Chapter 2
Philosophy of Science Meets Medicine (Again): A Clearer-Sighted View of the Virtues of Blinding and of Tests for Blinding in Clinical Trials

John Worrall

Abstract A clinical trial is double-blind if neither the administering clinicians nor the participants know to which arm of the trial – experimental or control – any particular participant has been assigned. Double-blinding controls for various possible biases that might otherwise affect the trial's result; and hence seems to be an unambiguous methodological virtue. And if blinding is a virtue, then so also, it would seem, is testing that blinding was retained throughout the trial's course. As a matter of fact, however, end-of-trial tests for blinding are relatively seldom performed and, when they are performed, frequently find that blinding has been lost. Rather than decrying this situation, as might have been expected, leaders of Evidence-Based Medicine have become cool, or outright negative, about tests for blinding (and occasionally even about blinding itself). This paper investigates this *prima facie* mysterious situation from the point of view of the general account of evidence supplied by philosophy of science. It argues that, although interesting and unexpected complexities and difficulties are associated with blinding, the correct response to them is not a negative view of end-of-trial tests for blinding, but rather a nuanced but still positive view.

Keywords Clinical trials · Evidence · Blinding · End-of-trial tests for blinding

2.1 Introduction

My perspective on philosophy of science is one of the very few things that I share with John Locke. Locke believed that philosophers in general should restrict themselves to being, in his famous phrase, "underlabourers to the sciences" ("Epistle to the reader" in his (1689)). I completely agree – there is no special philosophical way

J. Worrall (✉)
Department of Philosophy, Logic and Scientific Method, London School of Economics,
London, UK
e-mail: j.worrall@lse.ac.uk

W. J. Gonzalez (ed.), *Current Trends in Philosophy of Science*, Synthese Library
462, https://doi.org/10.1007/978-3-031-01315-7_2

of knowing: just logic (and mathematics) on the one hand, and the empirically-based sciences on the other. Hence, the only legitimate role for a philosopher is as an applied logician underlabouring in science: clarifying concepts and especially analysing arguments that are of relevance to science. So, when I am asked "what's new in philosophy of science?", I say "Nothing, or, at least, nothing at the fundamental level; it's just one damned underlabourer's job after another". The only thing in my view that has really changed during my 50 years or so in philosophy of science is that it now supplies underlabourers to a much wider range of sciences than it used to.

I am proud, for example, of having played a small role in extending the range of philosophy of science to include medicine.[1] Philosophy of medicine has become in recent years a small but energetic sub-field of philosophy of science. One focus of my work there has been Evidence-Based Medicine (EBM) and the hierarchies of evidence that it has spawned. Now, it surely ought to seem astounding to any right-minded person that there is an explicit Evidence-Based Medicine movement. *Of course*, medicine, like any rational discipline, should be based on the evidence. What else? And, indeed, the novelty and interest of the EBM approach lies, not in its general claim that medicine should be based on evidence, which is beyond sensible criticism, but rather in its specific claims about what exactly counts as evidence and, especially which kinds of evidence count as particularly strong or telling. These specific claims can be, and have been, criticised. EBM is seen as endorsing the view that evidence from clinical trials that have been randomized (division of participants into experimental and control arms made by some hidden random process) provides the "gold standard". But in fact, many in medicine, influenced by EBM, hold that, when it comes to evidence from a single trial,[2] the absolute pinnacle, the "gold plus" or "platinum" standard evidence is that supplied by trials that are not only randomized but also *double blind*: meaning of course that neither the clinicians involved in the trial nor the participants know to which arm, experimental or control, any individual participant has been assigned.[3] This paper does some underlabouring on arguments that have carried weight in medicine concerning the virtues of blinding and of testing to see that blinding has been retained during the course of a trial. Of course, "underlabourer" need not imply "underling". Philosophers of science are good at spotting detritus amongst the wonderful structures produced by science and encouraging its removal. In particular, they are good at identifying invalid or confused arguments that have nonetheless carried some persuasive weight in science. This is exactly what I try to do in this paper.

[1] Following the lead of Peter Urbach, see Urbach (1993) and, e.g., Worrall (2006, 2007a, b, 2010).

[2] EBM evidence hierarchies generally rank evidence from systematic reviews or meta-analyses (which attempt to amalgamate the results from many individual trials) as higher than evidence from any single trial.

[3] I shall count a trial as double blind only if everyone involved in the trial is blind to treatment arm for any particular participant, this includes both the participants and *all* the scientists: whether attending physicians, outcome assessors or data analysts. This cuts through the confusion noted, e.g., in Schulz and Grimes (2002, 697).

So, the most telling evidence for the effectiveness of some treatment, according to many in medicine (influenced by the original EBM position), is evidence provided by trials that are (a) randomized and (b) double-blind. My previous work in philosophy of medicine has concentrated largely on randomization and has argued that while randomizing has some epistemic virtues, it is far from the *sine qua non* of scientific validity that many influenced by EBM have taken it to be (see the references in footnote 1). However, no one could deny that RCTs may sometimes provide important evidence of effectiveness and such evidence definitely seems to carry extra weight if the trial was performed double-blind. The virtues of blinding will be examined in some detail soon, but roughly (and partially): if a clinical trial, for instance, is *not* blind, and if (as may be the case) the clinicians involved have a vested interest in achieving a "positive" result, then they may (perhaps subconsciously) bestow special care on the participants whom they know to be taking the experimental treatment, and those participants may improve because of that auxiliary care and the expectations raised by it, rather than because of any effect of the treatment under trial. More significantly, if the trial is not blind and the outcome involved is "subjective" ("participant showed some improvement in her symptoms") rather than "objective" ("participant died"), then an unblinded clinician (acting as outcome assessor) may – again perhaps subconsciously – be more inclined to declare a positive outcome for a participant that she knows to be receiving the experimental treatment.[4] Either of these possibilities, if actualised, would clearly bias the outcome: pushing that outcome toward a "positive" result whether or not the experimental treatment is actually better than the control treatment. If, on the other hand, it is the participants who are not blinded, then those knowingly taking the control treatment (let's suppose it's a placebo) may well, for example, break the trial-protocol and seek extra treatment for their condition outside the trial. This possibility, if actualised, would again bias the trial outcome but now in the opposite direction: because a number (or a disproportionate number) of participants on the placebo arm took concomitant therapies, a "negative" result might well be produced (meaning that the experimental treatment would be taken as "failing to outperform placebo") even though the experimental treatment is in fact more effective than placebo (but not more effective, or not much more effective, than placebo plus the concomitant treatment).

Double-blinding avoids these potential biases, as well as others that will be considered shortly, and hence appears to be a clear epistemic virtue. If the trial starts out unblinded (meaning that treatment allocation was not successfully concealed), then the scope for such biases to intrude is maximal; but, if a trial becomes unblinded

[4] This is far from being a merely theoretical possibility. Both Sackett (2011) and Schulz and Grimes (2002) cite the celebrated case of an RCT carried out on a triple concoction of cyclophosphamide, prednisone, and plasma exchange as a treatment for multiple sclerosis (Noseworthy et al. 1994). Sackett reports (2011, 675–676) that, in this trial, 'participants were periodically examined by two groups of neurologists, one group blind to their treatment groups and the other unblind. The blind neurologists found no effect ... on patient-outcomes, but the unblind neurologists concluded that [the] triple therapy was effective.'

at any stage before its end, then there is some scope for biases to intrude. It seems clear, then, that the epistemic ideal is that a trial should begin, and remain, double-blind.[5] And if so, the importance of *testing whether blindness was maintained throughout the trial* also seems clear: the failure of such a test may necessitate a re-evaluation of the weight of evidence produced by the trial, because it raises the possibility that biases have affected the outcome.

However, if blinding trials and testing for the retention of blinding at the end of the trial form the epistemic ideal, then it is an ideal more often honoured in the breach than the observance. Ney et al. (1986, 120) complained some time ago that "Although current medical science often relies on the double-blind trial to determine the value of a medication, there is very little evidence that the double-blind trial is blind for anybody, except those who read the report." This view has been endorsed by recent, more systematic research. For instance, a study by Fergusson et al. (2004) looked at 191 published reports of placebo-controlled trials and found that only 13 (8%) reported end-of-trial tests for blinding, and of those 13, blinding was discovered to have been lost in 9 of them (60%). Boutron et al. (2005) undertook a study that looked at 82 trial reports; while 54 of those trials tested for blinding of participants, in 22 of them blinding was reported to have been lost. As is reflected in the fact that these two studies came to quantitatively different (albeit similarly worrying) conclusions, there are issues about the representativeness of their samples of trial reports. A study by Hrobjartsson et al. (2007) not only was much larger but also made explicit and systematic attempts to take a representative sample of trials. This study ended up looking at 1599 trials and found that, of these, only 31 (2%) reported undertaking end-of-trial tests for blinding (of "key" trial personnel – whether clinicians, participants or outcome-assessors); of those 31 trial reports, 14 recorded that blinding was retained, 7 that blinding had not been retained, and in the remaining 19, the results of the end-of-trial test were either unclear or unreported.

