

# Profitable failure: antidepressant drugs and the triumph of flawed experiments

LINSEY MCGOEY

## ABSTRACT

Drawing on an analysis of Irving Kirsch and colleagues' controversial 2008 article in *PLoS* [Public Library of Science] *Medicine* on the efficacy of SSRI antidepressant drugs such as Prozac, I examine flaws within the methodologies of randomized controlled trials (RCTs) that have made it difficult for regulators, clinicians and patients to determine the therapeutic value of this class of drug. I then argue, drawing analogies to work by Pierre Bourdieu and Michael Power, that it is the very limitations of RCTs – their inadequacies in producing reliable evidence of clinical effects – that help to strengthen assumptions of their superiority as methodological tools. Finally, I suggest that the case of RCTs helps to explore the question of why failure is often useful in consolidating the authority of those who have presided over that failure, and why systems widely recognized to be ineffective tend to assume greater authority at the very moment when people speak of their malfunction.

*Key words* ambiguity, antidepressants, clinical trials, strategic ignorance, valuable failure

## INTRODUCTION

Before a packed lecture hall at the Institute of Psychiatry, King's College, London, four men and one woman sat on a podium, watching the crowd with uneasy smiles. The speakers were gathered for the 36th Maudsley debate, which took place on 4 June 2008. Named for Henry Maudsley, a pioneering British psychiatrist and founder of the hospital in South London that bears his name, Maudsley debates are a chance for practising clinicians, medical students and patients to listen to world-renowned experts discuss some of the most controversial issues in psychiatry today.

On 4 June the audience was gathered to hear two eminent British scientists, Lewis Wolpert and Guy Goodwin, address claims by an equally eminent American psychologist named Irving Kirsch about the usefulness of antidepressant drugs. 'The Drugs Don't Work' was the title of the debate. The motion was: 'The house believes that antidepressants are no better than placebo'. Arguing in favour was Kirsch, who had published a study in *PLoS* [Public Library of Science] *Medicine* four months before which suggested antidepressants such as Prozac are no more clinically efficacious than placebo in treating all but the most severe forms of depression. He was joined by Joanna Moncrieff, a clinical psychiatrist, senior lecturer at University College London (UCL), and recent author of *The Myth of the Chemical Cure*, a book that contests dominant perceptions of the mechanistic action of antidepressant drugs.

Arguing against the motion was Guy Goodwin, head of the Department of Psychiatry at Oxford University and former president of the British Association for Psychopharmacology, and Lewis Wolpert, emeritus professor of biology at UCL. Wolpert is a public figure in the UK, known for publishing his own experiences of depression in the book *Malignant Sadness: The Anatomy of Depression*.

The stakes of the Maudsley debate were high. Drugs such as Prozac have been developed, licensed and prescribed on the basis of the same methodological tools used to measure the safety and efficacy of all psychotropic drugs. If, as Kirsch suggested, psychiatrists had somehow overestimated the usefulness of Prozac, then they may have misperceived the usefulness of any number of routinely prescribed treatments.

Using the debate as a window into a disciplinary battle often reserved to sparring factions within psychiatric departments, this article explores the ramifications of recent debates over the efficacy of antidepressants. I suggest that, for those who insist on the usefulness of antidepressants and those who doubt it, each group is aware of methodological limitations that have made it difficult to determine the value of the drugs. This shared awareness of their own methodological limitations – their knowledge of their own ignorance – leads to conflict over how to address methodological shortcomings, and

ultimately to the perpetuation of the very limitations that have led to conflicting views on the drugs in the first place.

The structure of the article is as follows. First, I provide a summary of the Maudsley debate. Second, drawing on Kirsch's recent article on the efficacy of antidepressants, as well as work by Andrew Lakoff, David Healy and Ted Kaptchuk on the construction of clinical trials, I explore problems with the use of rating scales in trials for psychotropic drugs, as well as problems surrounding the recruitment of patients into antidepressant trials. Third, I explore how researchers and clinical trialists who are aware of the limits of RCTs must collaborate in ignoring or obfuscating those limitations in order to meet regulatory demands for scientific precision and specificity. Finally, drawing analogies to work by Michael Power, I suggest it is the very deficiencies of RCTs – their inadequacies in producing reliable evidence of clinical effects – that help to strengthen popular and scientific assumptions of their superiority as methodological tools.

Methodologically, the article draws on participant observation at numerous psychiatric meetings and conferences such as the Maudsley debate; textual analysis of the records of UK and US regulatory inquiries into the safety of antidepressant drugs; and over 20 in-depth interviews with psychiatrists, psychopharmacologists and regulators involved with efforts to determine the safety of selective serotonin reuptake inhibitors (SSRIs), from experts such as Kent Woods, CEO of the UK's Medicines and Healthcare Products Regulatory Agency (MHRA), to Tim Kendall, co-director of the UK's National Collaborating Centre for Mental Health, to Michael Shooter, former president of the UK's Royal College of Psychiatrists.

### 'I'D LIKE TO PUT MY TONGUE OUT TO ALL OF YOU': PASSIONATE SCIENCE AT THE MAUDSLEY

The Maudsley debate on 4 June lasted for just over an hour and half. The speakers were each granted 10 minutes to make their case, followed by questions from the audience. Below, I offer a summary of each presentation. Later on, I expand on some of the common points that emerged, such as questions over the usefulness of psychiatric rating scales, recruitment pressures, and problems surrounding the placebo effect in antidepressant trials.

The 4 June debate was chaired by Tom Fahy, a professor of forensic mental health at the Institute of Psychiatry. Introducing the speakers, Fahy stressed that, particularly as the NHS (the UK's National Health Service) 'spends £381 million a year on antidepressants', the motion was an important one. He then invited Kirsch, the first speaker, to the podium.<sup>1</sup> Kirsch said that he was uncomfortable speaking for the motion as it stood, because, as he has indicated in his article in *PloS Medicine*, antidepressants *are* statistically

superior to placebo – the question is whether this significance is of any use clinically. Tests of statistical significance indicate whether a result can be attributed to the inherent properties of the phenomena being studied, or whether a result stems from chance. Tests of clinical effectiveness, on the other hand, have to do with the size of that effect: they address the question of, as Kirsch said, ‘How much difference would it make in anybody’s life?’

