Françoise Baylis • Angela Ballantyne
Editors

Clinical Research Involving Pregnant Women

Springer

Chapter 11
Does My Bias Look Big in This?

David Healy and Derelie Mangin

Abstract *Randomised controlled trials (RCTs) are thought to be the gold standard in evidence. This review of their origins and adoption, highlights commonly ignored shortcomings with RCTs. If RCTs are used indiscriminately, their adverse effects may outweigh their benefits. This chapter focuses on antidepressants and how RCTs give the wrong message about safety, efficacy, and effectiveness. The arguments hold true in principle for all treatments, including all treatments for pregnant women. The received wisdom since thalidomide, that we should rarely if ever use drugs in pregnancy, increasingly is being eroded by arguments in support of the use of RCTs. In the case of antidepressants, this has made them among the most commonly prescribed drugs in pregnancy.*

There is a presumption that objectivity comes from the procedures of an RCT. We argue that objectivity comes from collective scrutiny of publicly available data and, in the case of pregnancy, this mandates the creation of pregnancy registries to generate sound evidence on the basis of which to make treatment decisions for pregnant women and women of child-bearing years.

There is a general belief that (RCTs) provide the best possible evidence regarding treatments and that RCTs are the only way to avoid biased judgements on the safety, efficacy and effectiveness of treatments. Allied to a position that pregnant women deserve access to the best quality evidence, this belief mandates an increased use of RCTs in pregnancy and in women of child-bearing years.

In contrast, articles on the value of observational studies invariably include disclaimers as to validity of such studies. The explicit message is that with a great deal of care, it might be possible to reduce the amount of bias to which such studies inevitably will be subject, but observational studies will never approach the quality of disinterested evidence that stems from RCTs.

D. Healy, MD (✉)
North Wales Department of Psychological Medicine, Cardiff University, Cardiff, Wales, UK
e-mail: david.healy54@googlemail.com

D. Mangin, MBChB, DPH, FRNZCGP
Department of Family Medicine, McMaster University, Hamilton, ON, Canada

University of Otago, Christchurch, New Zealand

F. Baylis, A. Ballantyne (eds.), *Clinical Research Involving Pregnant Women*,
Research Ethics Forum 3, DOI 10.1007/978-3-319-26512-4_11

We argue that RCTs are mechanical exercises, that in general they ***do not*** provide good quality evidence, that they are commonly ineradicably confounded, and that an unthinking application of RCTs in pregnancy would be a mistake. If women who are (or who might become) pregnant want the best quality evidence, researchers need to carefully think about the design of appropriate studies. Moreover, if researchers want to reduce bias, the focus should be on collective scrutiny of publicly available data, and a key source of such data will be pregnancy registries. In our view, objectivity comes from collective scrutiny and not mechanical exercises.

This chapter focusses on antidepressants and how RCTs give the wrong message on safety, efficacy, and effectiveness, but the arguments hold true in principle for all treatments.

11.1 The Origin of Randomised Trials

Ronald Fisher created the modern RCT in the 1920s, when investigating the effect of fertilisers. Many factors can confound fertiliser studies such as differences in soil drainage, exposure to wind or sunlight, and a myriad of soil elements. These known factors can be controlled for, but Fisher's insight was to control for unknown confounding factors by randomising fertilisers under study to alternate soil patches. Fisher tied significance testing to randomisation. If experimenters got the same result every time they repeated the intervention, they had designed a good experiment. There was a *Quod Erat Demonstrandum* quality to this strategy – shave a bit off one side of a coin and you can expect heads to come up nineteen times out of twenty. Randomisation in this sense is about leaving nothing to chance and it allows experimenters to show that they know what they are doing. This insight on what Fisher meant by an RCT has slipped out of view (Fisher 1935; Savage 1976; Marks 1997).

Randomisation was first used in medicine in Bradford Hill's trial of streptomycin for tuberculosis (Medical Research Council 1948). Earlier, non-randomised trials had established all that is known about streptomycin for tuberculosis – that streptomycin works in the short term, that the germ becomes resistant over time, and that treatment comes with significant risks such as ototoxicity (Toth 1998). As no new knowledge was gained from the RCT, Hill's streptomycin trial put randomisation rather than streptomycin on the map. If 'efficacy' means that trials accomplish something, while 'effectiveness' means that they work for the intended purpose, then Hill's trial demonstrated the efficacy of trials, but did not establish their effectiveness.