So, this most extensive and rigorous study found that end-of-trial blinding was tested-for and blinding clearly discovered to have been retained in only 14 out of a

[5] As we shall see, the "remains blind" condition is important. Schulz and Grimes (2002, 696), for example, complain that: "many medical researchers confuse the term blinding with allocation concealment... [In fact] the term blinding refers to keeping trial participants, investigators .... or assessors ... unaware of an assigned intervention, so that they are not influenced by that knowledge." The view turns out to be controversial however: the medical statistician, Stephen Senn, had earlier complained (2004) about investigators who, "consider a trial to be double-blind when the patient, investigators, and outcome assessors are unaware of the patient's assigned treatment throughout the conduct of the trial." But Senn states, they are "quite wrong to do so." The reason they are wrong, according to Senn, is that "The whole point of a successful double-blind trial is that there should be un-blinding through efficacy. That is to say that there should be no incidental reasons, apart from efficacy, as to why the treatments are distinguishable but that the treatments should reveal themselves through efficacy. If the treatments are not distinguishable at all, then the treatments have not been proved different." I will consider Senn's claim in detail later. But note for now that, in whatever way and at whatever stage it may be broken, breaking blind inevitably opens up a trial to *possible* bias.

total of 1599 trials. The study's conclusion (Hrobjartsson et al., 2007, 659) – that "There is an urgent need for improving the methods of evaluating the success of blinding" – seems, therefore, amply justified. Since making sure that blinding is instituted and retained in a trial is a matter of avoiding biasing the trial's outcome, and since a central driving force behind EBM was the elimination of bias (or confounding) in trials, it might be expected that the leaders of EBM would be at the forefront of those endorsing Hrobjartsson et al's plea for "improving the methods of evaluating the success of blinding". The facts are surprisingly at odds with this expectation.

David Sackett, who was often called "the father of EBM", writing in direct response to the study by Hrobjartsson et al., explained why "we vigorously test for blindness before our trials, but not during them and never at their conclusion" (Sackett, 2004, 1136). Later, in his paper (Sackett, 2011), he went so far as to describe testing for blindness at the end of a trial as "playing a mug's game". Although they don't express it quite so colourfully, Sackett's negative view is shared by other leading proponents of EBM. Schulz and Grimes, for example, in an influential (2002) paper in *The Lancet* had earlier "question[ed] the usefulness of tests of blinding in some circumstances".

This negative view has had important practical consequences. Schulz was one of the authors of the CONSORT guidelines for clinical trials (*www.consort-statement.org*). These guidelines are endorsed by prominent EBM-ers and widely accepted within medicine, to the extent that satisfying them is effectively a necessary condition for a trial report to be published in a "top" medical journal. The guidelines – rather oddly – govern the reporting, rather than the conduct, of trials. But reporting may well have knock-on effects concerning conduct. From their first appearance in 2004, the CONSORT guidelines were lukewarm both about the value of testing for blindness and even about blinding itself. Recommendations concerning the reporting of both were conditional. The checklist entry for blinding specifies only "If [blinding is] done, [an ideal trial report should state] who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes)": hardly a ringing endorsement of what seems a great epistemic virtue. And the entry concerning end-of-trial tests for blindness is again conditional and therefore lukewarm at best: "If [tests for blindness are] done, [an ideal trial report should state] how the success of blinding was evaluated". A 2010 revision of the guidelines (still in force) went further – the recommendation about blinding remains in conditional form, but mention of end-of-trial tests of blindness is dropped entirely. As its authors reported in one of several papers introducing the revised guidelines, in CONSORT 2010, "we have eliminated mention of how the success of blinding (masking) was assessed" (Schulz et al., 2010, 701). So there is currently no incentive for an investigator to report on whether or not blind was maintained in her trial and hence no incentive for her to institute a test to check if it was maintained. Indeed Sackett at least, as just noted, explicitly advises against making such a check.

This seems baffling. Double-blinding appears to be a clear evidential virtue and, if that's true, then it surely follows that there is also virtue in performing tests to check if a trial was in fact double-blind throughout its course. And yet some of the staunchest advocates of EBM feel that such tests are unnecessary and even that performing them may be a “mug’s game”! Why?[6]

In the next section, I analyse the arguments that have traditionally been taken as establishing that double-blinding a trial, both initially and throughout its course, adds weight to the evidence it provides. Then I analyse the claims presented by Schulz and Grimes, Sackett and others to underwrite their negative view of end-of-trial-tests for blinding. Their views ultimately evaporate as based on confusion but this is more than a mere clarificatory exercise. First, it turns out that some valid and interesting points underlie their analyses and show that the issues both of blinding and of end-of-trial tests for blinding are more complex than might initially be thought. But secondly, once these valid points are clearly articulated, it becomes evident that the correct conclusion from them is a nuanced but still positive view of the value of end-of-trial tests This in turn potentially has practical consequences: it mandates a re-revision of the CONSORT guidelines (as explained in the final section of this paper – Sect. 2.4).

2.2 The Traditional View: Evidence from Double-Blinded Trials Is More Telling

Despite some claims in philosophy of science,[7] the *fundamental* principles involved in weighing evidence are, I believe, both universal and simple. There is, in particular, one principle that is common to all serious accounts of confirmation and that, on its own, sheds a great deal of light on issues concerning evidence from clinical trials. The principle states that *evidence e tells more strongly in favour of a theory T if e not only conforms with T (*in the simplest case, follows from it deductively, given accepted auxiliaries and initial conditions), but also, at the same time, *tells against* plausible theories that rival T. The basic idea is that evidence tells more strongly in favour of a theory to the extent that it not only conforms with the theory but also picks out that theory from its competitors – in the ideal case, the evidence picks out the theory as its only currently available explanation.

The principle lies behind Popper’s idea that theories are particularly strongly confirmed (“corroborated” was the term he preferred) by passing “severe tests”, but

[6] Jeremy Howick first made me aware of this problem and gave me important guidance to the literature on it. We discussed the issues at length during the course of his doctoral studies at LSE. His own treatment of, and attempt to resolve the problem are contained in chapter 6 of his (2011) – that chapter being itself a revised version of part of his PhD thesis. Unsurprisingly there are a number of overlaps between his treatment and the one developed in this paper. But also a number of differences of approach and content (some of which I will allude to later). This paper is how chapter 6 of Jeremy’s book would have read had I been writing it.

[7] The most prominent advocate of the view that methodological principles are not universal, but instead have a complex and historically-evolving character, is Larry Laudan. For references to his work and my rebuttal of his claims see Worrall (1988).

it is especially clearly underwritten by the Bayesian account of confirmation. On this approach, the degree to which the discovery that potential evidence e actually holds supports the theory T is measured by the ratio p(e/T &B)/p(e/B), where B is background knowledge. In the simplest (positive) case where T, against the background B, deductively entails e, the numerator in this ratio (sometimes called "the Bayes factor") takes the value 1 and so the ratio reduces to 1/p(e/B). The denominator here measures the "prior probability" of the evidence e; so, remembering that probabilities vary between 0 and 1, the basic idea is that the Bayes factor is higher, that is, e tells more strongly in favour of T, the less likely e is otherwise; where "otherwise" means "on the assumption that some theory other than T were true". More precisely, if background knowledge B allows only a finite range of theories T_i of which T_n (=T) is the theory under consideration, then

$$p(e) = \sum_{i=1,\ldots,n} p(T_i)\cdot p(e|T_i)$$

That is, p(e) is the sum of the degrees to which e would be expected on each of the various rival theories, weighted by the plausibilities (prior probabilities) of those theories. In particular, p(e) is the smaller, and so 1/p(e) (and with it the Bayes factor measuring strength of confirmation) is the larger, the fewer plausible theories, aside from T itself, that attribute significant probability to e. Intuitively: the more unlikely evidence e would be, in the light of plausible theories other than T, the more the fact that T correctly predicts e tells in favour of it. Howson and Urbach suggest that this reflects "the everyday experience that information that is particularly unexpected or surprising unless some hypothesis is assumed to be true, supports that hypothesis with particular force" (2004, 97).

This simple principle underwrites the whole process of controlling clinical trials. Suppose that a trial produces a "positive result". This certainly *might* be explained by the greater effectiveness of the experimental treatment, but of course lots of alternative explanations are possible. For example, those in the control group may have been older, less fit, had more concomitant conditions ... than those in experimental group. Every time such an alternative is "controlled for", that is, every time the experimental and control groups are made similar with respect to such a "possible confounder", one alternative explanation of the positive result is undermined and the strength carried by the trial result in favour of the effectiveness of the experimental treatment is increased – in accordance with this basic principle.

But even if there is strong evidence that the two groups in a trial were initially similar in respect of relevant possible confounders, the evidence for effectiveness from an observed "positive" result becomes weaker if differences between the two groups emerged *during the trial* – differences that might plausibly have had an impact on the outcome. In case randomization has been used to create the control and experimental groups, such differences are often referred to as "post-randomization differences" or "post-randomization confounders".

