact play a role but which background knowledge supplies no reason to have suspect do so. The difference between this and the (known) 'baseline imbalance' case is of course that, by definition, trialists cannot check for imbalance in 'unknown' confounders.

An amusing example that demonstrates that randomization cannot guarantee that the two groups are equal in all relevant respects is provided by an article published in the *British Medical Journal* in 2001 by Leibovici [15]. This study identified 3393 patients who had a bloodstream infection of some sort while inpatients at the Rabin Medical Centre during 1990–1996. In July 2000 (so *at least four years after* these patients had been in hospital), a random number generator was used to divide them into two groups. Which of these two became the treatment group was decided by a coin toss. 1691 were randomized to the intervention group and 1702 to the control. A check was actually made in this case for 'baseline imbalances' with regard to main risk factors for death and severity of illness – that is, whether this pure randomized division without any prior matching had in fact produced groups that were significantly unbalanced with respect to 'known confounders'. None having been found, the names of those in the intervention group were given to a person 'who said a short prayer for the well being and full recovery of the group as a whole'.

Mortality, length of stay in hospital and duration of fever were then recorded from the hospital notes and compared in the two groups. The results were as follows. Mortality was 28.1% in intervention group and 30.2% in the control group; this was 'not significant' according to the usual significance testing/null hypothesis methodology. However, *both* length of stay in hospital *and* duration of fever were significantly shorter in the intervention group ($P = 0.01$ and $P = 0.04$)! Leibovici concluded – perfectly properly in accordance with accepted methodology in medicine – that:

> Remote, retroactive intercessory prayer said for a group is associated with a shorter stay in hospital and shorter duration of fever in patients with bloodstream infection and should be considered for use in clinical practice.

It should go without saying this is not to be taken seriously: even those who believe that God moves in mysterious ways are hardly likely to believe that they are mysterious as this! The reason why it is not to be taken seriously reveals a further aspect of this simple but striking case: that we are all naturally Bayesian. As Leibovici himself wrote [16]:

> If the pre-trial probability is infinitesimally low, the results of the trial will not really change it, and the trial should not be performed. This, to my mind, turns the article into a non-study, though the details provided (randomization done only once, statement of a prayer, analysis, etc.) are correct.

One simple lesson, then, is that Fisher's insistence on not bringing any prior information into the assessment of the impact of a stochastic experiment in order to guarantee objectivity was an understandable, but grievous and enormously deleterious error. Because she has very good reason to believe that 'remote intercessory prayer' *cannot* be effective, the way a sensible person reacts to this result is justifiably different from the case in which no such prior information is available. If only this lesson really went home within medicine! (Of course Bayesians don't help by calling the prior *information* 'subjective'!)

So far as its impact on the claim that RCTs are sufficient to establish causality, notice that Leibovici's study was fully and properly randomized (and indeed very large, $n > 3000$). Moreover there was no 'data mining' or the like. The fact that its results cannot be taken seriously means that there *must* have been some imbalances between the two groups, no doubt a whole series of independent ones that together account for the observed differences – though clearly these were imbalances in unknown factors, because, as Leibovici notes, the two groups had been checked for ('known') 'baseline imbalances'.

However, and before I'm drowned by cries of 'straw man', no one believes, do they?, that randomization inevitably guarantees similar groups and hence that a positive result in a properly randomized trial is *sufficient* for a treatment to be declared effective. Well actually I think lots of people in medicine *do* believe this, because this is what they think they are being told be the experts. And as we saw, many people, Mike Clarke included, certainly sometimes say that they believe it [14]. But I agree that it cannot be seriously believed. In so far as anyone seriously believes that

there is some guarantee of equivalence or similarity between the two groups in a randomized trial it is not belief in a sure-fire guarantee but rather in some sort of probabilistic quasi-guarantee – clearly, in what is admitted on all sides to be a stochastic domain, we could not reasonably expect any better.

