Chapter 27
Do We Need Some Large, Simple Randomized Trials in Medicine?

John Worrall

27.1 Introduction: Why Randomize?

In a randomized clinical trial (RCT), a group of patients, initially assembled through a mixture of deliberation (involving explicit inclusion and exclusion criteria) and serendipity (which patients happen to walk into which doctor's clinic while the trial is in progress), are divided by some random process into an experimental group (members of which will receive the therapy under test) and a control group (members of which will receive some other treatment – perhaps placebo, perhaps the currently standard treatment for the condition at issue). In a 'double blind' trial neither the patient nor the clinician knows to which of the groups a particular patient belongs. The results of double blind randomized controlled trials are almost universally regarded as providing the 'gold standard' for evidence in medicine. Fairly extreme claims to this effect can be found in the literature. For example the statistician Tukey wrote (1977, p. 679) "almost the *only* source of reliable evidence [in medicine] . . . is that obtained from . . . carefully conducted randomised trials". And the clinician Victor Herbert claimed (1977, p. 690) ". . .the only source of reliable evidence rising to the level of proof about the usefulness of any new therapy is that obtained from well-planned and carefully conducted randomized, and, where possible, coded (double blind) clinical trials. [Other] studies may point in a direction, but cannot be *evidence* as lawyers use the term evidence to mean something probative . . . [that is] tending to prove or actually proving". Finally, the still very influential movement in favour of 'Evidence Based Medicine' (EBM) that began at McMaster University in the 1980s was initially often regarded as endorsing the claim that only RCTs provide real scientifically telling evidence.

EBM now explicitly endorses a more guarded view involving a hierarchy of evidence of different weights. But although these hierarchies[1] explicitly allow that

J. Worrall (✉)
London School of Economics
www.ahrq.gov

[1] A 2002 study identified no less than 40 such systems of grading evidence (Agency for Healthcare Research and Quality. 2002. *Systems to rate the strength of scientific evidence*. Rockville MD:AHRQ,); while a 2006 survey found 20 more (Schünemann, Holger J., Atle Fretheim and

M. Suárez et al. (eds.), *EPSA Philosophical Issues in the Sciences: Launch of the European Philosophy of Science Association*,
DOI 10.1007/978-90-481-3252-2_27, © Springer Science+Business Media B.V. 2010

other forms of evidence can legitimately play a probative role, they still (all) unambiguously place evidence from RCTs at the top (sometimes along with systematic reviews or meta-analyses of RCTs).[2]

So although the extreme view that the only truly scientific evidence for the effectiveness of some treatment is that from an RCT seems to have been largely abandoned, nonetheless RCTs continue to be regarded as carrying special epistemic weight. Why? In previous work,[3] I identified five different types of answer:

1. Fisher's argument that randomization is necessary to underwrite the logic of the classical statistical significance test.[4]
2. Randomization controls for *all* possible confounders – known and unknown. This is sociologically speaking the argument that has carried most weight. Clearly a central issue in evaluating the weight of evidence supplied by any clinical trial in which the experimental group does better on average is whether there might be some other overall difference between those in that group and those in the control group – a difference that played a role (possibly the major role) in the 'positive' outcome. In principle the groups could be deliberately matched for 'known confounders' – factors, like age, sex, absence or presence of comorbidities, etc. that background knowledge makes it plausible might play a role in the outcome. But clearly this leaves open the possibility that there are 'unknown confounders' – factors that also play a role in outcome but that background knowledge gives us no reason to think do so – and which may be (of course, by definition, unknown to the clinicians) unbalanced between the two groups. Randomization's many admirers believe that it (and only it) solves this problem.[5]
3. Randomization controls for the *particular* possible confounder: 'selection bias'. Since selection bias is sometimes used in a number of (often very wide) senses, it is important to emphasise that by this term I mean specifically any bias that is, or may be, introduced as a direct result of the clinicians' having control over which group a particular patient goes into.
4. It is just an empirical fact that non-randomized trial designs exaggerate positive treatment effects.
5. Only an RCT can distinguish a real causal connection (between intervention and outcome) from a 'mere correlation' between the two. This argument arises from the burgeoning literature on probabilistic causality, but on analysis is quite quickly revealed to be simply argument 2 under a rather different guise.[6]

Andrew D. Oxman. 2006. Improving the use of research evidence in guideline development: 9. Grading evidence and recommendations. *Health Research Policy and Systems*. 4:21).