If a trial is performed without blinding, then it is open to a range of post-randomization confounders (or “biases”); whereas blinding a trial provides a safeguard against such confounders. Suppose, for example, that the participants in a placebo-controlled trial know to which treatment they have been assigned; then those on the experimental treatment might well develop higher expectations of a good outcome than those on placebo; this difference would introduce an “expectation bias”, which might, independently of any “characteristic” feature of the experimental treatment, lead to a better outcome, at least in trials on outcomes such as pain relief that are known to be subject to the placebo effect.[8] Blinding participants as to which arm of the trial they are on effectively eliminates the possibility of expectation bias; and hence, in accordance with our basic principle, increases the evidential weight of any observed ‘positive result’ in favour of the theory that the experimental treatment is effective.[9]

Similarly, at any rate in the case of a trial where the outcomes are “subjective” (perhaps dependent in part on what response a participant reports to the clinician), blinding eliminates potential “response bias”: if the participant knows that s/he has been given the experimental treatment and is inclined to please the clinicians, then s/he may be more inclined to report an improvement in her/his condition. If the participant knows that s/he is, instead, on the placebo arm of a trial, then s/he may be more inclined to drop out of the trial: blinding reduces the possibility of “attrition bias”. Moreover, a participant who knows s/he is on the experimental arm may feel a greater incentive to stick to the protocol when not being directly observed than a participant who knows s/he is on the placebo arm (after all, the latter participants may well share the common, but sometimes mistaken, view that taking a placebo is the equivalent of doing nothing) – in other words, blinding helps avoid “non-compliance bias”. Finally, as noted earlier, if a participant knows s/he is on the placebo arm of a trial, there is clearly a greater incentive to seek treatment for her/his condition outside the trial; if s/he is blind to treatment then any such differential incentive disappears – blinding undermines the possibility of “co-intervention bias”.[10]

If, on the other hand, it is the clinicians who are not blind and instead know which treatment each participant is given, then, because they often expect, or hope for, a result that favours the experimental treatment, they may communicate their expectations of success to those participants whom they know to be on that treatment; and those participants may then experience a greater expectation/placebo effect than

[8] The placebo effect being in general stronger if the person taking it believes that the treatment has some specific or characteristic effect beyond placebo (see, for example, Benedetti 2014).

[9] Of course, other methodological factors – such as the size of the study population – also play a role. The claim is not that if a trial was performed double-blind then its result is automatically telling; but only that a trial that is double-blind carries more weight than an otherwise methodologically similar one that is not double-blind.

[10] Of course, it may just happen that, in a particular trial, more people on the placebo arm, despite not knowing this, took concomitant treatment; but, unlike when the trial becomes unblinded, there would be no systematic reason to believe that they might.

those who are taking the placebo. Hence blinding controls against another form of expectation bias (“communication of expectation bias”). Also, an unblinded clinician may (perhaps unconsciously) give better, more solicitous treatment to those whom s/he knows to be on the treatment arm, and those participants may improve because of that auxiliary treatment rather than because of the intervention under test. Conversely, a clinician who knows that a participant is taking the placebo, and who prioritises the welfare of the participants over the scientific validity of the trial, may prescribe subsidiary treatment. Blinding eliminates systematic “differential subsidiary treatment bias”. Finally, where the outcome at issue is somewhat subjective, then, as already noted, a clinician who would prefer a “positive” result may be (perhaps sub-consciously) more inclined to judge that a participant’s symptoms have improved if s/he knows that that participant has been given the experimental treatment. Blinding helps avoid the possibility of “observer (or reporter) bias”.

And, in all these cases, by avoiding a potential bias, blinding rules out another possible explanation of the result. Hence, in accordance with the basic principle identified above, it increases the weight of evidence produced by the trial.

So, what’s not to like about double-blinding? And, if blinding is such an obvious epistemic virtue, then:

2.3 How Could Testing for Blindness at the End of a Trial Possibly Be a “Mug’s Game”?

Why, then, do some of the leading proponents of EBM sometimes at least seem to give the impression that they are rather lukewarm about blinding? And, more particularly, why are they at best cool, or even outright negative about testing for blindness at the end of a trial? How could it possibly be considered a “mug’s game” to check if blinding has been maintained throughout a trial? The period during which biases can creep in to affect the trial outcome is clearly longer the earlier blinding is lost (and so is maximal if initial treatment allocation is not concealed), but if blinding is lost at any stage before the trial ends then the possibility arises of identifiable biases affecting the result. End-of-trial tests seem, then, very important: if such a test indicates that blinding has been lost, then questions need to be asked about whether any of the biases that lack of blinding makes possible has been actualized and hence has affected the evidential weight properly carried by the trial result.[11]

[11] This is true however the trial works out. A “positive result” might seem to tell in favour of the effectiveness of the experimental treatment under trial (or rather its greater effectiveness than the control treatment) but post-randomization confounders may have instead been responsible. However, some such possible confounders, as we saw, can tell in the opposite direction: for example, concomitant treatment bias (participants who know they are on placebo seeking other treatments for their condition) can help produce a “negative” result even if the experimental treatment is effective (again, more accurately, more effective than control).

The “mug’s game” claim, it seems, cannot be correct. The treatments both of Schulz and Grimes and of Sackett are indeed, as we shall see, based on confusion in this respect. However, underlying those treatments and more or less explicit in them (particularly in the treatment by Schulz and Grimes) are three related important and under-recognised points. These show that the issues of blinding and of tests for blinding are not as straightforward as they might initially seem. Having first explained those underlying insights (Sect. 2.3.1), I then (Sect. 2.3.2) show how Schulz and Grimes and Sackett misrepresent them, fall into some confusion and end up endorsing their unjustifiably negative ‘mug’s game’ view of end-of-trial tests for blinding. In Sect. 3.3, I articulate the more nuanced, but still positive, account of the significance of blinding and of end-of-trial tests for blinding that is really justified by the insights outlined in Sect. 2.3.1. In the final part (Sect. 2.4) I spell out the re-modifications of the CONSORT guidelines that this more nuanced account underwrites.

2.3.1 *Blinding and Its Possible Loss Are More Complex Issues Than Might Appear: The Insights Underlying the Analyses of Schulz and Grimes and of Sackett*

2.3.1.1 Loss of Binding May Not Result from Methodological Sloppiness

However virtuous double-blinding may be from an epistemic point of view, it is not always easy to achieve in practice; and, moreover, once achieved, is not always easily maintained. Loss of blinding may not be the result of any methodological defect in the conduct of the trial; instead, however strong the clinicians’ commitment to the ideal of a double-blinded trial might be, their trial may become unblinded through circumstances beyond their control.

The natural home of double-blinding is the drug trial where it is *relatively* easy to make the control and experimental treatments observably indistinguishable both for the clinicians and participants. This makes allocation concealment, and hence initial double-blinding, reasonably simple; though even here, as we shall soon note, keeping the trial blind after the initial allocation may be an altogether different and more difficult matter. In non-pharmacological trials, even initial double-blinding may be difficult – indeed, in some cases, it may be impossible. To take one minor example, trials have been performed on the relative analgesic effects of real and sham acupuncture. The sham procedure involves an implement sometimes called the ‘Streitberger needle’ featuring a sheath into which the acupuncture needle can retract instead of penetrating the skin. The whole idea is to give the visual impression that the needle has penetrated the skin in the attempt to keep *participants* blind as to which of real or sham acupuncture they receive. But it is clearly difficult in such a trial to blind the administering acupuncturists (though it will be possible, and will clearly be a good idea, to try to ensure that those clinicians assessing outcomes are blind to interventions). To take another, even clearer, example: there is obviously no way that either participants or clinicians can be blinded in trials comparing a vigorous exercise programme with a course of antidepressant pills for the treatment of mild depression (though again outcome-assessors may, and should, be blinded).[12]

Even in its natural home of drug trials, although relatively easily achieved at the outset, double-blinding often cannot be retained as the trial progresses. Laying aside the possibility of outright malpractice *via* gaining access to the randomization code, there are two main ways in which blinding may be lost during the course of an initially blind trial: (a) through tell-tale side-effects and/or (b) through large positive clinical effects. To take an extreme case of the first: suppose that, in a placebo-controlled trial of, let's say, a treatment for migraine, several participants notice that their urine has become significantly discoloured; if so, then those participants and the clinicians involved will very likely, and very probably correctly, conclude that they are not taking the sugar pill, but are instead on the experimental arm of the trial. Suppose, alternatively, that the trial is an "active" one – that is, one in which the control is the currently accepted best treatment for the condition at issue. Because that treatment's side-effect profile will generally be well-known, if a participant experiences some different side-effect, then clinicians, and probably participants too, will likely, and probably correctly, conclude that they are being given the experimental treatment. On the other hand, those participants who had been taking the current treatment beforehand, and who have a similar experience as when knowingly taking the active control treatment, will likely, and probably correctly, conclude that they are in the control group. (And the clinicians who will be monitoring the participants' side-effects are likely to make the same inference.)[13]

So, double-blinding may be lost because of side-effects; it may also be lost because of a clear positive effect of the experimental treatment (though as we shall see, we should strictly speaking talk about "*apparent* positive effects"). To take an unrealistic "philosopher's" example – one that, nonetheless, makes the point: suppose that a placebo-controlled trial is being conducted on an experimental treatment for the common cold; all participants, who are being treated at the same time, have heavy colds of recent onset, but, within 5 minutes of administering the treatments, half of the participants have fully recovered while the other half continue sneezing and snuffling. In the light of background knowledge of the natural history of colds

[12] A supplementary insight is that such trials should not be dismissed as incapable of producing "real" evidence of (comparative) effectiveness just because they were not double-blind trials. Much, as we will reflect in Sect. 2.3.1.3, depends on the (apparent) size of the treatment effect. (I was introduced to the Streitberger needle example by Jeremy Howick.)