But what exactly does such a 'probabilistic guarantee' amount to? Surprisingly many people take what seems to me a surprising amount of solace in the phrase 'either the groups are equal or a chance event has occurred'. But in an area where it is acknowledged that we do not have control over all the factors that might play a role, then a chance event has always occurred – it just as 'chancy' if randomization produces equal groups as if it doesn't! This often repeated mantra seems to make the mistake of assuming that if a fair coin is tossed 10 times and nine heads result then this is a 'chance event' while if it produces five heads it is not.

Again looking at it from the perspective of basic philosophy of science, there seem to me to be two reasons to be suspicious of the credentials of any such 'probabilistic quasi guarantee'.

The first and more significant is that the reasoning underpinning these 'credentials' involves a slip from what is arguably true in the indefinite long run to a claim about what is true of a *particular random allocation*. An enormous amount of effort in the philosophy of science literature has gone into the attempt to make sense of single case probabilities on an objective view of probability. (This is in distinction to the 'subjective' Bayesian view of probabilities as degrees of belief for which the 'single case' presents no problem.) My own view is that no real sense can be made of the notion of single case objective probabilities. I cannot hope to argue this here; but I can indicate one central difficulty in supposing that there are such single case probabilities. The only sustainable objectivist view seems to be the frequency interpretation. But then the claim that there is a high probability that the experimental and control groups are balanced with respect to some particular factor really amounts to the claim that if one were to take some group and divide them into two by some random procedure and if one were then to randomize again and then again . . . keeping a cumulative total for the relative frequencies of patients exhibiting this factor in the two groups (and forgetting about the fact that these different trials would not be independent!) then in the indefinite long run the limiting frequency of this factor within both the experimental and control group would be the same and would be the same as the frequency with which that factor is exhibited in the experimental population as a whole. But we are never in the long run, we never randomize indefinitely often, medical researchers only randomize *once*. And in that one random allocation, the two groups can be as unbalanced with respect to the factor at issue as you like – as the Leibovici study establishes, but which is in any event obvious.

The second problem with the reasoning behind this probabilistic quasi-guarantee was pointed out by the Bayesian statistician Dennis Lindley. There is more than a hint of a quantifier fallacy here [17]. There seems to be some confusion between imbalance with respect to a *particular* factor and an *overall* imbalance. Even if one were to try to make some single-case probability argument work, it would be an argument that there is a very low probability that *some particular* factor (say age, or co-morbidity) is unbalanced between the two groups. But how is one to weigh this against the assumption driving this whole issue that the list of possible unknown factors is indefinite? In such circumstances, it seems that even if we suppose that there is a definite probability that the groups are unbalanced with respect to some particular specified factor, the 'probability that the groups are unbalanced with respect to *some* possibly confounding factor' is unquantifiable [10].

This argument for the special epistemic power of randomization – that by randomizing all confounders are controlled for at one stroke – is the one that has carried most weight within the medical community; and it at least is, to say the least, not clearly valid. When looked at from this more basic philosophy of science perspective, the other arguments for the special epistemic power of randomized trials also appear questionable [4]. The exception is the argument that randomization controls for one *very specific* possible confounder – selection bias, where this is understood not in the very general sense in which it is sometimes used but specifically as bias introduced when clinicians are able to choose the group to which a particular patient is assigned. However:

1 This control is not produced by the randomization itself but by the blinding – randomization is one way to take the powers of selection away from the clinicians and clearly *not the only such way*; and

2 As even the most ardent of randomizers concede, selection bias can often fairly readily be reduced to a level where it cannot plausibly be thought to have a large or even moderate effect. So only the smallest of effects are likely to be obscured by this bias and it is at least arguable that once side-effects are taken into consideration, those effects are not worth having. [18]

Nothing in these arguments is 'anti-RCT' (though they are anti the now largely discredited view that an RCT is a sine qua non of really scientific evidence). Still less is anything against applying scientific method in medicine. Instead, these arguments are aimed exactly at developing a properly scientific approach to evidence in medicine from a more basic philosophy of science perspective. That approach then allows an assessment of what RCTs can and can't do. One main consequence is a more optimistic view of the epistemic power of non-randomized studies [4]. This is also now reflected in the views of Michael Rawlins and the same message will independently arise from the final section.

