Controlled Trials & Adverse Events: Lessons from the History of Antidepressants & Suicide

David Healy MD FRCPsych

Introduction

Randomized controlled trials (RCT) have made critical contributions to the benefit of healthcare. But perhaps because their contributions are so celebrated, they are widely cited as a gold-standard in a manner that implies they will semi-automatically provide the best possible answers for almost any problem, in particular any issue to do with drug treatment . In contrast, while good clinical judgement once played a key role especially in delineating treatment related adverse events, when compared with RCT data clinical judgements are increasingly likely to be dismissed as anecdotal. On the basis of an apparent absence of evidence from trials about specific adverse effects, doctors and patients faced with these side effects are increasingly called on to doubt the evidence of their own eyes.

There are several problems with an uncritical approach to the benefits of RCTs insofar as they relate to treatment induced adverse events. Some of these problems stem from a series of inappropriate data-management strategies that are relatively widely known.

For instance, it is known that in pharmacotherapy trials the side effects of a drug may be coded under disparate headings. For example, suicidal acts may be coded under a variety of headings such as anxiety, agitation, akathisia, emotional lability, thinking abnormally, abnormal dreams, psychosis and others. This may happen unintentionally, but nevertheless such miscoding divides and conquers what may be a problem for a pharmaceutical company. Instances of this have been described for many adverse events.

The coding of data by pharmaceutical companies under diverse headings means that any tabulation of treatment related events that might appear on sites like clinicaltrials.gov will be necessarily suspect unless there is access to the raw data. Data in this case means information at the level of the individual patient rather than a listing of events.

In addition to miscoding, treatment related events may be mislocated. Many trials have a washout period (sometimes called a placebo run-in phase) lasting a week or two before randomization proper. This is a period where patients may be asked to stop prior antidepressants or other treatments. It is now clear that this an extremely hazardous period, owing possibly to the withdrawal effects of prior treatments.

In the initial trials of selective serotonin reuptake inhibiting (SSRI) antidepressants fluoxetine, paroxetine and sertraline, suicidal events that occurred during this washout period were later filed inappropriately under the heading of those randomized to placebo (Healy 2006). Events from the post-trial follow-up period that have occurred in patients previously on placebo but after randomization put on active agents have also been filed under the heading of placebo (Healy 2012). Mislocations of this sort have been detected not just for suicide related adverse events on antidepressants but also for heart attacks on treatments like rofecoxib (Vioxx) and rosiglitazone (Avandia).

A much greater number of events are hidden by an inappropriate reading of significance testing. Where a substantially increased rate of adverse events on drugs does not reach statistical significance for one reason or another (so that there is >5% probability that the findings could have arisen by chance), company rhetoric assumes that they did arise by chance and that as an increase in risk has not been conclusively demonstrated, there is in fact no increase in risk (Healy 2006; Healy 2012). This approach avails of a "doubt is our product" dynamic and has hidden a much greater number of suicidal acts in the case of the antidepressants, and heart attacks in the case of Vioxx, than was ever hidden by mislocating suicidal acts from washout periods to placebo.

Despite the miscoding of data, mislocating of events and misuse of statistical significance testing, the faith of most people in controlled trials remains unshaken. This paper outlines a further set of difficulties with controlled trials as these are used for the detection of adverse events. These problems stem from an interplay between the disease being treated and the effectiveness of treatment.

It's the Disease not the Drug

There are two ways in which a disease being treated can come into play to conceal treatment induced adverse events. One lies in disease heterogeneity and the other in the variable effectiveness of treatment.

1. Disease Heterogeneity

In the late 1980s, Lilly undertook a trial of fluoxetine (Prozac) in a group of patients with what is variously termed borderline personality disorder, intermittent brief depressive disorder or recurrent brief depressive disorder. In this trial placebo was sweepingly statistically superior to Prozac. The study was published four years later shorn of its data except for the broad claim that the numbers of suicide attempts in the fluoxetine and placebo groups were the same (Montgomery et al 1994).