[13] It is not quite true that unblinding through side-effects is always outside of the investigators' control. If side-effects can be predicted (or have emerged in earlier-phase trials) then a so-called "active placebo" can, in principle, be developed to reinstate control over blinding. Suppose, for example, that an experimental treatment is known to have the side-effect of discolouring the urine, then an agent can be added to the placebo substance that simply has the effect of discolouring the urine in a similar way. Such "active placebos" are sometimes used, but, where the side-effect is adverse, rather than neutral (persistent headache, say) there are clear ethical obstacles to the addition of a substance to the "placebo" designed to produce that adverse effect. Moreover, placebos cannot, by definition, be designed to be "active" with respect to *unexpected* side-effects.

(and background knowledge about the common cold's positive, but slight susceptibility to the placebo effect), it would be difficult not to form a view as to which participants had been given the experimental treatment, and which were the controls; and difficult to conceive that that view would be (at all substantially) incorrect. Certainly those participants themselves who experience this "miracle cure" and who presumably have all had colds before, will infer that they have been given the experimental treatment. The "hard-line" classical statistician would hold that there is no real "objective" evidence until the (properly-powered) trial has ended (let's assume that the trial design specified that treatment should continue for a week), the data have been analysed and a statistically significant result has been declared – until then we should speak of an *apparent* large positive effect. (This is no doubt why Schulz and Grimes, as well as Sackett, refer, as we shall see, to clinicians' "hunches" about treatment allocation in such circumstances.) But, as not infrequently happens, classical statistics seems to be out of line with educated scientific common sense. Certainly, if, as in this far-fetched case, the effect is marked enough (sadly seldom the case in recent clinical trials), then, given their knowledge of the natural history and of the likely extent of any placebo effect, it will be difficult to prevent clinicians and participants themselves making conjectures about which participants are on the treatment that is going to turn out to be effective. Of course, those conjectures *may* be wrong; but it seems, in a case like this, extraordinarily unlikely that they are.[14]

In sum, as Schulz and Grimes write (2002, 698–99):

> Disproportionate levels of side-effects might provide strong hints as to the intervention. Irrespective of the painstaking efforts to do double-blinded trials, some interventions have side-effects that are so recognisable that their occurrence will unavoidably reveal the intervention received to both the participants and the health-care providers. Even more fundamental than the hints from adverse effects are the hints from clinical outcomes. Researchers usually welcome large clinical effects . . .If they arise, health-care providers and participants would likely deduce . . . that a participant with a positive outcome received the active (new) intervention rather than the control (standard).

2.3.1.2 Loss of Blinding Leads to Possible Bias But Not All Possible Biases Are Actualised

So, blinding throughout a trial is more easily aspired to than achieved. Furthermore, even if blind is not maintained in a trial, this does not at all automatically entail that the outcome was subject to bias. Not every possibility for bias will be actualized. Suppose, for example, that the majority of clinicians involved in a trial have, on the basis of side-effects and/or marked positive effects, correctly "guessed" the treatment allocation for the majority of participants. But further suppose that – as happens in properly-conducted trials – the clinicians who assess the outcomes for each participant are different from those involved in the participants' care during the

[14] There are in fact some trickier issues here concerning the case of *apparent* significant clinical effects. These will be raised, and considered more carefully, *below* pp. 27–29, Sect. 2.3.1.3.

trial, and that, although the clinicians considered as a whole group, are unblinded, those clinicians charged with assessing outcomes can do no better than chance in guessing which treatment each participant was given. If so, then it seems that there can be no question of "observer or reporter bias" having affected the trial's overall outcome.[15]

Or suppose that the outcome measure involved in the trial is "objective" – say, "participant still alive 6 months after treatment" – then, even if outcome assessors have become unblinded, there is no scope for them to introduce "reporter bias" by fudging the outcome for particular participants.[16]

Or suppose, to take a final example, that participants, rather than clinicians, have become unblinded during the course of a trial. This certainly opens up the *possibility* of co-intervention bias: a significant number of those who discover that they are on the placebo arm of the trial may be tempted to seek treatments for their condition outside the trial – an extra drug, let's suppose. But whether or not a participant is taking some particular extra drug can often be tested for – through blood tests for example; if testing reveals that all, or the vast majority, of participants have resisted the temptation to seek additional treatment, then there is no concern about an *actual* co-intervention bias. While these are, perhaps, the clearest-cut cases, similar reassurances *may* be available concerning other potential biases resulting from loss of blinding. If such reassurances are available, then there is no reason to regard the evidential impact of the result of a trial as reduced, on the basis of a "negative" end-of-trial test for blinding.

2.3.1.3 Even If a Trial Has Actually Been Affected by Bias Because of Loss of Blinding, That Trial's Result May Nonetheless Supply Telling Evidence of Effectiveness

Consider a slightly different version of the "toy" example cited earlier. A group of people with heavy colds of recent onset are randomized (at the same time) into experimental and placebo groups. The trial is set to run for 2 hours in total, and the outcome of interest is recovery from the cold. One hour after they are all given a drug (experimental or placebo), roughly half of the whole group exhibit noticeably reduced symptoms; while, at the end of the trial after 2 hours, the participants in the improved half are completely better. The other half have improved barely, if at

[15] It seems not to be widely recognised, but this needn't always be the case. 'Blinded' outcome-assessors (in the sense that they have had no contact with the participants while the trial was running) may nonetheless ask questions about side-effects (indeed one would generally want the trial protocol to have them do so), and hence, if there are differences in side-effect profiles, those outcome-assessors may indeed 'guess' treatment allocations before producing their assessments of individual outcomes.

[16] "In general ... blinding becomes less important to reduce observer bias as the outcomes become less subjective, since objective (hard) outcomes leave little opportunity for bias" (Schulz & Grimes, 2002, 697).

all, and still have pronounced cold symptoms at the end of 1 hour and also at the end of the trial. No doubt the clinicians running the trial will, after the first hour has elapsed, have a good idea of which participants are in which groups; and will therefore have had the opportunity, during the remaining hour of the trial to, for example, give extra attention to those they suppose are taking the experimental treatment. This extra attention may have *some* differential effect, but, given what we know about the natural history of colds and about the (real but) small placebo effect on them, it is surely reasonable, surely "evidence-based" to judge that the effect of the extra attention cannot have been sufficient to have produced an overall result with such a striking effect; and hence not sufficient to invalidate the evidential impact of the result of the trial. That result will still be taken as strong (indeed, in this "toy" case, overwhelming) evidence for the effectiveness of the experimental treatment; and surely correctly so.

Schulz and Grimes (amongst others) talk in this connection of "large clinical effects" as leading to loss of blinding. Strictly speaking, of course, and as noted earlier, what is observed in such cases is an *apparent* clinical effect: a marked difference in the average value of some outcome variable in the two sub-classes of the study population. The "causal" judgement that this difference is the effect of an intervention is always, strictly, a theoretical one. However, so long as the apparent effect size is large, when judged against background knowledge – as it is in my imagined example – then no other explanation of the apparent effect seems remotely plausible.

The medical statistician Stephen Senn has argued even more strongly that, not only is it no problem for blinding to be lost via positive (apparent) clinical effects, it is the *aim* of trials that blinding should be lost in this way. Senn's (2004) was a direct response to the Fergusson et al. (2004) paper cited earlier: "According to Fergusson et al . . ., they [sic] 'consider a trial to be double-blind when the patients, investigators, and outcome assessors are unaware of the patient's assigned treatment throughout the conduct of the trial'. They are quite wrong to do so. The whole point of a successful double-blind trial is that there should be un-blinding through efficacy. That is to say that there should be no incidental reasons, apart from efficacy, as to why the treatments are distinguishable but that the treatments should reveal themselves through efficacy. If the treatments are not distinguishable at all, then the treatments have not been proved different."

It is heartening to hear a medical statistician rejecting the classical statistical hard line that efficacy cannot "reveal itself" during the trial but only after the properly powered and properly planned trial has formally ended and its results have been carefully analysed. Senn clearly goes too far however. First, it is certainly not a necessary condition of a successful clinical trial that it should be "unblinded through efficacy", as Senn suggests it is. There are plenty of very large trials that are generally regarded as positive (null hypothesis refuted at a high level of significance) but which suggest a very small effect of the experimental treatment that would not have been discerned (in any rational way) while the trial was running. Indeed, sadly, very few RCTs have results that are so clear-cut that they would "jump out" at clinicians (as in my imaginary cold example). Moreover, there is an obvious problem with Senn's requirement that "there should be no incidental reasons, aside from efficacy, as to why the treatments are distinguishable". As noted, the loss of blinding undeniably opens up a trial to the possibility of bias. So, there will always be other "incidental" factors that might possibly have an impact on the overall result of the trial. It is only when there is good reason to think that the effect of these factors will be small and the effect size being produced in the trial is large that it is safe to say that there is good evidence that the experimental treatment is superior to the control.[17]

Notice, by the way, that this third consideration points to an important distinction between the two ways in which blinding can be lost during the course of a trial. If the game is given away by unexpected side-effects, in the absence of any apparent large positive effect, then there is always an active worry that the trial may have achieved its positive result as a consequence purely of biases following the loss of blinding. This is especially true because a "positive" result simply means that the null hypothesis of no difference between experimental and control treatments has been "refuted", usually at the 5% level – and this can readily be achieved on the basis of a very small difference in effectiveness between the experimental and control treatments, especially if the trial is a large one. If, on the other hand, blinding is lost because of a large apparent positive effect on one half of the study population then, as we have seen, even if biases have been introduced, the trial outcome may still give good evidential reason to hold that the treatment under trial is indeed effective.