3. What is evidence, evidence *for*?

In section 2 a simple principle about theory testing was shown to shed light on issues about controls and evidential weight. A principle of philosophy of science (or really of educated commonsense) that is even more simple will take centre stage in this section.

Once we stray outside the rarefied atmosphere of theoretical physics where the theories to be tested are generally clear and into more complex, more 'empirical' domains, it is easy to fail to ask what *exact* theory we are looking for evidence for; and if we do so fail, then we are unlikely to have a sensible view of the import of whatever evidence we collect. This sounds (and is) entirely trivial and yet it has powerful practical implications – especially for medicine as we shall see.

Standardly, research reports in the medical journals will have titles like (these are taken from a recent edition of the *Lancet*) 'Efficacy and safety of ustekinumab . . . in patients with psoriasis . . .' or 'Active symptom control with or without chemotherapy in the treatment of patients with malignant pleural mesothelioma . . .' [19]. They will then report (usually randomized) trials on some

selected group of patients – where the selection involves a number of exclusion criteria (often over 65s will be excluded, so will those exhibiting risk factors for various conditions, those exhibiting certain co-morbidities and so on), generally using some *very precise treatment regimen* which the trialists are not allowed to alter or adjust, where the treatment is given for some *relatively brief period* (as Rawlins reports [8]: 'Most RCTs, even for interventions that are likely to be used by patients for many years, are of only six to 24 months duration.'). And the study will report that administration of substance S is (or is not) effective – meaning more effective than the treatment given to the control group (often placebo, sometimes the currently accepted treatment for the condition at hand).

So, assume that the trial outcome is positive, and that the trial is a pharmaceutical one testing substance S as treatment for condition C. Which exact theory has actually been tested? Not the (dangerously vague) claim that, say, substance S is 'effective' for condition C, but rather the more specific claim that substance S when administered in a very particular way to a very particular set of patients for a particular length of time is more effective than some comparator treatment (often, sadly, placebo). This is the claim that the RCT provides evidence for – let's assume for current purposes impeccable evidence.

But this is not, of course, the claim that the practising doctor would like to have evidence for. She would like to know whether the treatment is effective (in a wide sense that certainly includes factoring in any side-effects whether short or long-term) when prescribed to the sorts of patients she would like to prescribe it to. (She would also like to know whether it is more or less effective than the currently best available treatment for the same condition, *not* whether it is better than placebo.) This 'target population' is not very precisely characterized but will certainly include many types of patient excluded from the trial (the elderly perhaps or those with significant co-morbidity). Moreover, there will be the possibility of adjusting the dose in the light of individual patient's reactions. In the trial, care may be taken that the patient receives the allotted treatment, in the wild patients forget. Finally, if the condition is a chronic one then the doctor may want to prescribe S for a long time – certainly much longer than the trial itself is likely to have lasted.

This is often presented as the problem of 'external validity': does the evidence from the trial 'generalize' to the 'target population' (roughly the set of patients that doctors are likely to prescribe the treatment to if the trial is successful)? However, another lesson that I at least take from philosophy of science is that it is always better to reframe apparently inductive problems in a deductive way; and surely the most straightforward way to view this situation is in terms of the theory we would like to be tested not in fact being the theory that *is* tested by the clinical trial.

Sticking for the moment, however, to the usual formulation in terms of 'external validity', it is important to note that the issue is not what is sometimes dimissively called a 'purely philosophical' one. We are not here asking something on a par with 'does the fact that the sun has always risen in the past give us good grounds for thinking it will tomorrow?' Unlike Hume's case, we know on good specific grounds that the trial population and the target population are different. For example, a study looked at 25 recent RCTs on non-steroidal anti-inflammatory drugs (NSAIDs) and 27 recent RCTs on Statins and found that older people, women and ethnic minorities were (quite significantly) underpresented compared with the general (and therefore also presumably the 'target') population [20]. Moreover not only do we know that there are such differences, background knowledge, largely in the form of previous experience, provides good grounds for thinking that those differences may result in differences in outcome (and no reason to think that such differences will be small).