Kirsch began by describing the details of the *PLoS* article that sparked the controversy. He explained that his study, undertaken with colleagues in the USA and Canada and published in *PLoS Medicine*, was a meta-analysis (a statistical compilation of individual RCTs) of all the clinical trial data submitted to the US Food and Drug Administration (FDA) for the licensing of four commonly prescribed antidepressants. Kirsch said the results demonstrated that the antidepressants had statistically significant benefits over placebo, but that those results were of little clinical significance. Drawing on the stipulation of the UK’s National Institute for Health and Clinical Excellence (NICE) for a 3-point improvement in score on the Hamilton Rating Scale for Depression – a 51-point checklist used to determine the severity of a patient’s depression – as the threshold for clinical significance, Kirsch argued that an improvement in 1.8 points is not great enough to consider antidepressants more clinically effective than placebo.

Kirsch then flagged an objection to his own argument: the fact that NICE’s calculation of clinical significance for antidepressant trials is an arbitrary threshold point. There is no scientific reason to believe that a 3-point improvement in Hamilton scores is any more useful than an improvement of 10-point. Despite the arbitrary character of *all* tests of clinical significance, Kirsch stressed that they are necessary, because ‘you can’t make a decision without them’.

Kirsch turned to a second objection to his study: the suggestion that the individual clinical trials he examines in his meta-analysis are flawed, undermining its value. Kirsch said he agreed with that objection: the individual trials *are* flawed, with a major problem being that subjects recruited into the trials ‘are just not depressed enough’ – for ethical reasons that I describe later on. This argument, Kirsch suggested, simply proves his point. If the patient population recruited into RCT trials is not representative of the vast majority who will take them, then ‘clinically significant differences between antidepressants and placebo have not been established’, because the RCTs used to license the four antidepressants are not reliable indicators of the drugs’ clinical usefulness. On the other hand, if the trial subjects *are* representative, then, given the findings of his recent meta-analysis, his conclusion remains the same: ‘clinically significant differences between antidepressants and placebo have not been established’.

Guy Goodwin was the next speaker. He started by noting that he agreed with Kirsch: clinical trials are often not representative of clinical practice, and

there are now 'efforts to improve clinical trial methodologies stimulated by criticisms' from individuals such as Kirsch. Goodwin went on to detail some of the flaws within clinical trial methodologies that make it hard to interpret the results of RCT trials. He focused in particular on the problem of the placebo effect. In antidepressants' trials that have been carried out, a puzzling phenomenon has occurred: the number of patients who have responded well to the placebo arm of trials has often been almost as high as those who have responded well to antidepressants, leading some to suggest that a placebo effect – stemming from, for example, the cathartic value of sheer participation in a trial – is responsible for much of the therapeutic efficacy typically attributed to antidepressants.

Goodwin contested this, suggesting that there are a number of reasons why participants on the placebo arms of trials might appear to respond equally well to placebo, such as the problem of baseline rating inflation, where a patient with mild depression or even non-existent depression might be enrolled in a trial by overly enthusiastic trial investigators, who, for financial reasons, 'are motivated to enter lots of patients into trials'. Often, such participants might appear to become more agitated and less depressed as the trial goes on, something that, in Goodwin's view, stems less from the effect of placebo than from the fact that their initial 'baseline' level of depression might not have been as high as recruiters initially purported in order to enrol subjects in the first place.

Joanna Moncrieff took the podium next. She focused on problems with rating scales such as the Hamilton and their use in RCTs that make it difficult to determine the mechanistic action of antidepressants, as well as their therapeutic value. She stressed that as most antidepressants have sedative effects, and therefore can produce less agitation and less wakefulness, they might appear, on a clinical trial, to be alleviating depression, while in reality they may simply be increasing a subject's level of sedation: 'anything with sedative effects is likely to look good in a trial that uses a depression rating scale'.

As Moncrieff returned to her seat, Lewis Wolpert made his way to the microphone. 'Listening to the last speaker', he said, leaning over and heightening the microphone, 'clearly, she has never been depressed. In fact, I say to psychiatrists: many of you who treat depressives, but have not been depressed, haven't a clue what it's about.'

'I suffer from depression, and the idea that I take my antidepressant every morning – mine is Effexor, I take a low dose – the idea that I'm doing this as a placebo, well, I'd like to put my tongue out to all of you. Now, I know it's not scientific if I say that I know this antidepressant is working. But I do know that over the last 13 years I've taken different antidepressants, and when one doesn't work, I've changed to another.'

'I was suicidal', he continued. 'Why did I stop being suicidal? I think it was the antidepressants in the first place, and then later on, cognitive behavioural

therapy.’ Wolpert shifted slightly and addressed Kirsch directly. ‘I think this particular article of yours has done enormous damage to people with depression.’ He turned back to the audience. ‘Please vote against this absurd motion.’

Later on, I will return briefly to the Maudsley debate in order to analyse the implications of these comments. First, I turn to an elaboration of the problems with RCT methodologies pointed out by the speakers.

## ANTIDEPRESSANT EFFICACY AND THE METHODOLOGICAL LIMITATIONS OF RCTS

Kirsch and his colleagues’ study, ‘Initial Severity and Antidepressant Benefits: a Meta-Analysis of Data submitted to the Food and Drug Administration’, examined the clinical trial data submitted to the FDA for the licensing of four SSRI and SNRI antidepressants: fluoxetine (Prozac); venlafaxine (Effexor); nefazodone (serzone) and paroxetine (Paxil/Seroxat). They suggested in examining these trials, as Kirsch notes at the debate, that antidepressants are no more clinically effective than placebo except in treating severe depression.<sup>2</sup> In the following section, I focus on three things that complicate Kirsch’s findings: the usefulness of rating scales in psychiatric trials; problems surrounding how to determine clinical significance; and recruitment bias.

Since the 1960s, US and UK regulators have stipulated that all new medicines must be tested for efficacy and safety through RCTs in order to earn licences, a phenomenon attracting increasing attention from social scientists interested in the rise of standardized testing and its influence on drug regulation and delivery (see, for example, Abraham, 2007; Abraham and Sheppard, 1999; Corrigan, 2002; Dehue, 2002; Lakoff, 2005, 2007; Marks, 1997; Sunder Rajan, 2007; Timmermans, 2005; Timmermans and Berg, 2003; Vrecko, 2008; Wahlberg and McGoey, 2007; Will, 2007). The double-blind RCT, where neither the trial administrators nor trial participants know who is receiving the active treatment or who the control, has long been hailed as the gold standard for determining a new treatment’s worth, because the use of blinding, random allocation and control groups helps to determine whether the treatment itself or an outside variable – such as the individual’s physiology or an environmental factor – produces the treatment’s effects.