2.3.2 The "Mug's Game" Confusion

So, to summarize Sect. 2.3.1: (a) losing blind may not involve any methodological culpability; (b) losing blind means only that the trial becomes open to the *possibility* of certain kinds of bias, not that it necessarily has actually been affected by any particular bias/es; and (c) even if the outcome of some trial *was* affected by biases, then, at least under some circumstances, it can plausibly be argued that the outcome still gives good empirical support to the theory that the treatment under trial is

[17] Jeremy Howick in chapter 6 of his 2011 – see footnote 6 above – claims that 'When failure to successfully double mask a trial results from the dramatic effects of a treatment, the resulting factors arising from [for example] participant and caregiver knowledge are not confounding." But this is open to analogous objections to those just raised against Senn's analysis. First the assertion that the trial outcome was the result of a "dramatically" effective experimental drug is *not* an observational result, it is instead an interpretation of the outcome based on background knowledge. Secondly, the result *will* very likely be confounded by biases produced by the breaking of blind – in the sense that the result will (likely) be produced by a combination of the (supposed) effectiveness of the experimental treatment with biases resulting from the loss of blinding. The result may well have been statistically significant even if those biases had not played a role, but still the result itself was likely affected. Again, judgment based on background knowledge is required to deliver the view that the evidence, in particular the effect size, gives strong support to the theory that the experimental treatment is effective.

effective – so long as there is evidence-based reason to think that the effect of such biases was small, while the effect size "revealed" by the trial was large.

The first of these insights is explicit, and the second and third arguably implicit, in the treatments of Sackett and, particularly, of Schulz and Grimes. Those insights certainly give grounds for rejecting the naïve, but in fact quite widespread view that a "negative" result in an end-of-trial test entails that the result of the trial should be dismissed as of little or no evidential value (a point which Schulz and Grimes themselves emphasise – (2002, 698)). But the claim that we need to be more sophisticated in interpreting end-of-trial test outcomes, particularly if they are "negative", is quite a long way from the conclusion that to carry out such a test is to play a mug's game. We need to look more closely at how Schulz and Grimes and then Sackett arrived at their negative view of the value of end-of-trial tests for blindness.

First, we have talked blithely about blinding being lost as though this was a clear notion; but in fact it is by no means obvious when exactly blinding should be considered to have been lost in a trial. Assuming that the trial was blind initially, in other words that it began with treatment allocation successfully concealed from both clinicians and participants, then, short of illicit access to the randomization code, loss of blinding is not, objectively, a straightforward yes-or-no affair. Participants and clinicians, as noted, may, and often will, be making conjectures as the trial progresses about treatment assignment on the basis of side-effects or apparent positive clinical effects. But in order to say that blinding has been lost: (i) What proportion of participants or clinicians need to have arrived at conjectures about treatment assignments that are correct?; and (ii) How strong does the evidence for those conjectures have to be (and, relatedly one hopes, how sure do participants and clinicians have to be that their conjectures are correct)?

There is no one, objectively correct answer to these questions, but the issue has been decided by social convention. An end-of-trial test (if performed at all) always has the same form. Clinicians are asked to state which arm – experimental treatment or control treatment – they think each of the participants had been assigned to; the proportion of cases in which the clinicians correctly identify treatment arm is noted; and that proportion is compared to the proportion of times they would be expected to be correct if they were simply making random guesses about each participant's allocation. The trial is declared unblinded just in case the clinicians' identification of treatment arm is better to a statistically significant degree than the chance proportion. So, in the most straightforward case where only two arms (experimental and control) are involved, the trial is deemed to have become unblinded so far as clinicians are concerned if the proportion of their correct "guesses" of treatment-arm is statistically significantly different from 50%. Hence, any consideration of how confident the clinicians are in their "guesses", and of whether, and if so how, those "guesses" are based on evidence drops out of the picture (or is assumed dealt with by the requirement that the difference between the actual percentage of correct guesses and the percentage expected "by chance" is statistically significant).

The corresponding test for *participants* (in fact seldom – or, rather, even more rarely – performed) is obviously to ask them individually at the end of the test to state which arm of the trial they believe themselves to have been on – the test being "failed" and the trial regarded as having become unblinded if (and only if) the proportion of the participants who expressed correct views about their own allocation was statistically better than would be expected "on the basis of chance" – that is, in the simplest, two-arm case, where each "guess" had a probability of ½ of being correct. (There is in fact no uniform consensus concerning when to talk of an end-of-trial test as having been "passed" or "failed", having been "positive" or "negative". I will talk of such a test as "failed" whenever the result is deemed to have shown that the trial has become unblinded, even though this will mean that either clinicians or participants (or both) *succeeded* in identifying treatment arm at better than chance rates. To emphasise that this might seem odd (but then so would the opposite convention), I will continue to place "failed" (or "passed") in quotation marks.)

We arrive, then, at the following characterisation of when a trial has become unblinded (reflecting the obvious fact that this cannot apply only when a test for unblinding has actually been carried out): a trial has become unblinded if and only if either the number of clinicians who have correct conjectures about which participants are on which arm (or who would have such conjectures if they were explicitly asked to say which arm each particular participant was on), or the number of participants who have correct conjectures about which arm they themselves are on (or, again, would have such conjectures if they were explicitly asked) is statistically significantly different from what would be expected if they were "merely guessing" – in the simplest case: if they had a probability of 1/2 of being correct each time.

Both Sackett and Schulz and Grimes explicitly adopt this characterisation. Schulz and Grimes, for example, write (2002, 698): "Investigators can theoretically assess the success of blinding by directly asking participants, health-care providers, or outcome assessors which intervention they think was administered … In principle, if blinding was successful, these individuals should not be able to do better than chance when guessing the intervention …". They immediately add, however: "In practice, however, *blinding might be totally successful* [emphasis supplied], but participants, health-care providers and outcome assessors might nevertheless guess the intervention because of ancillary information … If indeed the active (new) intervention materialises as helpful … then [the clinicians' and participants'] deductions would be correct more often than chance guesses. Irrespective of their suspicions, end-of-trial tests for blindness might actually be tests of hunches for adverse [side-] effects or efficacy" (Schulz & Grimes, 2002, 698–699).

This is mysterious: first, it is not clear what precise "in principle"/"in practice" distinction is being appealed to, and, more especially, it is completely unclear how blinding can have been "totally successful" if participants and/or clinicians have been able to "guess" treatment assignment predominantly correctly. What exactly is the distinction between testing to see if blinding has been lost and testing to see if adverse side-effects and/or apparent effectiveness have given the treatment assignment away? Short of illicit access to the randomization code, side-effects and apparent positive effects are the (exhaustive, but non-exclusive) ways in which blinding can be lost; and so *of course* in testing for loss of blindness you are at the same time testing for adverse side-effects and/or (apparent) efficacy.

One conjecture might be that, by "successful blinding", Schulz and Grimes mean only that the initial treatment allocation was successfully hidden. But they themselves point out "[while] many medical researchers confuse the term blinding with allocation concealment… [In fact] the term blinding refers to *keeping* trial participants, investigators . . . . or assessors . . . unaware of an assigned intervention, so that they are not influenced by that knowledge" (Schulz & Grimes, 2002, 696). Their statement (2002, 698) that "blinding might be totally successful, but participants, health-care providers and outcome assessors might nevertheless guess [predominantly correctly] the intervention because of ancillary information" seems, then, to be an outright logical contradiction.

Of course, this cannot be what they intended since no one intends to assert a logical contradiction. But further consideration of what Schulz and Grimes might really have thought can be laid aside, because there is surely only one sensible claim here: that blinding may have been completely successful *in the sense that the clinicians organising the trial did everything that they could to maintain blindness*, despite which adverse side-effects and/or clear positive clinical outcomes gave the game away, and hence the trial became unblinded. This hardly helps their argument, however, since this sensible claim clearly does *not* entail their conclusion that end-of-trial tests are of questionable value. To the contrary: independently of how it happened, the fact that blinding has been lost during a trial entails (as they themselves clearly recognize) that identifiable biases *may* have affected its outcome. It then follows that, far from an end-of-trial test being valueless, significant information is in fact provided whatever such a test's result might turn out to be (assuming the result to be genuine, a point which I will take up soon). If the test indicates that blinding was maintained during the trial, then there is no reason to investigate whether certain specifiable "post-randomization" biases may have affected the result of that trial, and so its outcome can carry its full evidential weight concerning the effectiveness or ineffectiveness of the experimental treatment; if, on the other hand, the test indicates that blinding *wa*s broken, then there are reasons to check for biases – and so, before the result of the trial can be given its full evidential weight, questions need to be answered about whether or not biases did in fact affect that result.