Early Doubts About RCTs

The primacy of RCTs today as a method of evaluation stems not from greater rational or logical coherence, but from events centring on women and pregnancy – the thalidomide tragedy. The birth defects linked to the use of a sleeping pill,

thalidomide, created a political imperative to be seen to be doing something to make patients safer (see Langston 2016). As a result, in 1962 a change was made to the provisions of the US Food and Drugs Act requiring companies to demonstrate the effectiveness of new compounds, with an understanding that this would be done through placebo-controlled RCTs.

While RCTs initially appeared to be a way to contain pharmaceutical company claims, by the mid-1960s Hill (1966) noted that company salesmen were deploying RCT evidence to encourage doctors to use their company's products. At that time, Hill suggested that if RCTs ever became the only way to evaluate drugs that "the pendulum would not just have swung too far [away from physician judgement] it would have come off its hook" (Hill 1966, 113).

In 1962, at the time the US Food and Drug Act was amended, RCTs were a novel evaluation method whose suitability for the task at hand (i.e., improving patient safety) was uncertain (Healy 2012). Indeed, as early as the 1950s, some recognised that the philosophical basis of RCTs was uncertain; there was no agreement on the meaning of statistical significance, and no logical basis for randomisation had been elaborated (shortcomings that remain the case today) (Gigerenzer et al. 1990; Toth 1998). A powerful symbol of this uncertainty is the fact that in 1962, only one drug had demonstrated safety and effectiveness through a placebo-controlled RCT prior to marketing – thalidomide (Lasagna 1960).

Finally, there is a crisis today within drug development that sits poorly with claims that RCTs are an effective evaluation method. As Table 11.1 shows, most major drug groups were introduced in the 1950s without the benefit of RCTs, and the drugs that were introduced during this time remain more effective than drugs that have since come to market using RCTs. Empirically, therefore, it appears that RCTs are not necessary in developing an effective drug arsenal.

11.2 The Placebo Effect

RCTs of fertilisers are not controlled with placebos. The first RCTs in medicine were not placebo-controlled. The first placebo-controlled trials in medicine were not RCTs.

The marriage of RCTs and placebos gives the impression that a further set of confounding factors (biases) is being controlled for. But placebos introduce a systematic bias. An active drug simply needs to beat a placebo on some dimension for others to claim that it is effective. This can be achieved for ever weaker drugs by powering trials accordingly. Manipulations of this sort mean that recent antihypertensives, hypoglycemics and antidepressants are often less effective than drugs introduced without RCTs before 1962 (Table 11.1). Previously, the effects of a drug treatment on a patient had to be visible; today the effects can be invisible at the individual patient level, underpinning distinctions between clinical and statistical significance.

Table 11.1 Drug effectiveness with and without RCTs

Drugs introduced pre 1962	Exemplars of pre-1962 medicines	Post-1962 medicines: more effective or not
Analgesics	Morphine, Paracetamol	No
Antibiotics	Penicillins, Tetracyclines	No
Anticonvulsants	Barbiturates, Valproate	Possible
	Phenytoin	
Antidepressants	Tricyclics, MAOIs	No
Antihistamines	Chlorphenamine, Diphenhydramine	No
Antihypertensives	Thiazides	No
Antipsychotics	Clozapine, Haloperidol	No
Chemotherapies	Nitrogen mustards	Yes
	Cisplatin	
Contraceptives	Second generation COC	No
Diuretics	Furosemide	No
Hypoglycaemics	Metformin	No
Steroids	Prednisone	No
Stimulants	Dexamphetamine	No
	Methylphenidate	
Tranquilisers	Diazepam	No
Vaccines	Polio, Smallpox	No

RCTs & Efficacy

Currently, the design of RCTs to establish efficacy and effectiveness for an antidepressant, involves testing an active drug against a placebo in trials lasting six weeks, using rating scales as outcome measures. There is academic debate about whether a statistically significant finding of doubtful clinical significance constitutes efficacy. But in practice, regulators like the US Food and Drug Administration (FDA) license drugs for use on the basis of such data, and so for almost everyone debates about either efficacy or effectiveness are academic. The drugs are assumed to be effective and are put into ever increasing use as few doctors and no politicians want to deny patients, especially pregnant women, effective treatments.