RCTs emerged as a way for regulators to evaluate the safety of treatments, as well as a way to ensure human experimentation takes place in controlled, institutional settings that are themselves subject to regulatory oversight. The need for drug manufacturers to obtain evidence of the safety of new drugs through undertaking ‘adequate and well-controlled trials’ which are large enough to permit quantitative evaluation of treatment effects using ‘appropriate statistical methods’ soon became an obligatory passage point for the evaluation and licensing of new therapies (Marks, 1997: 130).

An impetus leading medical researchers in the mid-20th century to carry out clinical trials was their role in producing demonstrable results which could be shown to governmental authorities in order to procure more research funds. Austin Bradford Hill, for example, who directed the UK Medical Research Council's celebrated 1948 trial of streptomycin in pulmonary tuberculosis, viewed the trial as a way to 'speedily and effectively reveal the value of the treatment' in order to secure funds for its more universal provision (Doll, 2003: 930). RCTs serve as technologies of trust, to use Theodore Porter's term: a means to promote confidence in the outcomes of experiments within 'weak communities' where personal ties between regulators, researchers and clinicians are limited due to distance and the size of research communities (Porter, 1995: 225). They can also be considered inscriptions in Latour's sense: devices combinable with other inscriptions in a way that augments the power of the numbers or concepts behind the inscriptions, while simultaneously simplifying their visual representation (cf. Latour, 1986).

On the heels of Hill's successful trial demonstrating the usefulness of streptomycin in TB patients, the value of RCTs eventually came to be accepted by clinicians in the UK and USA, in part as a way to counter advertorial claims of therapeutic efficacy advanced by the pharmaceutical industry (Marks, 1997: 157).<sup>3</sup> By helping to generate visible evidence – accessible to both clinician and patient – RCTs were seen as providing a defence against arbitrary claims of a treatment's worth. A physician who refused to implement the evidence from a controlled trial was not simply acting parochially or archaically: she was acting *immorally* – unduly exposing patients to the unsubstantiated claims of industry, or worse, to the vagaries of her own personal whims. In recent years, despite the fact that the majority of RCTs are either conducted or funded by the pharmaceutical industry, the methodology itself has retained its early status as a defence against industry bias.

As I explore below, a conundrum has emerged alongside the justifiable faith in RCTs as a way to determine and demonstrate a treatment's worth. Partly because of the regulatory and clinical dependence on RCTs, problems within individual RCTs are rarely seen as indictments of the methodology as a whole: instead, they are used as evidence of the need to refine and perfect the methodology further. Secondly, when a treatment's effects are difficult to measure quantitatively, such as with psychotropic drugs, measures must be introduced in order to demonstrate commensurability among trial participants and to evaluate drug response, something that makes it difficult to know whether the effects of a drug are due to its efficacy, or are simply a construct of the measures adopted to demonstrate its worth. Third, what *can* be easily demonstrated through RCTs assumes disproportionate importance, something illuminated by the tendency of regulators and clinicians to equate statistical significance with clinical usefulness.

The limitations of RCTs are especially visible with psychiatric drugs, where effects are harder to measure than, say, a cancer therapy, where a blood test can reveal a treatment's value. Although the problem is particularly evident with psychiatric drugs, many of the problems I discuss below are relevant to *all* classes of drug, where a disproportionate privileging of statistical significance had led to the licensing of numerous drugs with adverse effects that were either innocently undetected or deliberately ignored because they could not be easily measured.<sup>4</sup>

The difficulty in measuring the effect of an antidepressant drug can be seen at every stage of its production and testing. In order to determine, for example, whether a participant is eligible for enrolment in an antidepressant trial, one first has to determine whether the patient *has* a psychiatric disorder, and its level of severity. This is a challenge for psychiatry, where the process of diagnosing patients is one that can vary widely among different clinical settings. In response to this problem, psychiatrists developed scales, such as the Hamilton rating scale, where behaviour and mood are codified and categorized according to standardized checklists. Patients receive a score for responses to questions about mood, insomnia, sexual function and so on – and the final score is used to help determine the severity of a patient's illness. As Andrew Lakoff notes, by translating 'subjective experience into quantitative cut-off points and outcome measures, [rating scales] make it possible to assemble and compare groups of patients across sites and between evaluators' (2007: 58). In psychiatry, there are a number of general domains of measurement: disease-specific rating scales, such as the Hamilton; patient-based disease-specific scales, such as the Beck Depression Inventory (BDI); and patient-based non-disease-specific scales of global functioning, such as the Quality of Life scale (Healy, 2001: 323).

One of the problems with rating scales is that, as Moncrieff reiterated on 4 June, they have numerous questions related to sleep and anxiety, items likely to show effective change with any drugs with sedative properties (Moncrieff, 2002). A second problem, as David Healy points out, is that changes in rating scale scores are sometimes presented as evidence of a treatment's efficacy, when, in reality, a drop in a Hamilton score from 38 to 19 points does not necessarily say anything about a treatment's clinical value (Healy, 2001: 323).

Healy's point refers to a question that has received a problematic dearth of attention in medicine, in drug regulation, and in the social sciences in general: the conflation of statistical significance with effectiveness in practice, and the tendency of regulators to view statistical significance as proof that a drug or treatment will be useful clinically. As Tim Kendall, joint director of the UK's National Collaborating Centre for Mental Health, notes in a recent interview, and as Kirsch pointed out on 4 June, too often with drug trials, regulators, clinicians and manufacturers equate statistical significance with clinical significance, when, in reality, a drug that is shown to be statistically



better than placebo or a comparator is not necessarily any better in a clinical setting (Graham, 2008; Kendall and McGoe, 2007). Misinterpretation of the real-world value of an effect shown to be statistically significant is not a problem limited to medicine. As Deirdre McCloskey and Stephen Ziliak argue, researchers in economics, psychology, pharmacology and education, among others, have overwhelmingly privileged the question of *whether* an effect is due to chance or inherent value, i.e. whether it is statistically significant, over questions of *how* large is the effect in practice, or *what* difference will it make if implemented (Ziliak and McCloskey, 2008).

Within medicine, tests of statistical significance give a yes or no answer to the question of whether a drug's effect is due to chance, but say nothing about the *size* of that effect (cf. Healy, 2006; Healy, 2009; Turner and Rosenthal, 2008). In contrast, effect sizes – or the measure of the strength of the relationship between two variables – do allow, in principle, the ability to determine clinical significance, and some researchers and policy-makers are beginning, only now, to grasp the importance of evaluating clinical usefulness.