In the early 1990s, SmithKline Beecham undertook a study of a closely related SSRI antidepressant paroxetine (Seroxat, Paxil) in the same hospital centre, in the same diagnostic group of patients, possibly with some of the same patients (protocol 106). This was terminated early, and the results were never published. The rate of suicidal acts on paroxetine was three-fold higher than on placebo (Data available from the author).

Several years later SmithKline Beecham undertook another trial (protocol 057) in a similar group of patients (Verkes et al 1998). There are several extant sets of figures from this study; the figures used in this article are one set, published by SmithKline Beecham.

In April 2006, GlaxoSmithKline issued a press release with the following figures for suicidal acts in the trials in their most important therapeutic area, Major Depressive Disorder. Conceding that there was a risk of suicide for this patient group was extremely problematic for the company.

This study of patients with Major Depressive Disorder showed a statistically significant increase in the risk of a suicidal act on paroxetine. The full press release, however, combined the patients from protocols 106 and 057 (IBDD trials) with patients from the MDD trials.

Table 1: Suicidal Acts in Major Depressive Disorder Trials

Major Depressive Disorder Trials (MDD)	**Paroxetine**	**Placebo**	**Relative Risk**
Number of Suicidal Acts /Number of Patients	**11/ 2943**	**0/ 1671**	**Inf (1.3, inf)**

The key point about including these intermittent depressive disorder patients is that the company could categorize them as being depressed and when two datasets, from Major Depressive and Intermittent Brief Depressive Disorder trials, are added together the increase in risk in Depression not only vanishes but paroxetine becomes apparently protective against suicide risk.

Table 2: Suicidal Acts in Major Depressive Disorder & Intermittent Brief Depressive Disorder Trials

	Paxil	**Placebo**	**Relative Risk**
MDD Trials			
Acts/Patients	**11/2943**	**0/1671**	**Inf (1.3, inf)**
IBDD Trials			
Acts/Patients	**32/147**	**35/151**	**0.9**
Combined Trials			
Acts/Patients	**43/3090**	**35/1822**	**0.7**

The IBDD group is a patient group who make regular suicide attempts – some IBDD patients make several attempts a week. This fact means that even if fluoxetine and paroxetine do not reduce the numbers of suicide attempts in these patients these trials can be useful for companies. Indeed even if there was an increase in the number of suicide attempts on active treatment the studies would still be astonishingly useful. In these two protocols (106 and 057), 298 patients had 67 suicidal acts between them – this is 100 times more suicidal acts than in the entire set of major depressive disorder trials

(In fact a handful of patients in 106 and 057 had close to half the suicidal acts between them). This works to hide the problem because of a variation on what has been termed Simpson's paradox (Cates 2002). Simpson's paradox arises when collapsing trials together based on the simple addition of all events leads to a reversal of the direction of effects seen in a majority of studies (Cates 2002). This is most likely if the event rates in studies differ markedl – as they do in the MDD and IBDD studies outlined here.

In the case of protocols 106 and 057, the timing of these studies is interesting. It is quite possible that some of those involved in the design of these studies had accepted that SSRIs like paroxetine cause suicide and embarked deliberately on a series of studies that used a problem the drug causes to hide a problem that the drug causes.

But whether deliberate or not in this case, something similar is possible in principle in single studies or combinations of studies if the clinical population recruited is heterogeneous in respect of the key adverse event. It would be seem quite possible to view a number of IBDD patients has having MDD and recruit them to an MDD trial and for the contribution from these patients to raise the background placebo rate, thereby concealing the adverse event.

Unless the adverse event, be it a respiratory, gastro-intestinal, rheumatological or other system event is fully understood and in particular its response to treatment, the heterogeneity of such populations makes it possible that problems triggered by treatment will not emerge as clearly linked to treatment.