[2] Meta-analyses and systematic reviews are attempts to amalgamate different studies on the 'same' intervention into one overall result. They face many interesting methodological problems.

[3] Worrall (2002, 2007a, b, 2008).

[4] For references and an especially clear account of this argument of Fisher's – together with an especially clear demonstration that the argument fails even on its own terms, see Howson (2000).

[5] So for example the Director of the UK Cochrane Centre, Mike Clarke, states on the Centre's Web-site that "[i]n a randomised trial, the *only* difference between the two groups being compared is that of most interest: the intervention under investigation." http://209.211.250.105/docs/whycc.htm. Accessed 18 December 2008.

[6] See in particular Worrall (2007a) and references therein.

Only argument 3 clearly survives critical scrutiny – or so I argued in the previous work alluded to. In this paper I want to look in more depth at argument 3 and its impact.

27.2 Selection and 'Treatment' Bias

The natural home of selection bias in the sense that I am understanding it is the non-randomized clinical trial, where the clinicians have direct control over which patients go into the experimental and which into the control group. Classic illustrations of how selection bias might operate to make the weight of a trial result highly questionable are provided by various early comparisons of patients treated surgically for some condition C and patients treated medically for the same condition. Of course, being considered operable forms a potentially very powerful selection bias – to be considered operable the patient will need to be in comparatively good condition, exhibit particular anatomical conditions and suffer from no major comorbidity. Hence any 'evidence' from such a trial that surgery is the better treatment for condition C would be clearly suspect.

Leaving it to clinicians to decide the group a patient goes into is clearly fraught with epistemological danger – especially, though not exclusively if they have a vested interest in a positive outcome. (Indeed as Hill [1937/71] pointed out there is even the 'opposite' danger that the clinician may 'bend over backwards' to be fair and produce a control group that is better overall in terms of positive prognostic factors and thus provide an overly severe test of the treatment, one which might well *underestimate* that treatment's virtues!) Moreover if the clinicians select then they will also know to which group a particular patient belongs, and may lavish particular attention on those they know to be in the experimental group (especially if the comparison is placebo). Hence what might be called '*treatment* bias' might be added to any baseline imbalances in the two groups ahead of treatment. (It is useful to have treatment bias as a separate category as we shall see.)

In historically controlled trials (sometimes also rather dismissively categorised as one kind of 'observational study'), the control group is provided by previous patients with the condition under investigation who were treated (preferably of course in the *recent* past) using the older 'standard treatment'. This means that all the patients actively involved in such trials are on the experimental treatment and the investigators know this. Of course there is always an attempt, and, in the more sophisticated historically controlled trials, a very great attempt, to match the historical controls with those currently being treated with respect to known prognostic factors for the condition at issue. However, it is unavoidable that all the patients actively involved in such a trial are given the experimental treatment and the clinicians know this. This seems to make the possibility of at least treatment bias inevitable in such trials.

Randomization as usually performed eliminates selection bias in this sense. So long as the protocol is followed, the decision about which group a particular patient belongs to is taken out of the clinicians' hands and is made instead by some random process (usually whether the next number in a table of random numbers is even or odd). Moreover if the trial is performed double blind then, so long as 'blind is maintained', neither the particular patient nor, more importantly in this respect, the clinicians know which group that patient is in, and so the possibility of treatment bias seems to be ruled out. (There are issues – often overlooked – about how long blind is in fact maintained in most clinical trials. But let's leave these issues aside for present purposes.) Notice however that there is nothing special about the randomization in this regard – nothing in the toss of the coin or the random number table really plays a role, instead randomizing is simply one way of taking control out of the hands of clinicians; and blinding (if effective) makes it impossible for the clinician to identify securely those in the treatment group and so no question of preferential treatment (treatment bias) arises. Notice also that nothing in the argument shows that the *only* way that selection bias can be eliminated is through randomization.[7]