There are numerous different ways to calculate effect sizes. In comparing two groups, such as placebo response and antidepressant response, a typical effect size index is a standardized mean difference, where the mean response on the placebo arm would be subtracted from the mean response on the active arm, and then divided by the standard deviation of either group. Standardized mean difference as an effect size measure was suggested in the 1960s by Jacob Cohen, who proposed values of 0.2, 0.5 and 0.8 to represent small, medium and large effects respectively (Huberty, 2002; Turner and Rosenthal, 2008). Drawing on Cohen's work, the UK's National Institute for Clinical Excellence (NICE), a body that formulates treatment guidelines for the UK's National Health Service, suggested that in antidepressant trials, effect sizes of 0.5 or greater should be considered clinically significant. In trials that employed the Hamilton scale, a 3-point difference or more in Hamilton change scores was viewed as the threshold for clinical efficacy (NICE, 2004).

In their recent analysis of antidepressant trials, Kirsch and colleagues (2008) found that the four antidepressants had a mean improvement of 9.60 points on the Hamilton scale. Scores on the placebo arms improved by 7.80 points. The mean drug–placebo difference was therefore 1.80 in Hamilton improvement scores, an effect size difference that Kirsch calculates as 0.32. Drawing on NICE's stipulation of 0.5 as the cut-off for clinical significance, Kirsch argues that a difference of 0.32 is not great enough to consider antidepressants more clinically effective than placebo.

Although the finding is logical when seen through the lens of NICE's recommendations for clinical significance, it also calls into question, as Turner and Rosenthal (2008) note, the usefulness, and universal applicability, of NICE's cut-off point of 0.5. Writing in the *British Medical Journal* in the aftermath of Kirsch's study, Turner and Rosenthal note that Kirsch's findings

were remarkably similar to their own study of the selective publication of antidepressant trials, published in the *New England Journal of Medicine* in January 2008. The main finding of both studies was that antidepressant drugs are far less effective than is apparent from the published clinical trial data that have appeared in medical journals. Turner and Rosenthal derived an overall effect size of 0.31 as a measure of the efficacy of antidepressants, a size similar to Kirsch's value of 0.32.

'Although these two sets of results were in excellent agreement', noted Turner and Rosenthal in the *British Medical Journal (BMJ)*, 'our interpretations of them were quite different.' They concluded that antidepressants *are* clinically superior to placebo, while Kirsch concluded antidepressants are ineffective. The difference stems from Kirsch's reliance on NICE's stipulation of a 3-point difference in Hamilton scores as a determination of clinical significance, something Turner and Rosenthal view as questionable:

On what basis did NICE adopt the 0.5 value as a cut-off? When Cohen first proposed these landmark effect sizes, he wrote, 'The terms "small", "medium", and "large" are relative . . . the definitions are arbitrary . . . these proposed conventions were set forth with much diffidence, qualifications, and invitations not to employ them if possible'. He also said, 'The values chosen had no more reliable a basis than my own intuition'. Thus, it seems doubtful that he would have endorsed NICE's use of an effect size of 0.5 as a litmus test for drug efficacy. (Turner and Rosenthal, 2008: 516)

Turner and Rosenthal go on to illustrate their point with an analogy. Imagine, they suggest, antidepressant efficacy in terms of litres of a fluid called 'd-juice'. In their study, they found 0.41 litres of d-juice in the 'glass' representing published clinical trials, and 0.31 litres of d-juice in the studies – both unpublished and published – held by the FDA. Although 0.31 litres of juice is less than the 0.41 litres suggested by the pharmaceutical industry in published data, it is still enough to suggest that antidepressants have some clinical usefulness. Kirsch and colleagues found 0.32 litres of d-juice, but unlike Turner and Rosenthal, they concluded, on the basis of NICE recommendations, that the 'glass' of antidepressant efficacy was entirely empty (Turner and Rosenthal, 2008).

Although Kirsch himself, speaking at the Maudsley debate, alluded to the arbitrariness of NICE's calculation of clinical significance, what he did not mention is that he had earlier published work *challenging* NICE's measurement – the same measurement that forms the basis of his argument in the recent study in *PLoS Medicine*. In an article contesting NICE's 2004 guidelines for the use of antidepressants in adults, Kirsch and Moncrieff note that 'no research evidence or consensus is available about what constitutes a clinically meaningful difference in Hamilton scores . . . NICE required a difference

of at least 3-points as the criterion for clinical importance but gave no justification for this figure' (Moncrieff and Kirsch, 2005: 155).

In the carrying-out of antidepressant trials and the interpreting of their results, uncertainty surrounding how well statistical significance translates into clinical usefulness, uncertainty surrounding patient response, and uncertainty surrounding how well rating scales can capture the nature of psychiatric drug response, are something well known to clinical trialists and to researchers such as Kirsch and Goodwin, both of whom pointed out on 4 June how *useless*, rather than useful, RCTs often are in revealing a drug's therapeutic value. Researchers are aware of the magnitude of what remains impossible to know and to demonstrate through RCTs. In presenting the results of trials to regulators or to journal editors, however, researchers must obscure their own recognition of the uncertainty surrounding their methodological tools.

Whether one is convinced of the efficacy of antidepressants, or one doubts it, the challenge is the same: to convince others of one's findings and arguments while fostering a lack of awareness of the many methodological weaknesses that have rendered one's conclusions questionable. This challenge is complicated by the fact that all are aware of the methodological limitations in the studies of those with competing points of view. But to admit knowledge of these limitations risks exposing the contingent nature of one's own results. Seen in this light, it is ignorance itself that is harnessed as a resource in strategizing how to make the most of one's knowledge.

The anthropologist Michael Taussig (1999) has written about the value of non-knowledge, and how 'knowing what not to know' is a key social and political tool of negotiation for both those in authority and those subject to it. Secrecy and non-knowledge are indispensable to the operation of power, 'Not only because power imposes secrecy on those whom it dominates, but because it is perhaps just as indispensable to the latter' (Taussig, 1999: 57). In debates over the efficacy of antidepressants, strategic ignorance has been useful to a range of parties, from researchers to manufacturers, and, thinking back to Lewis Wolpert's certainty of what a study such as Kirsch's could not know: his own personal reaction to antidepressants, to patients alike (for further discussions of the political uses of ignorance and secrecy, see Barry, 2006; Lamble, 2009; McGoey, 2007; Proctor, 1995).