This is in fact constructively demonstrated by a number of real cases in which a treatment endorsed by an RCT was later withdrawn because of significantly deleterious overall outcome. One case involved Benoxaprofen (Opren) [21]. This was an NSAID developed in the early 1980s for arthritis/musculo-skeletal pain. Its big attraction over other NSAIDs was that it was to be taken only once a day. A big RCT was performed in a trial restricted to 18–65 year olds. The trial had an impressively positive result and Opren was very aggressively marketed and duly cornered the market. It is, however, a fact that the population of people who suffer from arthritis and musculo-skeletal pain has an average age much higher than that of the general population. It turned out that in the elderly (who had not been represented in the trial population) Benaxaprofen has a significantly deleterious effect – causing a significant number of deaths from hepato-renal failure for example. And the drug was duly withdrawn. Rawlins [8] cites a total of 22 drugs that have been approved by RCTs in recent years only to be later withdrawn for safety reasons.

So the trial's 'external validity', or, as I would prefer, the fact that the trial is testing the wrong theory, is a genuine problem. When looked at in my way, the following result seems clearly to follow. Even were we to accept – as I have suggested we should not in fact – that the arguments for the special epistemic power of randomization establish RCT evidence as gold standard evidence for the wrong theory (that treatment T when administered to a specially selected group of patients in a particularly rigid, but high maintenance way for a relatively brief period, does better for condition C than placebo or standard treatment), it by no means follows that RCT evidence is also the best evidence for the right theory (that treatment T, when administered in the way it will be 'in the wild' to a much wider population, in a less systematic way and for possibly very long periods, does better than standard treatment – seldom, if ever, of course placebo).

Robert Truog [22] argued this thesis in the particular case of the introduction of the Extracorporeal Membraneous Oxygenation (ECMO) technique for persistent pulmonary hypertension of the newborn – largely in that case on the grounds of the fact that both the new treatment (ECMO) and standard treatment were evolving significantly and that therefore an RCT was at best able to give us a reliable comparative result about two treatments that were both out of date by the time the trial had ended. Simply by keeping systematic records of how some treatment fares and whose effectiveness varies as it evolves, and by trying hard to eliminate other plausible causes of difference between treatments by suitable post hoc matching, we are likely to get evidence that is much more telling and relevant for how to treat current patients than that provided by an RCT on treatments which, as Truog suggests, may be out of date before the result is achieved.

And Robyn Bluhm [23] has argued – to my mind convincingly – that similar considerations apply to chronic diseases, more or less across the board: this time in part because of the short-term nature of RCTs compared with the long-term nature of the condi-

tions. We are likely to get more convincing evidence of the superiority of some new treatment by looking at its long-term track record compared with the long-term track record of the previous treatment on patients that were (at least approximately) comparable to the current ones in other relevant respects, than we are from a 'snap shot' RCT whose length is considerably shorter than the average period for which patients take the treatment concerned.

Once we view the issue as one of making sure we have the right evidence for the right theory, one would expect cases of exactly the kind commented on by Truog and Bluhm. Notice, then, in particular that these are not – in line with the evidence hierarchy idea – cases in which we cannot or do not have RCT evidence and therefore settle for evidence from observational studies as 'next best'. On the contrary, once we have identified accurately the theory we want evidence for, observational studies are likely to provide the best evidence according to some very simple ideas from basic scientific method or philosophy of science.

Conclusion

Of course the issue of the role of evidence in medicine is a complex and multifaceted one. I have argued here only that at least some light can be shed on some facets by returning to the basic ideas about theory-testing studied in the philosophy of science.