27.3 How Large an Effect Is Selection Bias Likely to Produce?

Suppose I am right that controlling for selection bias is randomization's only unambiguous epistemic virtue. (I believe that this was in fact the view of Austin Bradford Hill, who is credited as the first to import Fisher's randomizing methodology into medicine.[8]) The next question – especially given that it is conceded on all sides that randomization may involve some 'ethical cost' – is surely: *how large* an effect is selection bias likely to produce if not controlled for? (Questions of likely effect size are very often underemphasised in medicine, I would argue, in favour of the question of simple statistical 'significance'.)

There has been increasing recognition, even amongst arch-advocates of RCTs, that the answer to this question may well be 'very small'. This recognition goes back at least to a brief letter to the Editor of the *British Medical Journal* in 1980 by Doll and Peto (1980). They allow that selection bias is 'hardly likely to produce a tenfold artefactual effect' though, they insist, it 'may well produce a twofold artefactual error'. It is unclear exactly what metric they are presupposing here, but we can get by just in qualitative terms: selection bias is likely to be quite small. I can only think that they are referring here to what might be called 'practically ineliminable' selection bias; since if performed badly enough trials subject to selection bias can be as biased as you like (think about the comparison mentioned above between surgical and medical interventions for the same condition). So we are talking about trials in which sensible efforts have been made to match the two groups,

[7] So for example Bartlett and colleagues who introduced ECMO as a treatment for PHSS simply switched from treating all babies admitted to their hospital (U of Michigan) with the condition with the previously standard treatment to treating all babies admitted to their hospital with ECMO. No selection! (Though certainly the issue of treatment bias is a genuine one.) See Worrall (2008).

[8] See Bradford Hill op. cit.

without randomizing. Doll and Peto's concession that selection bias is 'likely' to be small, at once admits some intuitive Bayesianism (we are allowed to judge prior likelihoods) and implicitly admits that randomization is not needed, that historically controlled trials may be sufficient when the treatment effect (revealed by the historically controlled trial) is large. (But this concession went unnoticed in several important cases including I would argue the famous ECMO case.[9])

The concession was made entirely explicit in Peto et al. (1995) which allows that "… randomized trials may be unnecessary… For example, randomization is not needed to show that prolonged cigarette smoking causes cancer…" (p. 32) And more recently and more explicitly still, Paul Glaziou et al. (2007) write "'Some treatments have such dramatic effects that biases can be ruled out without randomised trials." (p. 351) Of course these concessions were not made before time – it is easy to produce a long list of treatments that are (i) established in medicine, which (ii) no sane person could deny are effective and yet (iii) have never been subjected to an RCT.[10] Nonetheless the concessions are important and welcome.

27.4 How Doll, Peto and Others Turn the Smallness of Selection Bias into an Argument for RCTs

But the central aim of Doll and Peto (1980), of Yusuf et al. (1984) and of Peto et al. (1995) was not at all to argue for the virtues of some historically controlled trials. On the contrary, their aim was to use the fact that historically controlled trials are bound to suffer from the possibility of selection bias, *even if* that bias is small, as a further argument for the necessity of *RCT*s! Their aim was to argue in effect that we need RCTs even more not less. Hence the title of the (1984) paper: 'Why do we need some large, simple, randomized trials?' – from which I in turn took the title of this paper.

One crucial premise of the Doll/Peto argument is that the romantic age of medicine – the days of the great breakthroughs producing new treatments with 'dramatic' effects – is over (or at least very largely so). Doll and Peto wrote (op. cit., p. 44) "most of the really important therapeutic advances of the past decade have involved the recognition that some particular treatment for some common condition yields a *small but important* improvement in the proportion of favourable outcomes." To which Yusuf et al. (1984, p.410) added: "if any widely practicable intervention had a very large effect, … then… these huge gains in therapy are likely to be identified more or less reliably by simple clinical observation, by 'historically controlled' comparisons, or by a variety of other informal or semi-formal non-randomized methods". Hence (op. cit., p.411) "if there remains some controversy about the efficacy of any widely practicable treatment, its effects on major endpoints may well be either nil, or moderate…."