In the essay 'Intentional Ignorance: a History of Blind Assessment and Placebo Controls in Medicine', Ted Kaptchuk (1998) notes that the spectre of purposeful non-knowledge has been integral to the evidence produced by clinical trials. The emphasis, for example, on blinding – where neither a trial administrator nor a trial participant knows who has received the placebo or who has received an active substance – emerged from the belief that even the most respected biomedical researchers were capable of distorting trial outcomes if their personal biases were left unrestrained:

Previously, the taint and accusations of bias, prejudice, overenthusiasm, credulity, and delusion were reserved for deviant healers; now, what was once a fringe threat was internalized. Even the judgements of the most senior clinicians concerning the efficacy of new therapeutics were suspect. 'Bias' now haunted medicine. (1998: 430)

Because of its seeming rigour in controlling the personal influences both of researcher and patients on trial outcomes, the double-blind RCT soon became the epitome of morally incontestable experimentation: 'the adoption of blind assessment in medicine has had as much to do with shifting political, moral and rhetorical agendas and technical research design issues as with scientific standards of evidence. [Blinding] has been a vehicle to confer social authority and moral legitimacy' (Kaptchuk, 1998: 432).

Blinding was introduced to trials to help temper bias, to help free clinical trials from political or personal influence. But, as the Maudsley debate illustrates, commercially and ethically motivated decisions about which patients to include in trials and which standards of clinical efficacy to adopt remain integral to the construction and implementation of clinical trials. Bias has persisted, and taken on new forms, despite efforts to eradicate it through blinding mechanisms – or perhaps because of need for creative means to evade such mechanisms.<sup>5</sup>

In the next section, I turn to a separate problem highlighted in the debate on 4 June – the fact that often subjects enrolled in an antidepressant trial are not representative of the patients who will take the drug, something that again calls into question the value of clinical trials as tools for determining therapeutic usefulness. I then examine the paradox raised in the introduction: why is it that RCTs command *greater* regulatory and rhetorical value the more individuals point out the limitations of individual studies?

## RECRUITMENT DEMANDS: SELECTING OUT SEVERELY DEPRESSED PATIENTS

When designing an RCT to test a new antidepressant, the demands of selecting a group of standardized subjects eligible for enrolment in a trial are complicated by the requirements of ethical review boards in North America and Europe, which often stipulate that the more severe a person's disorder, the greater the ethical duty to avoid placing such a patient in a randomized trial (Miller and Brody, 2002; Miller and Silverman, 2004). This is because if they are randomized to the experimental treatment, they are potentially denied access to the best, proven therapy, something that contravenes international statutes such as the Declaration of Helsinki that seek to protect the human subjects of medical experimentation from harm or exploitation. In

recent years, ethical review boards in the USA and the UK have sought to adhere to international statutes such as Helsinki through demanding that researchers curtail the use of placebo in clinical trials, as well as avoid recruiting subjects at the severe end of a disorder, as those patients are more susceptible to suffering adverse reactions (see Petryna, 2007).

A result of the stipulation against placebo use, as a UK psychiatrist based at Oxford University described to me in an interview in February 2005, is that much of the available RCT data for SSRI antidepressants such as Prozac is 'necessarily skewed towards the relatively mild, trouble-free end. Because they're the only people that it's actually ethical to put in.' This informant's point was returned to a number of times by psychiatrists and epidemiologists I have spoken with. A former chair of the UK Royal College of Psychiatrists' Faculty of Child and Adolescent Psychiatry noted, for example, that 'you almost have to be well in the States to get on a trial. Because if you've got anything wrong with you, you get excluded.' An epidemiologist with whom I spoke in March 2005 noted that he found it unsurprising that early antidepressant trials had not shown the suicidal risk later detected by practising clinicians because 'they select out all the suicidal people. That's what you do in clinical trials. There's nothing underhand about that.'

There may be nothing underhand about this tendency, but it does raise the question of the usefulness of RCTs in determining clinical effects among patients routinely excluded from trials for ethical reasons. One of the difficulties raised by the systematic selection of individuals at the less severe end of a disorder or a disease for participation in clinical trials is, as Healy has stressed, the likelihood of a disconnect between a treatment's performance during a trial, and its performance when distributed clinically (Healy, 2001). Complicating things further is the fact that, in the placebo-controlled trials that have taken place despite increasingly strict ethical constraints, placebo response has been, as both Goodwin and Kirsch noted at the Maudsley debate, surprisingly high (see Wahlberg, 2008; Wilson, 2008).

Given this, a challenge for antidepressants' manufacturers is proving to national regulators that an antidepressant's benefits are statistically significant in comparison to placebo (Lakoff, 2007). To overcome this problem, manufacturers have adopted a number of practices geared at eliminating research subjects who are likely to respond to placebo from the trial. As Lakoff was told by a biostatistician, 'The biggest problem is getting the right patients', through looking at clues that might indicate one's susceptibility to responding to placebos – age of onset of depression, family history and so on – and labelling those patients as ineligible for a trial. A second strategy has been the adoption of the 'mousetrap technique', the name for a single-blind placebo run-in period where doctors are aware of which patients are receiving a placebo, while patients are unaware that the trial has not yet begun. With this technique, experimenters stage a mini-trial, giving all patients a placebo for

a week, and then eliminate those who have responded to the placebo (Lakoff, 2007: 67).

The development of tailored run-in periods is just one of the many tactics adopted by manufacturers who struggle to meet regulatory requirements that demand statistical evidence of efficacy through quantifiable measures such as ratings scales. Even though manufacturers 'are quite sceptical about the capacity of the standard rating scales to produce a consistent patient population for testing . . . and also that the scales are applied inconsistently by raters', they must collude with regulatory demands to demonstrate a consistent drug response (Lakoff, 2007: 65). To researchers and clinicians, these points are not surprising. If anything, the opposite is true: what is surprising is how mundane such methodological obstacles appear to those conducting and implementing RCTs. The question, though, is why such mundane difficulties are difficult to challenge except with recourse to the very methodological tools found wanting to begin with.

### PROFITABLE FAILURE AND THE CONSTRAINTS OF METHODOLOGICAL MIMESIS

As the Maudsley debate highlighted, numerous practitioners are frustrated with the limitations of RCTs in psychiatry and their inability to provide reliable evidence of the value of a treatment when distributed clinically. At the same time as people point out their limitations, clinical trials command increasing political importance, particularly with the emergence of bodies such as the UK's NICE, a body that devises treatment guidelines that doctors and nurses working in the National Health Service must adhere to. In formulating guidelines, bodies such as NICE explicitly privilege the evidence from RCTs over other forms of experiential evidence, in part because RCTs are perceived as the most rigorous form of experimentation in medicine (see Heimer *et al.*, 2005; Kendall and McGoey, 2007). Even though NICE guidance is often questioned by practising clinicians, who suspect guidelines are less reliable – and more politically vulnerable – than is purported by NICE staff, clinicians must take care, through the need to avoid suggestions of malpractice, in appearing publicly compliant with a system they often privately object to.