Consider the following thought experiment. A company does a series of RCTs on alcohol as an antidepressant. It uses its current abilities to hide data from RCTs, to only publish data from selected RCTs, and to ghost-write all publications. Using these strategies, the company could achieve an identical outcome as is achieved with RCTs of SSRIs (selective serotonin reuptake inhibitors) as antidepressants. It is clear, however, that designating alcohol as an antidepressant would not be a good outcome for pregnant women (or any other patient group). With this thought experiment, we are able to pit other knowledge about the use of alcohol during pregnancy against knowledge that might stem from RCTs of alcohol as a possible effective antidepressant. That is, we can put the RCT knowledge in perspective. But in most

other cases we can't do this, and we assume that because the claims have come through an RCT, as in the case of comparable claims about SSRIs, that this is good quality data. In real life, for instance, the example of burgeoning stimulant use driven by RCTs suggests that we are not able to deploy wisdom or even common sense against the treatment imperative that flows from RCTs.

The received wisdom since thalidomide, that we should rarely if ever use drugs in pregnancy, is being eroded by RCTs that have made drugs such as antidepressants among the most commonly prescribed drugs in pregnancy. If safety in pregnancy improved in the 1960s, following the thalidomide tragedy, this had nothing to do with the introduction of RCTs, and everything to do with the fact that pregnant women and their doctors became reluctant to use medications during pregnancy.

RCT Crisis

It is important to distinguish the failings of RCTs that we have linked to a crisis in drug development, from a much more commonly referred-to crisis concerning the conduct of RCTs. This latter crisis is linked to the use of surrogate outcomes, in trials of inadequate duration, against a regulatory background that will license products on the basis of biased trial data.[1] Our goal in this chapter is not to offer another list of the many failings of the conduct of RCTs. We recognise that RCTs have an important place in therapeutics but, we maintain that even if they are carried out impeccably, their adverse effects may outweigh their benefits. Adapting Muir Gray's dictum that all screening is harmful, we might say that all RCTs are harmful but, in some instances, there are also benefits that warrant taking the unavoidable risks involved (Raffle and Gray 2007).

Mediculture or Medicine?

Our argument in brief is: People and their diseases, and the treatment of those diseases, are not uni-dimensional in the same way as Fisher's soil patches and growing crops. As a result, transforming a chemical into a medicine is a different matter to demonstrating a chemical is an effective fertiliser.

For example, a fertiliser has only one action we need pay heed to, but a medicine may have a hundred effects all of which need attention. It is not problematic to designate a primary effect in an RCT of fertilisers and ignore other effects. The fact that a small proportion of ears of corn might die prematurely because of the fertiliser is of no consequence. But medicine is critically concerned with potential benefit to an

[1] For a discussion of licensing in the face of inadequate, contested data see Healy's discussion of a decision to license Zoloft on the basis of ghost-written publications stemming from these two positive RCTs, when there were ten or more negative RCTs (Healy 2012).

individual patient, and average effects are only useful insofar as they might be of help to an individual patient. Average effects that obscure potential harm to an individual patient entail risks that may not be worth taking. Clinical practice wants to manage heterogeneity, not act as though it doesn't exist. This is especially true in pregnancy (see Baylis and MacQuarrie 2016).

Randomisation aims at eliminating sources of objective bias. But in the case of SSRI trials, for instance, the possible effects of these drugs on mood are designated as primary effects, when such effects are less likely than effects on sex and bowel function. The trial process then means that these more common effects are ignored. Is bias being eliminated here or systematised?

11.3 Antidepressants and Suicide

A Thought Experiment

The confounders that randomisation can introduce in testing a medication (that are not confounders in testing a fertiliser), can be drawn out through two examples involving antidepressants and suicide in depression. The lessons about confounders, however, apply to all drug groups and all drug effects.