[9] See Worrall (2008).

[10] See Worrall (2007b) and the list in Rawlins (2008).

So a clinical trial, in the current situation in medicine, will need to be able to distinguish between a null and a 'moderately' (they really mean small) positive effect. And this is exactly what an historically controlled trial cannot do – even though the (practically ineliminable) bias to which it is subject is admittedly itself small. Hence we need an RCT to do this. Moreover, we need *large* RCTs: "It is chiefly because one [nowadays] usually needs to be able to distinguish reliably between moderate and null effects that trials need to be *strictly* randomized ... and much, much larger than is currently usual" (Yusuf et al. 1984, p. 410). The need for the trial to be large is based on the fact that randomization, despite what some of its advocates often seem to claim, cannot be guaranteed to equalize the experimental and control groups in terms of other potentially prognostic factors. Trials no matter how carefully randomized are, instead, subject to 'random error' – and the claim is that 'random error' is likely to be small if, but only if, the trial is large. As Doll and Peto (op. cit.) put it "the small randomized trials that are regrettably commonplace nowadays have random errors which are often far larger than the real differences to be detected."

So we need large RCTs to distinguish the small effects that are all we can reasonably expect. In order to be *very* large, practically speaking they need to be multi-centre; and this in turn means that the trials need to have a very *simple* protocol since complexity may produce differences between treatment centres that obscure the true effects. Finally – and importantly – small effect sizes are not to be scoffed at: treatments yielding small effects on common conditions may well finish up saving more lives overall than dramatic treatments for much rarer conditions. For example, Peto et al. (1995) claim that the ISIS-2 study which found an absolute risk reduction of heart attacks of under 2% and whose results were published in 1988 had probably by 1995 "avoid[ed] about 100,000 vascular deaths in developed countries alone." (p. 26).

This, then, is 'Why we need some large, simple, randomized trials'. And this view has proved very influential – especially in cardiology. Large trials on reducing heart attacks and stroke, for example, have included (with numbers of patients in parentheses):

ASSET (5,200)
GISSI-2 (12,700)
GISSI-3 (19,500)
CURE (12,200)
ISIS-2 (17,000)
ISIS-4 (58,000)[11]

This is certainly an interesting and seemingly powerful argument for the special epistemic power of randomized trials – not one that is usually cited and not one that I analysed in my earlier papers.

[11] These numbers are taken from (and my treatment influenced by) Penston (2003).

27.5 Analysis of the Argument

Clearly a crucial premise of this argument is that it is at least unlikely that the treatments being tested in the current situation in medicine will have large effects. Doll and Peto and collaborators give no argument for this beyond the rather strange claim that any big effects out there would probably already have been discovered. This sounds rather like physicists in the nineteenth century holding the view that Newton had made the big breakthrough and that all that was left for physicists to do was to fill in details. There have been some recent quite major breakthroughs – for example in the treatment of leukaemia and of HIV Aids. And it is difficult to see the general grounds for the pessimism involved in their assumption. (And it should be carefully noted, I believe, that, despite being arch-advocates of the epistemic superiority of RCTs, they are admitting that (sophisticated) historically controlled trials are sufficient to reveal anything other than 'moderate' (really: small) effects.)

However their argument can be re-gigged so as to avoid reliance on this pessimistic premise by making it more local. There may of course *in particular cases* be good reasons to think that some proposed treatment aimed at, say, reducing the risk of myocardial infarction or strokes is unlikely to have a really 'dramatic' effect. Hence the argument would now suggest that, in those cases where we have good prior reason to think the effect of the treatment, if any, is 'moderate', we need to perform RCTs, since the selection bias inherent in non-randomized studies may produce effects of the same order of magnitude as (or higher than) the likely effect.

Can there be any reason to question even this more measured claim? There seem to me to be two such reasons.