To reiterate what was said above in Sect. 2.3.1.2, if an end-of-trial test provides evidence that blinding was lost during the trial, then this does *not* automatically entail that the trial was in fact biased – only that it might have been. And if the trial was in fact affected by bias, this does not automatically entail that its result cannot provide telling evidence for the effectiveness of the treatment under trial. But these are reasons for interpreting end-of-trial tests and their outcomes sensibly, not for rejecting such tests as without value.

Although, as we shall see, he ends in the same confusion as do Schulz and Grimes, David Sackett's argument for the lack of value of end-of-trial tests is worth separate examination for reasons that will become apparent. Sackett's most sustained and pointed version of the argument is in his 'Clinical Round' paper (2011); and it is there that he explicitly branded performing end-of-trial tests as playing a "mug's game". He presented the argument in that paper alongside reminiscences of his involvement in what came to be known as “The Canadian Aspirin Trial” – a trial now regarded as having been the first to provide evidence that a regular dose of aspirin reduces the risk of strokes and death among people who have suffered from transient ischemic attacks (“mini-strokes”).

That trial was a comparative one involving, not just aspirin, but also another potentially effective drug – sulfinpyrazone; and it used a “double-dummy design”, which involved participants being randomized into *four* groups – those in the first were given both experimental drugs (aspirin and sulfinpyrazone), those in the second received sulfinpyrazone plus placebo (“aspirin-placebo”, a “dummy” pill intended to mimic aspirin), those in the third got aspirin plus (sulfinpyrazone-) placebo and those in the fourth were given both the placebos.

Sackett records (2011, 674) that, as the results emerged, he was “dreaming of a lead article in the *New England Journal of Medicine*” (the world’s most prestigious medical journal), largely because the overall outcome was turning out to contradict pre-trial expectations: humble aspirin had not been thought likely to reduce the risk of stroke but the trial had found evidence that it did, whereas sulfinpyrazone was believed, ahead of the trial, to be much the more likely to prove effective, but in fact the trial found no evidence of this.

Sackett recalls that one of the “remaining odds and ends” before a paper could be sent off for publication was the analysis of the results of the end-of-study questionnaire that had been used to check that the clinicians involved in the trial had remained blind to which patients were in which groups. This – in line with usual practice, as noted above – had involved asking the clinicians involved to state which group they thought each participant had been in. Sackett and collaborators assumed that, because the trial involved four equal-sized groups (remember the trial had a “double-dummy” design), 25% correct “guesses” would be expected on the basis of chance. Sackett continues the story (2011, 674):

> ‘I felt the bullet enter my heart’ when our . . . statistician tracked me down on the ward to tell me that our clinicians’ correct guesses were . . . statistically significantly different from 25%. Had our triumphant lead article been reduced to an apologetic Letter to the Editor? And why did my [statistician colleague] have a big grin on his face?

The reasons for the grin, apparently, were (a) that the clinicians’ predictions about which participants were in which groups were statistically significantly *wrong*; and (b) those clinicians’ pre-result predictions, also recorded by the statistician, about which of the two “active” drugs – sulfinpyrazone or aspirin – if either, would prove effective were equally wrong: “[m]ost of them predicted that aspirin would be worthless but sulfinpyrazone would be effective” (Sackett, 2011, 675). So, most clinicians had assumed, in line with their prior conjectures, that those participants whom they identified (notice: identified predominantly correctly!) as faring better in the trial were likely to have been given sulfinpyrazone – an assumption that, it turned out, was at 180 ° to the actual result of the trial.[18]

Sackett spins a fine yarn. But why exactly is its moral supposed to be that in testing for blindness at the end of a trial you are playing a "mug's game"?

First, notice that what Sackett takes to be the crucial aspect of his story – namely that the guesses of the clinicians in his study were statistically significantly *wrong* in terms of which drug, aspirin or sulfinpyrazone, was the effective one – in fact carries no methodological weight whatsoever. The important methodological question is whether those clinicians could, while the trial was still ongoing, distinguish (predominantly correctly) between participants on the treatments that turned out to be effective (whichever they turned out to be – as it happened, aspirin with either placebo or sulfinpyrazone) from those participants on "ineffective" treatments (whichever *they* turned out to be, as it happened, sulfinpyrazone with placebo and placebo with placebo).[19] According to Sackett's story, the test revealed that the clinicians could indeed make that distinction at better than chance rates. It makes no difference from the methodological/evidential point of view if they also (predominantly and now incorrectly) believed that the effective drug would turn out to be sulfinpyrazone rather than aspirin. All that counts is that, since they had, before the end of the trial, (predominantly correctly) identified which participants were on the effective drug, the trial was opened up to all the possibilities of bias we have noted – subsidiary treatment (or "co-intervention") bias, reporter bias and the rest. Although I am happy that he was happy, the grin on the statistician's face was completely misplaced.

Sackett, however, sees the mistaken identification of the likely effective treatment as crucial: "With a 'prior' belief that sulfinpyrazone was effective, when a patient fared well throughout the trial, it was clinically sensible for their neurologist to suspect that they were on it. Similarly, if a patient suffered a stroke during the trial, it was clinically sensible for their neurologist to suspect the double placebo or the aspirin they thought was probably worthless. Thus, our end-of-study test for blindness was exposed as a test for (incorrect [sic]) hunches about efficacy . . . So, the first 'pearl' on offer . . . is that testing for blindness at the end of your trial is a mug's game, because it cannot distinguish the failure of your blinding tactics from their correct [sic] guesses about which treatment was received, based on their experiences of pharmacodynamics, side-effects and trial outcome" (Sackett, 2011, 675).

Taken literally, Sackett here ends in essentially the same inconsistency as Schulz and Grimes did. If clinicians and/or participants are able to guess treatment

[18] There is a story to tell about whether the 25% model is correct in this case, but it would only complicate matters unnecessarily to tell it here. The general points about unblinding and end-or-trial tests are independent of that issue.

[19] This is what suggests that the 25% hypothesis was the wrong model. But. to repeat, we shouldn't allow this to spoil the story. Let's assume that Sackett is right that blind was broken, the question at issue is what follows logically from this assumption and this question about logical entailment is, of course, independent of the question of whether or not the assumption is true in the case of some particular trial.

assignment at significantly better than chance rates then, by definition, the trial has become unblinded (despite initial hidden treatment allocation). Hence, except when blind has been broken by illicit access to the randomization code, how *could* an end-of-trial test differentiate the success of their "blinding tactics" from the issue of whether clinicians developed predominantly correct conjectures about which participants were on the effective treatment? If the end-of-trial test finds that the trial has become unblinded, then it would seem that, by definition, the clinicians' blinding tactics have failed (albeit through no fault of their own).

In order to free Sackett's analysis from contradiction, some meaning needs to be found for the phrase "the blinding tactics were a success" that does not entail that blinding need be maintained (and hence an end-of-trial test "passed"). Sackett in fact provides interesting detail on just how painstaking the efforts to keep a trial blind may be in carefully planned trials. He identifies three potential biases (a) contamination of the comparison [control] group with the experimental treatment, (b) cointervention bias and (c) outcome assessment bias. Possible bias (c) was avoided by using a "panel of adjudicators who are blind to treatment" (so they at least would "pass" any end-of-trial test – that is, do no better than chance in identifying treatment allocation). As for (a) and (b) "treatment allocations had been concealed; active drugs and their corresponding placebos were identical in size, color, taste, smell and flotation; we'd told everybody to use acetaminophen for pain; we'd purged uric acid results from all lab reports (sulfinpyrazone is uricosuric); and we'd kept our periodic platelet function test results secret" (Sackett, 2011, 675).

Not all of the considerations that Sackett raises are in fact to do with biases that might be introduced if blind is broken. For example, having clinicians use acetaminophen rather than aspirin to treat any patients in the trial who were in pain was obviously aimed at preventing participants who had been assigned to one of the non-aspirin groups obtaining aspirin as subsidiary treatment alongside the non-aspirin treatment to which they had been assigned. But, for example, not allowing the clinicians access to uric acid results on participants – results which would have given grounds for inferring which participants were being given sulfinpyrazone (which, as Sackett points out, is uricosuric) – *was* an attempt to prevent blind being broken. This is a good illustration of another, subsidiary insight: that if certain side-effects are *expected* from any treatment involved in a trial, then there may be ways in which they can be concealed. If such ways exist, then it is surely good methodological practice to employ them.

Similarly, Sackett makes it clear that sophisticated triallists will pay great attention to making the various treatments similar not only in looks but in other ways that might not initially be considered but which might allow inquisitive clinicians (or participants) to differentiate them – for example, if two types of drug looked the same but smelled differently; or if one type "sank" while the other "swam" when thrown into water (usually in the toilet!).

So, it could be that when claiming that "blinding tactics may be successful" even though blind was broken, Sackett had in mind that all the precautions that were aimed at keeping the trial blind *in so far as possible* were successful. But, even on this reading of his claim, it obviously fails to follow that testing for blindness is a "mug's game". The fact – as we are taking it to be – that, despite all the clever precautions taken in advance, blind was broken in Sackett's Canadian Aspirin trial is important information. It means that a proper methodological analysis of the impact of the result of the trial requires questions to be asked about whether or not the "knowledge" involved in the breaking of blind of which participants were on the effective drug arrived at before the trial had ended had any biasing effect on the eventual outcome. Had the result of the end-of-trial test been "positive" (that is, had the clinicians done no better than chance in discerning treatment assignment) then those questions did not need to be considered. As before, it is certainly important to recognise that "failure" in such a test does not entail that the result was biased, only that it might have been; and, also as before, it is important to recognise that, even if biases did creep in, this does not entail that the outcome can give no solid evidence of effectiveness. But neither of these points means that performing end-of-trial tests is of little or no value.