References

1. Sackett, D. L., Strauss, S. E., Richardson, W. S., Rosenberg, W. & Haynes, R. B. (2000) Evidence Based Medicine. How to Practice and Teach EBM, 2nd edn. Edinburgh and London: Churchill Livingston.
2. Hill, A. B. (1952) The Principles of Medical Statistics, 1st edn (12th Edn 1991). London: Livingstone.
3. Sackett, D. L. (1996) Evidence based medicine. What it is and what it isn't. *British Medical Journal*, 312, 71–72.
4. Worrall, J. (2007) Evidence in medicine and evidence-based medicine. *Philosophy Compass*, 2 (6), 981–1022.
5. Barton, S. (2000) Which clinical studies provide the best evidence? The best RCT still trumps the best observational study. *British Medical Journal*, 321, 255–256.
6. Agency for Healthcare Research and Quality (2002) Systems to Rate the Strength of Scientific Evidence. Rockville, MD: AHRQ.
7. Schünemann, H. J., Fretheim, A. & Oxman, A. D. (2006) Improving the use of research evidence in guideline development: 9. Grading evidence and recommendations. *Health Research Policy and Systems*, 4, 21.
8. Rawlins, M. D. (2008) De Testimonio: On the Evidence for Decisions about the use of Therapeutic Interventions. The Harveian Oration 2008. London: Royal College of Physicians.
9. Glaziou, P., Chalmers, I., Rawlins, M. & McCulloch, P. (2007) When are randomised trials Unnecessary? Picking signal from noise. *British Medical Journal*, 334, 349–351.
10. Howson, C. & Urbach, P. M. (2007) Scientific Reasoning: The Bayesian Approach. Chicago, IL: Open Court.
11. Popper, K. R. (1958) The Logic of Scientific Discovery. London: Hutchison.
12. Mayo, D. (1996) Error and the Growth of Experimental Knowledge. Chicago, IL: University of Chicago Press.
13. Duhem, P. (1954) The Aim and Structure of Physical Theory. Princeton, NJ: Princeton University Press.
14. Clarke, M. (2004) *Cochrane Collaboration – Systematic Reviews and the Cochrane Collaboration*. Available at: http://209.211.250.105/docs/whycc.htm (last accessed 12 February 2010).
15. Leibovici, L. (2001) Effects of remote, retroactive, intercessory prayer on outcomes in patients with bloodstream infection. *British Medical Journal*, 323, 1450–1451.
16. Leibovici, L. (2002) Author's reply. *British Medical Journal*, Letters, 324, 1037.
17. Lindley, D. V. (1982) The role of randomisation in inference. In Philosophy of Science Association, Vol. 2 (ed. A. Fine), pp. 431–446. East Lansing, MI: Philosophy of Science Association.
18. Worrall, J. (2009) Do *we* need some large, simple randomized trials in medicine? In European Philosophy of Science, Vol. 1 (ed. M. Suarez), pp. 289–301. Dordrecht: Springer.
19. Papp, K. A., Langley, R. G. & Lebwohl, M. (2008) Efficacy and safety of ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with psoriasis: 52-week results from a randomised, double-blind, placebo-controlled trial (PHOENIX 2). *Lancet*, 371, 1675–1684; Muers, M. F., Stephens, R. J. & Fisher, P. (2008) Active sympton control with or without chemotherapy in the treatment of patients with malignant pleural mesothelioma (MS01): a multicentre randomised trial. *Lancet*, 371, 1685–1694.
20. Bartlett, C., Doyal, L., Ebrahim, S., Davey, P., Bachmann, M., Egger, M. & Dieppe, P. (2005) The causes and effects of socio-demographic exclusions from clinical trials. *Health Technology Assessment*, 9, 1–152.
21. *Lancet*. (1982) Lessons from the benoxaprofen affair. *Lancet*, 320 (8297), 529–530.
22. Truog, R. D. (1992) Randomized controlled trials: lessons from ECMO. *Clinical Research*, 40 (3), 519–527.
23. Bluhm, R. (2009) Further Lessons from ECMO: The Epistemology and Ethics of Chronic Disease Research (forthcoming).