27.5.1 The Issue of 'External Validity'

As is frequently conceded, the issue of 'external validity' is one that can always be raised for any trial – though the conceder usually then goes on to categorise external validity as a difficult problem and practically to ignore it! Suppose it is agreed that the RCT is the most reliable means of arriving at the 'right' result so far as the set of patients in the study in the study is concerned. (This is usually called 'internal validity'.) It is still reasonable to question whether that result is likely to generalise to the 'target population' (that is, the set of people who will be treated if the treatment is declared 'effective' in the trial). It should be noted that, contrary to a fairly widespread myth, there is no guarantee (even of a classical statistical sort) that a randomized study's result will generalise in this way, since there is no sense in which the initial study group is a random sample from any specified population.

Standardly, research reports in the medical journals will have titles like (taken from a randomly chosen 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. . .”.[12] 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). The trials will generally involve some *very precise treatment regimen* which the trialists are not allowed to alter or adjust and will generally run for some *relatively brief period* (as Michael Rawlins, the head of the UK National Institute for Clinical Excellence, reports “Most RCTs, even for interventions that are likely to be used by patients for many years, are of only six to 24 months duration.”[13]) And the study will report that administration of substance S is (or is not) effective – meaning more (or no 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 for efficacy in treating 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[14] than some comparator treatment (often, as I say, placebo). This is the claim for which the RCT provides evidence – let’s assume for present purposes impeccable evidence.

But this is not, of course, the claim that the practising physician would like to have evidence for. She would like to know whether the treatment is effective (in a wide sense that certainly involves factoring in any side-effects, whether short or long term) when prescribed to the sorts of patients she would like to prescribe it to. This ‘target population’ is not very precisely characterised but will certainly include many types of patient excluded from the trial (the elderly perhaps, or those with significant co-mordibity). 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 physician may want to prescribe S for a long time – certainly much longer than the trial itself is likely to have lasted.

Note that the issue of external validity is not what is sometimes dismissively 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 David Hume’s case, we often know on good specific grounds that the trial population and the target population are different. For example, a study by Bartlett et al. (2005) looked at 25 recent RCTs on NSAIDs and 27 recent RCTs on Statins and found that older people, women and ethnic minorities were (quite significantly) under-represented compared to the general (and therefore also presumably the ‘target’ population). Moreover not only do we know

[12] Lancet **371**, 2008, pp. 1665 and 1685.

[13] Rawlins (op. cit., p. 16).

[14] Of course ‘effectiveness’ is a tricky notion too – positive effect on the ‘target disorder’ is only part of the story, side effects need to be taken into account too.

that there are such differences, background knowledge, largely in the form of previous experience, lends good grounds for thinking that those differences may result in differences in outcome (and it lends no reason to think that such differences will be small).

Nor do we need appeal here merely to logical possibility: there are a number of real cases in which a treatment endorsed by an RCT had to be withdrawn later because of significantly deleterious overall outcome. One such case involved Benoxaprofen (Opren). 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 and hence was likely greatly to increase patient compliance. A large RCT was performed in a trial restricted to 18–65 year olds. The trial had an impressively positive result; Opren was very aggressively promoted and duly cornered the market. Now, it is 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 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. Michael Rawlins cites a total of 22 drugs that have been approved by RCTs in recent years only to be later withdrawn for safety reasons (2008, p. 22).

The issue of external validity arises especially sharply, I believe, in the case of the very large randomized trials recommended by Doll and Peto. If you are performing a large trial, you are (as Doll and Peto suggest) expecting no more than a small effect (and a trial would in practice never get to be large if the effect were itself at all large). While intuitively it's quite unlikely that a therapy that produces, say, a 50% reduction in absolute risk even in a small RCT will not prove of positive benefit in the target population as a whole, this seems altogether more plausible in the case of tiny "effects" "revealed" by mega-trials. Like all trials, these trials involve specific 'selection criteria' (partly, though far from exclusively, with ethical considerations in mind). Those meeting these criteria *may well* suffer from fewer side-effects or have a different response than is typical within the overall target population.