Sackett's story had a happy ending: his study was published by the *New England Journal of Medicine* (albeit not as the lead article, but certainly as a full article, not the feared Letter to the Editor – (The Canadian Cooperative Study Group, 1978). Sackett records (2011, 675) that "We successfully explained [the situation] to the journal's Editor (if not to one persistently confused referee) and got the trial published." Since what they "explained" to the Editor was that their study "exposed" end-of-trial tests as a test for hunches about side effects and efficacy, and since the assumption that this means there is no need to worry about loss of blinding is itself based on confusion, it seems that it was the Editor that they confused. And it seems likely that the "persistently confused referee" was in fact seeing things clearly: Sackett's colleagues, according to his own story, had become unblinded during the course of the trial and so questions should have been asked about biases, which is not to say, to repeat one final time, that the result *was* biased, only that it might have been.

Sackett develops a secondary argument for downgrading end-of-trial tests for blindness: that revealing the fact that blinding was lost may affect the *perceived* impact of the trial's result. Remember that he "felt the knife go into his heart" when the statistician told him the end-of-trial test had been failed – his reaction being "Had our triumphant lead article been reduced to an apologetic Letter to the Editor?" (Sackett, 2011, 674). Sackett is clearly assuming here that the evidence from a trial known to have become unblinded would have its impact immediately downgraded by the clinical research community in general. And performing tests that are likely to downgrade the impact of your study, *for no good reason*, does indeed seem like a mug's game![20]

[20] Indeed, reintroducing the confusion about the allegedly crucial nature of the fact that the prior view of the clinicians was that it would be sulfinpyrazone rather than aspirin that would turn out to be effective, Sackett envisages a "bone-chilling alternative ending [to his Canadian Aspirin story]. What if the neurologists in our … trial had begun it with the reverse set of hunches, this time thinking that aspirin would probably work and sulfinpyrazone probably wouldn't? Testing for blindness at the end of that trial – forcing us to weaken our conclusions about efficacy and dashing

The qualifier "for no good reason" is, however, clearly crucial here. If trial clinicians were to seek to avoid performing an end-of-trial test knowing that trial had become unblinded and that consequently biases had indeed crept in substantially to affect the result, then this would not count as reasonably refusing to play a mug's game, but rather as evading one's responsibilities to good scientific evidential practice. Presumably Sackett believed that the Canadian Aspirin Trial had been so carefully planned, and the effect-size revealed by the trial large enough (though the latter is arguable – to say the least), that it was legitimate to rule out biases introduced by the acknowledged loss of blinding as an alternative explanation of the observed outcome. But even supposing this is true, to react by in effect recommending that end-of-trial tests are *never* done seems an obvious and significant error: many shoddily performed more recent trials could have hidden (and perhaps have hidden) behind Sackett's "mug's game" claim. The correct reaction is instead to challenge the prevalent community view that loss of blinding immediately and automatically adversely affects the weight of the trial result; and carefully explain why, in your trial at least, provisions were in place to keep the outcome effectively bias-free despite the loss of blinding. It is surely never a good idea to collude in a mistaken view just because it has become socially entrenched. Instead one should aim to re-educate the community and so change its view. But the community's confusion is not to do with whether or not the test separates blinding and side/positive effects (it doesn't because it can't), but instead stems from the assumption that a trial that has become unblinded automatically fails to supply good evidence for the effectiveness of the experimental therapy.

Schulz and Grimes also present a secondary argument, related to, but slightly different from Sackett's. Consideration of this argument will finally lead to a coherent (and dispiriting) concern about the value of end-of-trial tests. Schulz and Grimes write: "Furthermore, individuals [when being quizzed as part of an end-of-trial test] might be reluctant to expose any unblinding efforts by providing accurate responses to the queries – in other words, if they have deciphered group assignments, they might provide responses contrary to their deciphering findings to disguise their actions. That difficulty, along with interpretation difficulties stemming from adverse side-effects and successful clinical outcomes, leads us to question the usefulness of tests of blinding …" (2002, 699). Laying aside the "interpretation difficulties" as already shown to be confused, Schulz and Grimes are suggesting that clinicians who have in fact discerned which participants are on the experimental treatment have an incentive to hide this fact by not giving true responses to the questions in an end-of-trial test. This is because they fear that if they give accurate responses and hence the end-of-trial test reveals that their trial had become unblinded, then the medical

our hopes of prominent publication – certainly would have been a mug's game!" (Sackett, 2011, 676). This shows how deep the worry that the community will judge on the basis of "unblinded therefore as good as worthless" in forcing them "to weaken [their] conclusions about efficacy". It also, however, aexhibits just how confused Sackett's analysis is: the real story (sulfinpyrazone thought likely to be effective) entails exactly the same concern with lack of blinding as the "blood curdling" alternative.

community is likely to dismiss their trial result as carrying little or no (or, at any rate, greatly reduced) evidential weight. If so, then we cannot rely on the accuracy of end-of-trial tests and this seems – at last – like a good reason to question the utility of such tests.

Schulz and Grimes suggest that clinicians might dissemble in response to an end-of-trial test to cover up the fact that they had correctly discerned treatment allocation (at least to some significant extent). But, again, there are two importantly different reasons that a clinician might have for dissembling. *First* she might know (or, more accurately, have good evidence) that various aspects of the trial protocol made it difficult for bias to affect the result and have good evidence that, if any biases *have* intruded, they would have had only a small effect at most – an effect insufficient to produce a false positive result in the clinical trial. By covering up the loss of blinding, such a clinician, aware of the attitude toward non-blinded trials prevalent in her community, would be trying to avoid an automatic downgrade of the evidential value of their trial's result – a downgrade that she judges (probably implicitly) to be unjustified in this case, and judges it to be unjustified on the basis of evidence. The response to this possibility is the same as the one just given to Sackett's similar but somewhat different concern: namely, that there would be no reason for providing misleading test responses if the clinical trials community were educated into more enlightened evidential practices – into accepting that, in appropriate circumstances, the outcome of a trial that has become unblinded can still provide significant evidence of effectiveness. It would seem, then, that the best advice to such a well-intentioned clinician is to respond truthfully to the end-of-trial questionnaire (as Sackett's colleagues presumably did) and, at the same time, try to re-educate the community.

In this first case, so we are supposing, any bias that affected the trial's outcome as a consequence of unblinding was minimal and so the fact that the end-of-trial blinding test was fudged provides no significant reason to downgrade the evidential value of that outcome. Hence, if we could be sure that all cases of dissembling on end-of-trial tests were like this one, then this possibility of dissembling would form no sort of argument for failing to perform such tests. Not everyone will dissemble, and if they don't, then the test supplies accurate and useful information; while if they do dissemble, then, although the result of the blinding test will be inaccurate, no real evidential harm is done so far as the impact of the outcome of the clinical trial is concerned.[21]

However, there is a *second* reason why a clinician might be motivated to cover up loss of blinding in her trial and hence dissemble when subjected to an end-of-trial test; and this reason is undoubtedly problematic. Clinical triallists generally are under great pressure to produce "positive" results – i.e. outcomes that involve refutations of the null hypothesis of no difference between experimental and control treatments and hence are taken to support (sometimes even to 'establish') the

[21] Except, as indicated earlier, that an opportunity has been missed to help eradicate an evidential mistake (that unblinded trials are automatically non-telling).

effectiveness of the experimental treatment. Despite being exposed any number of times as nonsense, the implicit belief that "Not statistically significant means insignificant" – i.e. that "negative results" are of no interest – still seems deeply-embedded in the collective medical psyche. For many years it was next to impossible to get a study with a "negative" result published in medicine. And, although editors have recently wised up somewhat, the problem has by no means entirely disappeared. An even greater worry is the influence of "Big Pharma" which of course has a vested interest in clinical trials yielding "positive results" for any treatment it wants to market. Anyone who thinks this is an exaggerated worry should rush to read (Angell, 2004). Angell was on the editorial staff of *The New England Journal of Medicine* for over 20 years, rising to Editor-in-Chief and writes (2004, xxvi–xxvii): "As I saw industry influence grow, I became increasingly troubled by the possibility that much published research is seriously flawed, leading doctors to believe new drugs are generally more effective and safer than they actually are". So, suppose such an investigator under pressure to produce a "positive" result has discerned treatment allocation during the course of a trial and has used that information to bias the trial outcome – by for example giving special care and attention to those she has (it turns out correctly) conjectured are on the experimental treatment. Such an investigator would have a very good reason to dissemble when completing end-of-trial questionnaires to test for the retention of blinding.