Imipramine, the first antidepressant, was launched in 1958 without RCTs. In 1959, at a meeting convened to discuss its effects, several clinicians reported having witnessed patient agitation after exposure to Imipramine, how the agitation cleared after stopping the Imipramine, and then reappeared on re-exposure – a medical testing protocol known as challenge-dechallenge-rechallenge. These clinicians decided that, wonderful though Imipramine was for many patients, it could trigger suicidal and homicidal ideation in some people (Davies 1964). The challenge-dechallenge-rechallenge protocol offers as convincing a demonstration of cause and effect, as the statistical significance testing of the type advocated by Fisher. Both tests show greater replicability than is found with the statistically significant findings reported from most RCTs today.

Imipramine and related tricyclic antidepressants are serotonin reuptake inhibitors. They are more clinically potent than most SSRIs, 'beating' SSRIs in patients with melancholia (Healy 1998). Melancholic patients are 80 times more likely to commit suicide than mildly depressed patients (Hagnell et al. 1981). Accordingly, comparing Imipramine and placebo in an RCT of melancholic patients would likely show less suicides and suicidal acts on Imipramine than on placebo. The relative risk might be as low as 0.5. Thus, a drug that causes suicide will also appear to protect against suicide, in some clinical trials.

In contrast, various meta-analyses of suicides and suicidal acts in SSRI and post-SSRI RCTs indicate a relative risk that SSRIs will cause suicide and suicidal acts of roughly 2.0 (Fergusson et al. 2005). This different outcome results, in part, from the fact that SSRIs weaker than Imipramine were tested in people who were at less risk

Table 11.2 Suicidal acts in Major Depressive Disorder trials (MDD) trials

Major Depressive Disorder trials (MDD)	Paroxetine	Placebo	Relative risk
Suicidal acts/Patients	11/2943	0/1671	Inf (1.3, inf)

of suicide. As a result, the rate of suicidal acts on placebo is reduced – making the risk from SSRIs more noticeable. When a drug such as Imipramine is put into this assay system, it would show the same excess of suicidal acts.

It is common to hear claims that RCTs demonstrate cause and effect. But this thought experiment shows that if a trial is not designed to look at an issue, it cannot show cause and effect with respect to that issue. These RCTs of SSRIs say nothing about causality, except insofar as there could not be an excess of suicides and suicidal acts if the SSRIs don't cause suicide. Better evidence that SSRIs can cause suicide in some people comes from Teicher's 1990 paper on Prozac and suicide that demonstrated the challenge-dechallenge-rechallenge relationships (Teicher et al. 1990). Another implication of the thought experiment is that RCTs do not generate reliable data on frequency. Even in studies designed to look at antidepressants and suicide, we cannot in fact have any idea from RCTs how often antidepressants might trigger suicidality in clinical practice. The next example will buttress this point.

Actual Experiments

In the early 1990s, SmithKline undertook a study of paroxetine (Study 106) in patients with Intermittent Brief Depressive Disorders (IBDD). The study terminated early, and was never published. The rate of suicidal acts on paroxetine was threefold higher than on placebo.[2] SmithKline then undertook study 057 in a similar group of patients (Verkes et al. 1998). Neither trial supported using paroxetine for IBDD.

In April 2006, GlaxoSmithKline issued a press release that presented the following data for patients in paroxetine Major Depressive Disorder (MDD) trials (Table 11.2).

The MDD patients on paroxetine showed a significant increase in the risk of suicidal acts as compared with placebo (GlaxoSmithKline 2006). The press release also contained data on suicidal acts from the previous IBDD trials (Studies 106 and 057), despite the fact that these studies did not support using paroxetine for IBDD. When the data from all three studies were aggregated, surprisingly the risk of suicidal acts on paroxetine in depression trials vanished (see Table 11.3).

It is possible to add 16 more suicidal acts to the paroxetine IBBD column in Table 11.3 (viz., 48/147), increasing the relative risk of an adverse event on paroxetine to 1.4 (viz., the combined paroxetine suicidal act number increases to 59 (viz.,

[2]Data available upon request from David Healy.