2. Treatment Related Effect Modification

In 1990 concerns about a suicide risk of antidepressants arose with the publication of case studies in which suicidality emerged on fluoxetine, cleared when treatment stopped and re-emerged on the reinstitution of treatment. These reports fulfilled all the standard canons for determining cause and effect at the time (Healy 2004). However the field was swayed instead by arguments that the clinical trial data showed no risk, even though the trial data showed a clear increase in risk on active treatment but this increase in risk was not statistically significant.

Over 15 years later when a sufficiently large number of trials were assembled, and rates of suicidal acts on antidepressants such as the selective serotonin reuptake inhibitors (SSRIs) did show a statistically significant increase in the relative risk of a suicidal act compared to placebo, FDA officials stated that this demonstrated a causal effect.

Causality was conceded as since the 1980s and a series of legal cases involving breast implants, there has been a de facto medico-legal convention that a demonstration of cause and effect requires a statistically significant doubling of the risk of the adverse event in question (Angel 1997).

The first tricyclic antidepressant imipramine was introduced in 1958. At a meeting in Cambridge in 1959, several participants stated on the basis of clinical observations involving the emergence of the problem on exposure to the drug (challenge) and clearing up of the problem on discontinuation of treatment (dechallenge) that imipramine could directly cause suicide by increasing agitation. There was no dissent (Davies 1962).

There is considerable clinical trial evidence that imipramine, clomipramine and other tricyclic antidepressants are more effective than SSRIs. Specifically, they are effective in melancholic depressions where SSRIs are not. This means that they are therefore effective in a patient group at a substantially higher risk of suicide than those outpatient or primary care depressed patients entered into SSRI trials (Healy 2004).

It follows from this that in a putative placebo controlled trial of imipramine in melancholia the rate of suicidal acts in the placebo arm would be higher than in SSRI trials and the extent to which imipramine lowered the rate of suicidal acts by successful treatment of melancholia would be much greater than the rate SSRIs may have lowered suicidal acts in the largely primary care depression trials these drugs were studied in. From this it follows that the relative risk of a suicidal act on imipramine or other tricyclic agents in such trials might well be less than 1.0, perhaps as low as 0.5.

On the basis of RCT data where the relative risk was less than 1.0, many academics and regulators like FDA would not be prepared to concede that the drug being tested could cause suicide, although using the criteria still embodied in standard textbooks of adverse event causality – namely challenge-dechallenge and rechallenge along with dose-responsiveness - these older serotonin reuptake inhibiting drugs unquestionably do cause suicide.

Epistemological Problems

The notion that a treatment that on the one hand causes a problem might in good faith trials, without any manipulation of the data or statistical artefact, give rise to a relative risk < 1.0 poses a conundrum. The possibility has been noted in abstract terms, but to the best of our knowledge has not been specified as baldly as here

(Greenland and Robbins 1988; Lanes1999; Rothman et al 2008). This raises several epistemological issues.

First, giving a primacy to RCT data over other data throws us into the unusual position of having to concede that the problems which the SSRIs induced only came to light as a result of the artefact of their being tested in populations at minimal risk of suicide.

Second, given what we now know about the behaviour of an effect like suicide, we can construct studies to make it appear or disappear but knowing about an effect to this extent raises the question as to what exactly RCTs establish. Where an adverse event is as well understood as suicide on antidepressants it seems that it might be possible in some circumstances to design trials to produce any predetermined relative risk between 0.1 and 10.0.

Third, in the case of a relative risk of a suicidal act in depression trials, if the drug has also been effective and helped some of those with illness linked suicidality, the relative risk is in fact a compound of risks. The treatment induced component of that compound is therefore almost certainly greater than the relative risk suggests, but we have no way of knowing by how much greater it is.

We can map out the dilemmas in the case of the antidepressants and suicide because these drugs and this problem are relatively well characterized. Comparable scenarios can be mapped out for some arrhythmias on anti-arrhythmics, for beta agonists given for asthma, as well as for certain vaccines. In principle such difficulties will potentially arise in every case in which both an illness and its treatment give rise to at least superficially similar problems. But if the adverse effect and the treatments are not well understood the results that emerge from these trials become impossible to interpret other than to say these are the data that emerged from this particular assay.