For example, in the GISSI-3 study assessing a proposed treatment for ischaemic heart disease, only 45% of the 43,047 people admitted to the coronary care units in the hospitals involved in the trial were randomized. A back-up study showed the excluded group to have roughly twice the mortality of the included group.[15] Notice that the absolute risk reduction allegedly found by the GISSI-3 study was 1.4%. It is not as if the reasons for exclusion are always clear-cut (so that the 'target population' could be more precisely defined on the basis of the study). For example, one exclusion 'criterion' employed in the ASSENT-2 trial was "any other disorder that the investigator judged would place the patient at increased risk"! The ISIS-2 trial listed any further reason for exclusion "not specified by the protocol but by the responsible physician".[16] These trials also generally involve a 'run in' period meant to test for compliance, side effects, and, in statin cases, increased creatinine and

[15] For details and references see Penston (2003).

[16] Taken from Penston op. cit.

hyperkalaemia. It may *well* be that a therapy that has a tiny positive effect even in a large trial population that is unusually compliant, shows fewer immediate side-effects, have normal levels of creatinine and hyperkalaemia, etc., has a negative effect in a population where partial compliance, existence of side-effects and so on is the norm.

Of course on any account – Bayesian, as well as classical frequentist, and even commonsense – the larger the trial, other things being equal, the stronger the evidence. But other things never are equal, and here in particular the two factors (i) large population, but (ii) small effect pull in opposite directions. When we bring in the inclusion and exclusion criteria which lead to the study population satisfying special conditions not shared by the target population, it seems difficult to form a reasonable view about what the study result is telling us about the effect in the target population.

27.5.2 Are Such Small Effects Worth Having?

So, one problem with the argument of Doll and Peto (and collaborators) is the issue of external validity when such small effects (if any) are likely to be involved. The second issue is whether such small effects as may or may not be revealed in the mega-trials that they advocate are worth having if they exist at all.

For example, several such trials have investigated the effect of various statins on subsequent mortality from stroke and heart attack (LIPID, CARE, etc.). These have uniformly found absolute risk reductions of less than 2%. Here are some representative results.

Study	**Outcome**	**Abs RR**	**NTI**
LIPID	mortality	1.9%	98.1
CARE	stroke	1.2%	98.8
GISSI-3	composite	1.4%	98.6

Here the third column gives the absolute risk reduction. Put plainly, these results are telling us that if we go ahead and use these drugs for treatment, then even if the trial result happens to generalise (that is, the treated population turns out to reflect the study population – and this is certainly questionable, as we just saw), then more than 98% of those treated will get *no* benefit (see the fourth column representing '*Number Treated Ineffectively*').

It is crucial when trying to make a serious assessment of the (likely) impact of some treatment on a condition to ignore all talk of *relative* risk reduction: one hears figures of 30% or even 50% risk reductions bandied about, which sound striking, but are in fact systematically misleading since they suppress the base rate. Suppose, unrealistically but for sake of a particularly telling example, only 1 in a million of those whom medics propose to treat with some prophylactic medicine will on average develop some outcome (say a stroke within the next 5 years) if left untreated.

Then, if that medicine reduces the average rate to zero then this will of course represent a 100% relative risk reduction. It by no means follows, however, once we factor in side effects, that this is a treatment that can rationally be recommended. It is always *absolute* risk reduction that we need to know in order to make a rational decision about the use of some treatment. An equivalent statistic sometimes (laudably) used is the NNT, standing for ‘number needed to treat’. This is an expectation value: the number of patients you would need to treat on average in order to produce one positive event (recovery or amelioration of symptoms or whatever). So in the unrealistic example just cited, the NNT is 1 million. However surely the statistic (entirely analytically equivalent to either absolute risk reduction or NNT) that is likely to have most (rational) rhetorical impact is NTI – ‘number treated ineffectively’. This is just NNT *minus* 1, and measures the average number of people who will be treated ineffectively in producing just one positive event. Hence the NTI column in the above table.

But what, will go up the cry, if you are that 1 in a 100 (or whatever) who will benefit? Surely if the benefit is no myocardial infarction or no stroke in the next 5 years you want to reap that benefit. And of course if there were no ‘downside’ then treatment would be the rational course even with such high NTIs. But there always is a downside and it is this that Doll and Peto entirely ignore when producing their plausible argument for the importance of even small positive effects from treatments.