Table 11.3 Suicidal acts in Major Depressive Disorder trials (MDD) and Intermittent Brief Depressive Disorders (IBDD) trials

	Paroxetine	Placebo	Relative risk
MDD Trials Suicidal Acts/Patients	11/2943	0/1671	Inf (1.3, inf)
IBDD Trials Suicidal Acts/Patients (Studies 106 and 057)	32/147	35/151	0.9
MDD and IBDD Trials Combined Suicidal Acts/ Patients	43/3090	35/1822	0.7

43 + 16)), and still get the same apparently protective outcome overall with paroxetine (paroxetine 59/3090; placebo 35/1822). This paradoxical outcome is predictable. Knowing what a drug can do makes it possible to design placebo-controlled RCTs that use a problem the drug actually causes, to hide that same problem. In this respect, a medicine is unlike a fertiliser.

It is clearly bad meta-analytic technique to lump the datasets for the IBDD and the MDD trials, but the example points to a deeper problem for RCTs undertaken in heterogeneous clinical populations – from back pain to Parkinson's disease. Just as IBDD patients can meet the criteria for MDD, so too diverse patients with back pain or Parkinson's disease or Type II diabetes can meet criteria for different illnesses. Provided there is more than one IBDD patient entered into MDD trials, randomisation will ensure these patients will hide the effect of an SSRI on suicidal acts. Similarly, drugs for back pains, parkinsonian syndromes, or diabetic states of one type may mask what may be beneficial treatment effects on other types of back pain, Parkinson's disease, or diabetes.

The only way to overcome this bias and get a result that would have Fisher agreeing that we know what we are doing is, in fact, to understand the pathophysiology of the clinical condition we are treating and the pharmacogenetics of the drug we are using. It is only then that RCTs would take on the quality of a demonstration that was Fisher's original intention.

11.4 Pharmagnosia

Unlike fertilisers, medicines have a hundred or more effects. When we design an experiment employing randomisation to manage the unknown unknowns for one of these effects, we risk generating ignorance about ignorance regarding most of what the drug does. The process is akin to hypnosis, where holding a subject's attention to one focus can lead him/her to miss more important material out of focus, especially when for the sake of 'objectivity' patient reports are essentially ignored.

In the case of the SSRIs, the choice of endpoint was dictated by business considerations. This meant powering studies to produce a statistically significant outcome on rating scales that measure clinical changes in a very rough fashion. But because the Hamilton Ratings Scale for Depression (HAMD) was the primary endpoint

there was a focus on scores from this, data on sexual functioning and the other 99 effects these drugs have, were either not collected or poorly collected, letting companies claim afterwards that less than 5 % of those taking SSRIs had a disturbance of sexual functioning, when the true figure is closer to 100 %.

Thus, trial design has generated an agnosia for most of the effects of these drugs. This agnosia has been compounded by a rhetoric that gives the impression that since these drugs have been through RCTs, most of what needs to be known about them is known. We become ignorant of effects that are termed adverse solely because they are not the primary effects being looked at and, in so doing, we compromise safety and reduce the rate of discovery of new drugs.

Atom-Agnosia

The clinical encounter is a relationship, and good care involves close attention to the individual (the 'atom'). RCTs have affected this relationship by treating individual variation as inconsequential. As a result, clinical encounters have become an industrial process, like agriculture, that aims at implementing impersonal algorithms and guidelines, leaving clinicians practising mediculture rather than medicine.

In agriculture, RCTs work and (perhaps until the advent of genetically modified crops), there was little attempt to hide the data. In contrast, in mediculture, RCTs don't work, and treatment guidelines are based on data that are often miscoded and always inaccessible, in ghost-written publications from trials that are not designed to detect many of the significant effects of treatment. The key point is that the individual has vanished, and it is becoming progressively more difficult for any of us to form a genuine relationship with a doctor.

Efficacy and Effectiveness of RCTs

The problems from pharmacognosia to randomisation-induced confounders are not an inconvenience that stem from some oddity to do with antidepressants or suicide. They are intrinsic to RCTs within medicine and can be expected every time a treatment and an illness produce similar or superficially similar outcomes – whether a benefit or a harm.