If a practitioner *does* want to challenge the authority of an individual RCT, or the reliability of individual treatment guidelines, he or she must have recourse to corroborative evidence to support his or her dissent, preferably in the form of more RCTs. It is the very methodological weaknesses of RCTs that imbues them with the authority they hold: for to deny the reliability of a particular study, one must reach for more data, more studies, larger RCTs, in order to justify the validity of one's objections. Of course, individuals are

free to suggest that RCTs themselves are incapable of arbitrating in the debate before them. But what data do they possess, what representation, what visuals, what inscriptions (cf. Latour, 1986) does such a dissenter have at hand to convince others of the value of her or his interpretation over others'? The problem is not that individuals are incapable of or restrained from challenging RCTs, but that, unless they have the resources to defend the scientific rigour of their objections, preferably through RCT evidence, their interlocutors are equally free to remain deaf.

The mimetic authority of RCTs – the way that the solution to flawed RCTs is to conduct more of them – can be compared to Michael Power's work on the productive failure of auditing systems. Power has demonstrated how failed audits tend to produce calls for more audits, rather than for reconsiderations of auditing systems in general. The failure of audit, in other words, is the inability to blow the whistle on the failure of audit, something that is useful for considering the questions of why systems in general – whether administrative, political, or methodological – tend to assume greatest authority and, arguably, remain *most* impervious to effective dissent or challenge, at the very point when individuals speak of their widespread inefficacy:

Indeed, the great puzzle of financial audit is that it has never been a more powerful and influential model of administrative control than now, when many commentators talk of an auditing crisis. (Power, 1994: 7)

Why is the authority of RCTs augmented, rather than diminished, the more people point out the limitation of individual studies? What has rendered RCTs for psychotropic drugs so invulnerable to the widespread recognition of their inadequacies in determining clinical effects? One answer lies in the influence of methodological mimesis. The solution to failed audit is more audit. The solution to failed RCTs is more RCTs, their shortcomings magnified through techniques such as meta-analyses which aggregate individual studies (McGoey, 2009).

This point helps to illuminate why controversies from seemingly antithetical parties – such as the explicit opposition between those who doubt the efficacy of SSRIs and those who defend the usefulness of the drugs – tend to strengthen the authority of the methodologies that failed to supply adequate evidence of the drug's efficacy in the first place. A reason for this is that the personal *illusio* of individuals, or their shared sense of investment in the rules of a game and its outcome, tends to structure resistance to the game itself according to certain tacit assumptions (Bourdieu, 1992: 66; McGoey, 2009).<sup>6</sup>

In science and medicine, the shared *illusio* of clinicians, policy-makers and regulators alike is faith in the moral authority of objectivity, or a belief that, as Nikolas Rose notes of the governing influence of impartiality and objectivity in science, 'impersonality rather than status, wisdom or experiences' should dictate the delivery of medical care and direction of scientific

invention. In cases where objectivity remains elusive, the ability to *appear* objective carries a degree of political and moral capital; capital which compounds social and political belief in the value of objectivity, regardless of how unattainable personal detachment is in practice (Daston, 1994; Daston and Sibum, 2003; Porter, 1995; Rose, 1999: 208). Even when researchers are aware of the contingency of RCT results, their *illusio* cues them to the importance of appearing ignorant of that contingency. In science and medicine, all parties invest strategically in the need to appear as divested of personal interest as possible.

Although Kirsch himself questioned the reliability of NICE's numerical threshold for clinical significance in an article published with Moncrieff in 2004, he is forced to adopt that threshold in order to justify his own pronouncements on the efficacy of antidepressants, strengthening the rhetorical authority, and seeming universality, of thresholds that were arbitrarily constructed to begin with. This may, at first, seem unproblematic. If anything, Kirsch's flexibility in alternatively questioning and then adopting NICE's threshold seems to illuminate the ability of individual dissenters to sway a crowd and convince others through appropriating the very methodologies they object to. Science, in short, always furnishes the tools to undermine its own truths.

That interpretation seemed to be the perspective shared by the audience on 4 June. At the beginning of the debate, Fahy took a poll of the room, asking the crowd to vote for or against the motion: 'The house believes that antidepressants are no better than placebo'. Of the 200 or so attendees, almost 150 were against the motion: unsurprising in a room full of psychiatrists charged with the duty to prescribe the drugs each day. After the panellists presented their cases and answered questions from the floor, Fahy retook the poll. Though the vast majority in the room still sided with Wolpert and Goodwin, at least 40 either abstained or voted in favour of the motion – seemingly persuaded by the fact that Kirsch and Moncrieff had presented the most detailed evidence to back their views, while Wolpert relied solely on personal experience, and Goodwin managed only to question the reliability of RCT methodologies. If the debate had been titled 'The RCTs Don't Work', rather than 'The Drugs Don't Work' it is possible not a single dissenting voice might have been heard.

Here is where the key paradox appears. Never before have the inadequacies of RCTs been so apparent to so many. Yet, equally, never before have those in positions of authority – from regulators, to NICE policy-makers, to doctors – relied so extensively on RCT evidence. Because of the seeming ability of RCTs to control personal and professional bias, their results are regularly viewed as more neutral than other forms of inquiry. Yet the vast majority of RCTs are conducted by entities with a commercial stake in their outcome. The ability of Kirsch to challenge his interlocutors through an



adoption of their methods seems to indicate the democratic flexibility of science: its openness to feisty iconoclasts such as Kirsch or Moncrieff. Yet, had his dissent been voiced through a *rejection* of RCT methodology – had he offered, as Wolpert did, mere personal reflections – his objections would have won him little credibility.

Many forms of dissent are *speakable*, but in order to be *audible*, Kirsch's objections had to be voiced in a manner comprehensible to his rivals. His dissent was dictated and stipulated by the very authorities that engendered that dissent, ultimately solidifying the authority of RCTs that failed to provide reliable evidence in the first place. Science may furnish the tools to undermine itself, but they must be *scientific* tools: ensuring that regardless of the radicalism of the critique, the triumph of science – the illusion of science's detachment from politics – is cemented, fuelled by the very challenges it provokes.