This is certainly worrying. But, although it might provide reason to doubt any "all clear" produced by an end-of-trial test for blinding, it is clearly not at all specific to such tests: it simply adds to the (depressingly long) list of reasons to be suspicious of clinical trial results, especially when financed (and often controlled) by pharmaceutical companies. If the possibility that clinicians will give inaccurate replies to the questions in an end-of-trial test is to be taken as a reason to downgrade the significance of such tests and to leave them out of the guidelines for good practice (as in CONSORT), then, to be consistent, all the usual requirements for good practice should be similarly downgraded. Might clinicians lie when reporting that their trial was properly randomized? Of course they might. Might clinicians lie when reporting that initial treatment allocation was hidden from all clinicians and participants? Of course they might. Might clinicians report outcomes for individual participants inaccurately? Of course they might. But no one, I take it, would infer from these possibilities that imposing requirements such as proper randomization, hidden treatment allocation, accurate reporting of data, etc., is playing a series of mug's games. There is nothing special then, in this regard, about requiring that end-of-trial tests for blinding are performed and their results reported.

Science relies on the honesty of scientists. Of course, self-correcting mechanisms are sometimes in play – principally repetition of the trial by different scientists; but in general it just needs to be taken as the default position that the scientists reporting a result have acted honourably and reported accurately – not allowing the mere possibility of malfeasance to detract from the impact of published results. The fact that clinicians may not respond truthfully to end-of-trial test questionnaires is worrying but is no reason at all for a general downgrade of the methodological value of such tests.

Summarizing, then, my response to the analyses by Schulz and Grimes and by Sackett: the "hunches" issue supplies, on analysis, no coherent reason at all to devalue end-of-trial tests for blinding; and downgrading them is the wrong response to the secondary worry about the reputation, and hence evidential impact, of the trial if such a test is "failed". Instead of holding that performing such end-of-trial tests is a "mug's game", the correct inference from the valid points that we identified earlier in Sect. 2.3.1 as underlying those analyses is surely that it is always a good idea to do end-of-trial tests on trials that begin double-blind but that we should be careful not to over-interpret the results of such tests – in particular, if they turn out to be "negative". More exactly: if the result of the end-of-trial test is that clinicians or participants do not do significantly better than chance in "guessing" treatment assignments there is no issue of post-randomization bias (unless they are dissembling – and we cannot let this mere possibility instil a generalised scepticism). If, however, the result is that clinicians and/or participants do perform at significantly better than chance, then thought needs to be given to the question of whether the biases that breaking of blind makes possible have plausibly been actualized. If there is indeed reason to think that biases have crept in, then the impact of the evidence from the trial's result should be duly downgraded. But if, on the contrary, either (a) the protocol of the trial left unblinded clinicians or participants with little or no scope to use their knowledge of treatment assignment to introduce bias, or (b) the effect size "revealed" by the trial was large and background knowledge endorses the view that biases could, at worst, have contributed only little to that overall effect, then the trial result should still carry its full evidential impact.

2.4 The CONSORT Guidelines Re-revised?

It was always a mistake for the CONSORT guidelines to be as lukewarm about blinding and about tests for the retention of blinding as they were; and the 2010 revision, by omitting mention of end-of-trial tests entirely, made matters worse. The above analysis, if correct, surely mandates a re-revision of these guidelines in this regard.

The current recommendation concerning blinding reads (*www.consort-statement.org*).

Blinding: If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how.

An additional clause (in bold below) seems, however, to be required if an accurate assessment is to be made of the weight of evidence provided by a trial's outcome:

Blinding: If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how; **if not done, then reasons should be provided for why the biases that lack of blinding makes possible can reasonably be thought, on the basis of the protocol of the**

trial, of background knowledge and of other evidence, to have had at most a minor effect on the trial's outcome.

As for tests for the retention of blinding, the recommendation in the original version of CONSORT was conditional (essentially: "if such tests were done, state who (participants, clinicians, outcome assessors ..) were tested"); while the deliberate omission of any mention of such tests in the 2010 revision surely encouraged the belief that those tests have little or no role in assessing the weight of evidence from the result of the trial, and hence the belief that performing them is barely, if at all, worthwhile. As we have seen, this omission was based on confused thinking, and in fact tests for blinding should ideally always be done (unless the trial is from the outset an unblinded study). This is because those tests always deliver information that is important in the assessment of weight of evidence: if the test is performed and the outcome of the test is "positive" (that is, the trial remained blinded throughout) then the result of the trial carries full evidential weight; if the test is carried out and it is "negative" (that is, blind was likely to have been broken), then issues arise about *possible* biases having affected the result.

As mentioned earlier, the CONSORT team deliberately restricted itself to setting guidelines for how clinical trials should be *reported* – rather quixotically, it might be thought, since how trials are performed is clearly a more important issue than how they are reported (even though the two, of course, interrelate in important ways). However, even restricting attention to the reporting of trials, an entry for 'Tests for Blindness' should surely be reinstated in the guidelines; and the text might read as follows:

Tests for Blindness (At the End of the Trial) If done, then who was tested for retention of blindness (for example, participants, care providers, those assessing outcomes), and with what result; if either not done, or done and the outcome was "negative" (blinding not retained), then state what aspects of the protocol, if any, or of the result provide warrant, given background knowledge, for thinking that the biases that lack of blindness makes possible did not in fact affect the outcome of the trial or affected it only to a negligible degree.

Although it is implicit in this wording, it might be best to make the requirement explicit that trialists should also ideally report, in the case that the end-of- trial test reveals loss of blinding, on whether the people questioned based their discernments of treatment arm on unusual side effects or on large (apparent) positive effect (or on both). This is because, as noted earlier (but invariably ignored in medicine), quite different considerations apply to the two ways in which blind may be broken. If this has happened entirely through noting unusual side-effects, then, there was no clear indication ahead of the loss of blindness that the treatment under trial is effective. Hence the "positive" result in the trial may be entirely the product of the biases that loss of blindness has allowed – especially if the effect size is small and the sample size large. On the other hand, if loss of blinding was brought about by a "large positive effect" of the treatment under trial, then, although strictly only an "apparent" effect, it may well be reasonable to judge on the basis of evidence-based background knowledge, that possible biases introduced in the wake of loss of blinding could not alone have produced an effect of anywhere near that size.

I am sorry that, if implemented, this would make the guidelines more complicated; but accuracy is of course the dominant virtue here. Albert Einstein famously averred: “Physics should be as simple as possible, but not more so”. Guidelines too.

References

Angell, M. (2004). *The truth about the drug companies: How they deceive us and what to do about it* (Paperback ed.). Random House.

Benedetti, F. (2014). *Placebo effects* (2nd ed.). Oxford University Press.

Boutron, I., Estellat, C., & Ravaud, P. (2005). A review of blinding in randomized controlled trials found results inconsistent and questionable. *The Journal of Clinical Epidemiology, 58*, 1220–1226.

Ferguson, D., Glass, K. C., Waring, D., & Shapiro, S. (2004). Turning a blind eye: The success of blinding reported in a random sample of placebo controlled trials. *The British Medical Journal, 328*, 432–436.

Howick, J. (2011). *The philosophy of evidence-based medicine*. Wiley-Blackwell (BMJ Books).

Howson, C., & Urbach, P. M. (2004). *Scientific reasoning: The Bayesian approach* (3rd ed.). Open Court.

Hrobjartsson, A., Forfang, E., Haahr, M. T., et al. (2007). Blinded trials taken to the test: An analysis of randomized clinical trials that report tests for the success of blinding. *The International Journal of Epidemiology, 36*, 654–667.

Locke, J. (1689). *An essay concerning human understanding*. Thomas Basset.

Ney, P. G., Collins, C., & Spensor, C. (1986). Double blind: Double talk or are there ways to do better research? *Medical Hypotheses, 21*(2), 119–126.

Noseworthy, J. H., Ebers, G. C., Vandervoort, M. K., Yetisir, E., & Roberts, R. (1994). The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology, 44*, 16–20.

Sackett, D. L. (2004). Why we don’t test for blindness at the end of our trials. *The British Medical Journal, 328*, 1136.

Sackett, D. L. (2011). Clinician-trialist rounds: 6. Testing for blindness at the end of your trial is a mug’s game. *Clinical Trials, 8*, 674–676.

Schulz, K. F., & Grimes, D. A. (2002). Blinding in randomized trials: Hiding who got what. *The Lancet, 395*, 696–700.

Schulz, K. F., Altman, D. G., Moher, H., & for the CONSORT group. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *The British Medical Journal, 340*, 698–702.

Senn, S. J. (2004). A blinkered view of blinding. *The British Medical Journal, 328*, 1135–1136.

The Canadian Cooperative Study Group. (1978). A randomized trial of aspirin and sulfinpyrazone in threatened stroke. *The New England Journal of Medicine, 299*, 53–59.

Urbach, P. M. (1993). The value of randomization and control in clinical trials. *Statistics in Medicine, 12*(15–16), 1421–1431.

Worrall, J. (1988). The value of a fixed methodology. *The British Journal for the Philosophy of Science, 39*, 263–275.

Worrall, J. (2006). Why randomise? Evidence and ethics in clinical trials. In W. J. Gonzalez & J. Alcolea (Eds.), *Contemporary perspectives in philosophy and methodology of science* (pp. 65–82). Netbiblo.

Worrall, J. (2007a). Why there's no cause to randomize. *The 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.

Worrall, J. (2010). Evidence: Philosophy of science meets medicine. *Journal of Evaluation in Clinical Practice, 16*(2), 356–362.