Similar scenarios unfold with cardiac rhythm problems in trials of anti-arrhythmics, with breathing difficulties in trials of anti-asthmatics, and with vaccines and viral infections that cause brain or other damage. Controlled trials show that ACE-inhibitors improve renal function in patients on diabetes but, in a proportion of cases, they can make renal function worse by aggravating renal artery stenosis. The interaction between the heart attack producing effects of both diabetes and rosiglitazone obscure the adverse effects of rosiglitazone (Cohen 2010). Exenatide and sitagliptin produce pancreatitis, but diabetes can too (Cohen 2013). These

problems happen as much with the effects of treatments termed benefits as with those termed harms.

RCTs were introduced to the regulatory apparatus in an attempt to enhance safety by demonstrating effectiveness. The only time RCTs are unambiguously effective at enhancing safety is when they demonstrate that a drug is inefficacious or ineffective. An example of this is the Women's Health Initiative study of hormone replacement therapy, where a proposed reduction in cardiovascular risk specifically, and mortality in general, was not found and indeed the opposite was found (Women's Health Initiative 2002). With this kind of outcome, the risks inherent in RCTs of a medicine are warranted because the demonstrated effects led to a reduced rather than an increased exposure to unknown effects (see Kukla 2016). It is worth noting that the hormone replacement therapy studies were helpful, but not because they generated the right answer. The negative answer they generated put the onus to back up claims being made where it belonged – namely on those who might make money from vulnerable people.

11.5 The Best Evidence

Evidence-based medicine has become synonymous with RCTs even though such trials fail to tell the physician what she wants to know which is which drug is best for Mr Jones or Ms Smith – not what happens to a non-existent average person (Lasagna 1998).

The verdict of RCTs is often pitted against clinical judgement, despite the fact that an RCT may not be able to show an antidepressant causes suicidality, whereas the exercise of clinical judgement within an RCT can. Clinicians and patients often can distinguish between depression-induced and drug-induced suicidality.

Patients can distinguish between the beneficial effect of a drug and the effect of that benefit on outcomes as when, for instance, patients make it clear that an SSRI is producing a useful emotional numbing but this is not leading to recovery. This information is important, if doctors want to introduce another drug with a different mode of action into the mix, or want to stop the original drug and start another.

As things stand, because RCTs de facto discourage the engagement of doctors with patients, they obscure any specific effects that quite different therapeutic principles might have. In the case of antidepressants or hypoglycemics, RCTs make diverse drugs acting selectively on different systems look exactly the same. This leads to patients being put on drug cocktails because all have been shown to 'work' without any effort to match a therapeutic principle to a patient's needs. This clinical approach prevents doctors finding the right drug for the patient in front of them, and blocks the possibility of insights into the nature of the syndromes they are treating.

If we are to reverse this, there needs to be a focus on safety rather than efficacy. A focus on safety will increase the chance that a pregnant woman will end up on a drug that in fact works for her and, of even greater importance, she will only end up on a treatment she values.

Registries and Objectivity

Given that the challenge-dechallenge-rechallenge protocol is not a reasonable option for pregnant women, there is an urgent need for comprehensive pregnancy registries. The only explanation for the lack of such registries would seem to be the fact that until recently, drug use in pregnancy was minimal. As this changes, women (including pregnant women), and administrators, midwives, doctors and nurses working in antenatal and postnatal services need to work together to put such registries in place (see Ballantyne and Rogers 2016).

Part of the appeal behind RCTs is that when contrasted to individual judgements by patients or doctors, they appear to offer objectivity. But as this exposition of some of the limitations of RCTs reveals, RCTs are largely a mechanical process. This means objectivity must come from elsewhere. In our opinion, objectivity generally results from the ability to bring many points of view to bear on an issue – it results from a collective exercise.

The challenge of objectively establishing whether a treatment causes birth defects is one of the greatest challenges in science, given that the challenge-dechallenge-rechallenge strategy is unavailable and RCTs are not likely to be helpful. When it comes to minimising any bias in registry data, objectivity is most likely to be achieved if we have the greatest possible input from the widest range of sources – including from those whose bias might be thought to be biggest because of the outfits they are currently in (institutions they work for).

References

Ballantyne, A., and W. Rogers. 2016. Pregnancy, vulnerability, and the risk of exploitation in clinical research. In *Clinical research involving pregnant women*, eds. F. Baylis and A. Ballantyne, 139–159. Cham: Springer.