## CONCLUSIONS

In this article, I used the recent controversy over Kirsch's study in order to examine the larger question of why RCTs have tended to command greater regulatory and popular legitimacy the more that people point out their failures. I described how, regardless of whether they have value clinically, the erroneous perception of trials as being value-free politically increases the rhetorical capital they contain. If anything, the *more* useless RCTs are in practice, the more their strength is augmented, as more and more practitioners rally around a call for more RCTs in order to remedy the failings of previous trials. In some ways, my analysis reminds me of a problem long rehearsed in political theory and sociology: the question of why political resistance rarely leads to effective political change, and why, as Bourdieu reminds us in essays such as 'Rethinking the State: Genesis and Structure of the Bureaucratic Field' (1999), critique is so easily appropriated by the authorities that have elicited objections to their rule. The contribution of this article has been to examine the ways that the profitable failure of clinical trial methodologies helps facilitate the ability of dominant authorities to appropriate resistance itself.

The tendency of dominant authorities to absorb challenges to their rule *particularly* during periods of crisis or failure can be seen in a variety of political or economic arenas, such as the current financial crisis, where the failure of the finance sector is simultaneously the opportunity of those who have presided over that failure: their expertise becomes even more valued in the aftermath of the collapse of an architecture they helped to build, if only because the deployment of new forms of experiments is deemed too risky for desperate times. This point helps to shed light on the question of why catastrophic situations rarely produce much systemic change, and why, in

considering how it can be that failed technologies or methodologies retain the political clout they do, it is useful to examine which authorities gain the most from the perpetuation of useless methodologies.

## NOTES

My thanks go to Andrew Barry, Scott Vrecko, Martyn Pickersgill and Ayo Wahlberg for comments on an earlier draft. The article was made possible by an ESRC post-doctoral fellowship carried out at Oxford's School of Geography. An earlier version of this article was presented at a workshop at Harvard University, co-hosted by the European Neuroscience and Society Network and Harvard's Department for the History of Science. I am grateful to participants for their feedback.

- 1 The debate is available by podcast on the Institute of Psychiatry's website: <http://www.iop.kcl.ac.uk/podcast/?id=238&type=item>
- 2 For a discussion of Kirsch's earlier studies on antidepressant efficacy, see Wilson, 2008.
- 3 Although Marks's study of the adoption of RCT techniques by US physicians and regulators is justly celebrated as a seminal study of the history of RCTs, he misinterprets the influence of Ronald Fisher's statistical theory on the adoption of randomization in trials conducted by Bradford Hill. Iain Chalmers's work is useful on this point (Chalmers, 2005).
- 4 A seminal example of this is the case of Vioxx, where manufacturers and regulators dismissed early evidence that the drug led to cardiac failure because the evidence was not presented as being statistically significant. One Vioxx trial, for example, showed a ratio of five heart attacks on the Vioxx arm, versus one on the control arm. Despite the 80 per cent increased risk of heart failure on Vioxx, the published study of the trial stated there was *no* difference in effects between treatment and control groups, because the difference was not statistically significant. See Ziliak and McCloskey for further discussion (2008).
- 5 Randomized trials are, of course, not the sole experimental settings where experimenters must play creatively with techniques geared at eliciting reactions from participants most amenable to the desired outcome of the experiment, while simultaneously appearing as unobtrusive and non-directional as possible. Javier Lezaun has examined, for example, the efforts of moderators carrying out focus group research to mitigate 'the conflicting burdens of objective detachment from, and natural empathy with the research subjects' (Lezaun, 2007). With both focus groups and RCTs, the success of the experiment is contingent on how effectively its artificiality is concealed; on how successful moderators are at effacing their own presence.
- 6 In *The Logic of Practice*, Bourdieu describes '*illusio*' as faith in the rules of a game inculcated through one's adjustment to the demands of a field – his term for any setting where agents and their social positions are located – through one's immersion within the field. *Illusio* is 'the sense of investment in the game and the outcome, interest in the game, commitment to the presuppositions – *doxa* – of

the game'. *Illusio* is a shared faith in the obvious sensibility of a game's rules; a faith which is unrecognized as belief because players are unaware their mere participation confirms an investment in and acceptance of the game's structure and consequences (1992: 66–7).

## BIBLIOGRAPHY

- Abraham, J. (2007) 'Drug Trials in International Regulatory Context', *Biosocieties* 2: 41–56.
- Abraham, J. and Sheppard, J. (1999) 'Complacent and Conflicting Scientific Expertise in British and American Drug Regulation: Clinical Risk Assessment of Triazolam', *Social Studies of Science* 29: 803–43.
- Barry, A. (2006) 'Technological Zones', *European Journal of Social Theory* 9: 239–53.
- Bourdieu, P. (1992) *The Logic of Practice*. Stanford, CA: Stanford University Press.
- Bourdieu, P. (1999) 'Rethinking the State: Genesis and Structure of the Bureaucratic Field', in *State/Culture: State Formation after the Cultural Turn*, ed. G. Steinmetz. Ithaca, NY: Cornell University Press, pp. 53–75.
- Chalmers, I. (2005) 'Statistical Theory was not the Reason that Randomization was used in the British Medical Research Council's Clinical Trial of Streptomycin for Pulmonary Tuberculosis', in G. Jorland, A. Opinel and G. Weisz (eds) *Body Counts: Medical Quantification in Historical and Sociological Perspective*. Montreal and Kingston: McGill-Queen's University Press.
- Corrigan, O. P. (2002) 'A Risky Business: the Detection of Adverse Drug Reactions in Clinical Trials and Post-Marketing Exercises', *Social Science & Medicine* 55: 497–507.
- Daston, L. (1994) 'Baconian Facts, Academic Civility, and the Prehistory of Objectivity', in *Rethinking Objectivity*, ed. A. Megill. Durham, NC: Duke University Press.
- Daston, L. and Sibum, O. H. (2003) 'Introduction: Scientific Personae and Their Histories', *Science in Context* 16: 1–8.
- Dehue, T. (2002) 'A Dutch Treat: Randomized Controlled Experimentation and the Case of Heroin-Maintenance in the Netherlands', *History of the Human Sciences* 15: 75–98.
- Doll, R. (2003) 'Fisher and Bradford Hill: Their Personal Impact', *International Journal of Epidemiology* 32: 929–31.
- Graham, J. (2008) 'Facilitating Regulation: The Dance of Statistical Significance and the Clinical Meaningfulness in Standardizing Technologies for Dementia', *Bio-Societies* 3: 241–63.
- Healy, D. (2001) 'The Dilemma posed by New and Fashionable Treatments', *Advances in Psychiatric Treatment* 7: 322–7.
- Healy, D. (2006) 'The Antidepressant Tale: Figures signifying Nothing?', *Advances in Psychiatric Treatment* 12: 320–7.
- Healy, D. (2009) 'Trussed in Evidence? Ambiguities at the Interface between Clinical Evidence and Clinical Practice', *Transcultural Psychiatry* 46: 16–37.
- Heimer, C. A., Petty Coleman, J. and Culyba, R. J. (2005) 'Risk and Rules: the "Legalization" of Medicine', in B. Hutter and M. Power (eds) *Organizational Encounters with Risk*. Cambridge: Cambridge University Press.