Returning to the trials on statins, these trials were regarded as the justification for introducing mass prescription of statins as prophylaxis for stroke and heart attack. In 2003, well over 5% of the entire US population were taking statins as prophylactic medicine. According to our “best” evidence, 98% of those will get no benefit (even assuming the results generalise).[17] This is a lot of people and means pretty good business for the pharmaceutical companies!

Once you factor in side-effects (and surely just being on long-term medication should count as a side-effect), it is surely at least questionable whether this treatment policy is sustainable. Of course some side-effects will (generally) be revealed in the trial and can be taken into account in deciding whether to treat or not, but the worrying thing is surely longer term side effects that do not (cannot) show up in the trial. Remember that, as Michael Rawlins points out, almost all trials last for between 6 and 24 months (and relatively few seem to be even close to the upper end). But statins, like puppies, are for life!

Again we are not dealing here with mere “philosophers’ logical possibilities”. One particular statin, Cerivastatin, was ‘sanctioned’ in an RCT but then quickly withdrawn because of an unexpectedly high number of deaths amongst those treated. It seems to me sobering to think that of those who died probably more than 98%, even on the most favourable interpretation of the trials on the basis of which the drug was introduced, were receiving no benefit from the drug.

[17] Figures again taken from Penston op. cit.

Doll and Peto, as we saw, make the apparently very cogent point that many lives may be saved by discovering treatments that have small effects provided the condition is common. But they are – clearly – only doing half of what ought to be the *expected utility calculation*! They entirely ignore the 'downside'. Suppose that some drug is in reality 1% effective for some condition, then the expected utility of using it as a treatment for that condition is:

P(helps) × utility(helps)+ P(doesn't help) × disutility(taking it ineffectively)

Given that the first probability is only 0.01 and the second 0.99 and given that there *is* a downside in terms of side effects, it cannot simply be assumed that this expected utility is positive. Medicine should surely beware the drive to treat at all costs.

27.6 Conclusion

Neither this, nor my earlier arguments about the evidential weight of various types of clinical trial, is at all aimed at denigrating RCTs in general, let alone questioning the application of scientific method in medicine. On the contrary they are aimed at encouraging the correct application of science in medicine. Randomization can sometimes be of epistemic value, so long as it is not regarded as an evidential *sine qua non*. The main thing is to keep one's critical, philosophical-commonsense faculties at full power: this new argument by Doll and Peto at least carries less weight than might first meet the eye.

References

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 Technol Assess 9:1–152

Doll R, Peto R (1980). Randomised controlled trials and retrospective controls. Br Med J 280:44

Glaziou P, Chalmers I, Rawlins M, McCulloch P (2007) When are randomised trials unnecessary? Picking signal from noise. Br Med J 334:349–351

Hill AB (1937) Principles of medical statistics, 1st edn. in 1937, 9th edn in 1971. Livingstone, London

Herbert V (1977) Acquiring new information while retaining old ethics. Science 198:690–693

Howson C (2000) Hume's problem: Induction and the justification of belief. Oxford University Press, Oxford

Penston J (2003) Fiction and fantasy in medical research. the large scale randomised trial. The London Press, London

Peto R, Collins R, Gray R (1995) Large scale randomized evidence: Large simple trials and overviews of trials. J Clin Epidemiol 48:23–40

Rawlins M (2008) De Testimonio: On the evidence for decisions about the use of therapeutic interventions. Royal College of Physicians. http://www.rcplondon.ac.uk/pubs/brochure.aspx?e = 262. Accessed 18 December 2008

Tukey JW (1977) Some thoughts on clinical trials, especially problems of multiplicity. Science 198:679–684

Worrall J (2002) What evidence in evidence-based medicine? Philos Sci 69:S316–S330

Worrall J (2007a) Why there's no cause to randomize. Br J Philos Sci 58:451–488

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

Worrall J (2008) Evidence and ethics in medicine. Perspect Biol Med 51:418–431

Yusuf S, Collins R, Peto R (1984) Why do we need some large, simple randomized trials? Statist Med 3:409–420