Baylis, F., and R. MacQuarrie. 2016. Why physicians and women should want pregnant women included in clinical trials. In *Clinical research involving pregnant women*, eds. F. Baylis and A. Ballantyne, 17–31. Cham: Springer.

Cohen, D. 2010. Rosiglitazone. What went wrong? *BMJ* 341: c4848.

Cohen, D. 2013. European drugs agency clashes with scientists over safety of GLP-1 drugs. *BMJ* 347: f4838.

Davies, E.B. (ed.). 1964. *Depression: Proceedings of the symposium held at Cambridge, 22–26 September 1959*. Cambridge: Cambridge University Press.

Fergusson, D., S. Doucette, K. Cranley-Glass, S. Shapiro, D. Healy, P. Hebert, and B. Hutton. 2005. The association between suicide attempts and selective serotonin reuptake inhibitors: Systematic review of randomized controlled trials. *BMJ* 330(7488): 396–399.

Fisher, R.A. 1935. *The design of experiments*. London: Macmillan.

Gigerenzer, G., Z. Swijtink, T. Porter, L. Daston, J. Beatty, and L. Kruger. 1990. *The empire of chance: How probability changed science and everyday life*. Cambridge: Cambridge University Press.

GlaxoSmithKline, 2006. Paroxetine adult suicidality analysis. GlaxoSmithKline. Updated 5 April 2006; originally cited 8 August 2007. On file with and available from the authors.

Hagnell, O., J. Lanke, and B. Rorsman. 1981. Suicide rates in the Lundby study: Mental illness as a risk factor for suicide. *Neuropsychobiology* 7(5): 248–253.
Healy, D. 1998. *The antidepressant era.* Cambridge, MA: Harvard University Press.
Healy, D. 2012. *Pharmageddon.* Berkeley: University of California Press.
Hill, A.B. 1966. Reflections on controlled trial. *Annals of Rheumatic Diseases* 25(2): 107–133.
Kukla, R. 2016. Equipoise, uncertainty, and inductive risk in research involving pregnant women. In *Clinical research involving pregnant women*, eds. F. Baylis and A. Ballantyne, 179–196. Cham: Springer.
Langston, L. 2016. Better safe than sorry: Risk, stigma, and research during pregnancy. In *Clinical research involving pregnant women*, eds. F. Baylis and A. Ballantyne, 33–50. Cham: Springer.
Lasagna, L. 1960. Thalidomide – A new non-barbiturate sleep-inducing drug. *Journal of Chronic Diseases* 11(6): 627–631.
Lasagna, L. 1998. Back to the future. Evaluation and drug development. In *The psychopharmacologists*, vol. 2, ed. D. Healy, 135–166. London: Arnold.
Marks, H.M. 1997. *The progress of experiment: Science and therapeutic reform in the United States, 1900–1990.* Cambridge: Cambridge University Press.
Medical Research Council. 1948. Streptomycin treatment of pulmonary tuberculosis. *BMJ* 2(4582): 769–782.
Raffle, A., and M. Gray. 2007. *Screening: Evidence and practice.* Oxford: Oxford University Press.
Savage, L.J. 1976. On rereading R.A. Fisher. *The Annals of Statistics* 4(3): 441–500.
Teicher, M.H., C. Glod, and J.O. Cole. 1990. Emergence of intense suicidal preoccupation during fluoxetine treatment. *The American Journal of Psychiatry* 147(2): 107–210.
Toth, B. 1998. Clinical trials in British medicine 1858–1948, with special reference to the development of the randomised controlled trial. Bristol University PhD dissertation.
Verkes, R.J., R.C. Van der Mast, M.W. Hengeveld, J.P. Tuyl, A.H. Zwindermann, and G.M. Van Kempen. 1998. Reduction by paroxetine of suicidal behavior in patients with repeated suicide attempts but not major depression. *The American Journal of Psychiatry* 155(4): 543–547.
Women's Health Initiative (Writing group for the Womens' Health Initiative Investigators). 2002. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized controlled trial. *JAMA* 288(3): 321–333.