- Huberty, C. J. (2002) 'A History of Effect Size Indices', *Educational and Psychological Measurement* 22: 227–40.
- Kaptchuk, T. (1998) 'Intentional Ignorance: a History of Blind Assessment and Placebo Controls in Medicine', *Bulletin of the History of Medicine* 72(3): 389–433.
- Kendall, T. and McGoey, L. (2007) 'Truth, Disclosure, and the Influence of Industry on the Development of NICE Guidelines: an Interview with Tim Kendall', *BioSocieties* 2: 129–41.
- Kirsch, I., Deacon, B., Heudo-Medina, T., Scoboria, A., Moore, T. and Johnson, B. (2008) 'Initial Severity and Antidepressant Benefits: a Meta-Analysis of Data submitted to the Food and Drug Administration', *PLoS Medicine* 5: e45. doi: 10.1371/journal.pmed.0050045
- Lakoff, A. (2005) *Pharmaceutical Reason: Knowledge and Value in Global Psychiatry*. Cambridge: Cambridge University Press.
- Lakoff, A. (2007) 'The Right Patients for the Drug: Managing the Placebo Effect', *Biosocieties* 2: 57–73.
- Lamble, S. (2009) 'Unknowable Bodies, Unthinkable Sexualities: Lesbian and Transgender Legal Invisibility in the Toronto Women's Bathhouse Raid', *Social & Legal Studies* 18: 111–30.
- Latour, B. (1986) 'Visualization and Cognition: Thinking with Eyes and Hands', *Knowledge and Society* 6: 1–40.
- Lezaun, J. (2007) 'A Market of Opinions: the Political Epistemology of Focus Groups', *The Sociological Review* 55(2): 130–51.
- Marks, H. M. (1997) *The Progress of Experiment*. Cambridge: Cambridge University Press.
- McGoey, L. (2007) 'On the Will to Ignorance in Bureaucracy', *Economy and Society* 36: 212–35.
- McGoey, L. (2009) 'Resistance Scripts: the Authority of Method within Regimes of Evidence', in C. Calhoun and R. Sennett (eds) *Creating Authority*. London: Routledge.
- Miller, F. G. and Brody, H. (2002) 'What makes Placebo-Controlled Trials Unethical?', *American Journal of Bioethics* 2: 3–9.
- Miller, F. G. and Silverman, H. J. (2004) 'The Ethical Relevance of the Standard of Care in the Design of Clinical Trials', *American Journal of Respiratory and Critical Care Medicine* 169: 562–4.
- Moncrieff, J. (2002) 'The Antidepressant Debate', *British Journal of Psychiatry* 180: 193–4.
- Moncrieff, J. and Kirsch, I. (2005) 'Efficacy of Antidepressants in Adults', *British Medical Journal* 331: 155–7.
- NICE (2004) 'Depression: Management of Depression in Primary and Secondary Care', *National Clinical Practice Guideline* Number 23: online (last accessed March 2008) at: <http://www.nice.org.uk/nicemedia/pdf/cg023fullguideline.pdf>
- Petryna, A. (2007) 'Clinical Trials Offshored: On Private Sector Science and Public Health', *BioSocieties* 2: 21–41.
- Porter, T. (1995) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.
- Power, M. (1994) *The Audit Explosion*. London: White Dove Press.
- Proctor, R. (1995) *Cancer Wars: How Politics Shapes What We Know and Don't Know about Cancer*. New York: HarperCollins.

- Rose, N. (1999) *Powers of Freedom: Reframing Political Thought*. Cambridge: Cambridge University Press.
- Sunder Rajan, K. (2007) 'Experimental Values: Indian Clinical Trials and Surplus Health', *New Left Review* 45: 67–88.
- Taussig, M. (1999) *Defacement: Public Secrecy and the Labor of the Negative*. Stanford, CA: Stanford University Press.
- Timmermans, S. (2005) 'From Autonomy to Accountability: the Role of Clinical Practice Guidelines in Professional Power', *Perspectives in Biology and Medicine* 48: 490–501.
- Timmermans, S. and Berg, M. (2003) *The Gold Standard: The Challenge of Evidence-Based Medicine and Standardization in Health Care*. Philadelphia, PA: Temple University Press.
- Turner, E. and Rosenthal, R. (2008) 'Efficacy of Antidepressants is not an Absolute Measure, and it depends on How Clinical Significance is defined', *British Medical Journal* 336: 516–17.
- Vrecko, S. (2008) 'Capital Ventures in Biology: Biosocial Dynamics in the Industry and Science of Gambling', *Economy and Society* 37: 50–67.
- Wahlberg, A. (2008) 'Above and Beyond Superstition – Western Herbal Medicine and the Decriminalising of Placebo', *History of the Human Sciences* 21: 77–101.
- Wahlberg, A. and McGoey, L. (2007) 'An Elusive Evidence Base: the Construction and Governance of Randomised Controlled Trials (Introduction to Special Issue)', *BioSocieties* 2: 1–11.
- Will, C. (2007) 'The Alchemy of Clinical Trials', *Biosocieties* 2: 85–101.
- Wilson, E. (2008) 'Ingesting Placebo', *Australian Feminist Studies* 23: 31–42.
- Ziliak, S. and McCloskey, D. (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*. Ann Arbor: University of Michigan Press.

## BIOGRAPHICAL NOTE

LINSEY MCGOEY is a research fellow in science and technology studies at the Institute for Science, Innovation and Society, University of Oxford. She completed her PhD at the LSE's BIOS Centre, followed by an ESRC post-doctoral fellowship at Oxford's School of Geography. She is currently completing a book on the uses of strategic ignorance in political and economic life.

*Address:* Saïd Business School, University of Oxford, Park End Street, Oxford OX1 1HP, UK. [email: linsey.mcgoey@sbs.ox.ac.uk]