A "blame the illness" dynamic feeds into a clinical bias to see problems as stemming from diseases rather than their treatments. Perhaps recognizing the merits of this defence, companies marketing antipsychotics in recent years faced with elevated rates of diabetes argued that schizophrenia gives rise to diabetes without any evidence to support their position, and found this defence worked (Le Noury et al 2008). Almost anything it seems can be portrayed as a risk of the illness.

Remedying the Problem

There are a number of possible remedies to the problems outlined above, some of which involve RCTs and others that do not.

RCT solutions:

First it must be noted that a great number of these problems arise in RCTs where the adverse event in question is not the primary outcome. They arise in trials therefore not designed to look specifically at this issue. One option is to simply say that if a study has not been designed to look at the issue in question, we have in fact no good clinical trial evidence on the issue. A failure to do this, risks compromising the credibility of RCTs.

A further set of RCTs have the capacity to reveal adverse events without the confounding or effect modification that may stem from an associated illness. These are the phase one or healthy volunteer trials that companies and universities conduct. At present clinical trials in patient groups (phase 2 and 3 trials) are registered so we know what trials are happening even though we cannot get the data from these trials. But there is no register for phase 1 studies. And where access to the data from clinical trials remains problematic because of patient confidentiality, there should in principle be no problem with access to the data from healthy volunteer studies.

Phase 1 studies can be surprisingly powerful. Consider a never-published study of 12 volunteers conducted in Leeds in 1983, 9 years before the antidepressant sertraline was marketed, and 21 years before FDA required it to carry suicide warnings. This study of sertraline was terminated early because all of the women on sertraline had become anxious or apprehensive, noting thoughts of aggression and related difficulties. Pfizer concluded that sertraline had caused these changes. One of the senior investigators further noted that comparable results had been seen in healthy volunteers for other SSRIs then in development. There are in fact several known healthy volunteer suicides and episodes of violence in the phase one studies of SSRIs.

In order to detect an adverse event like suicide, the event must exceed the base rate in the untreated clinical population. The salience of the event is therefore much more marked in a healthy population. This is the reverse of the situation with protocol 057 and 106.

Non-RCT Solutions:

1. A credibility index

With adverse events that stem from both an illness and its treatment, the question is what weight to put on observations from controlled trials that have not been designed to investigate the issue but give rise to a non-significant increase or decrease in the relative risk for the event versus observations that incorporate challenge, dechallenge and rechallenge (CDR) relationships, along with evidence of dose-responsiveness, and reversal by antidote.

One doctor who reports that a patient develops an adverse event on treatment and who, because of CDR, dose response and other relationships, links this to treatment, might not be believed. If, however, a thousand doctors outline similarly good quality reports (and even more so, if each knows there are 999 other reports) the field is likely to believe the outcome. The question is therefore, where do we cross the credibility threshold for believing reports like this – is it 5 good reports, 10 or 100 reports? What weight should be put on such reports as compared with data from RCTs where the event in question has not been the primary outcome measure?

In the case of antidepressants and suicide, 6 good clinical observations from one clinical centre turned out to be correct, but made no difference to the general perception of the issues. There is an interesting cognitive bias in the case of this problem in that events like suicidal acts and violence seem much less likely to be attributed to prescription drugs than dependence and withdrawal for instance. In the case of suicide on antidepressants this bias was not overcome even when the clinical trial evidence points to a risk, although by now a lot of other issues have clustered around this linkage.

2. A toxicity index:

Current antidepressant trials run for approximately six weeks and involve changes on rating scales such that a collection of side effects could give an apparent benefit. In the case of the antidepressants some of the side effects include sedation and increased appetite, and these side effects will lead to a lowering of the depression rating scale scores. When patients complete quality of life scales in contrast the benefits for antidepressants are not found.

In addition, the FDA currently licenses drugs on the basis of two positive trials, even if such trials are nested among a larger number of negative trials. The FDA concedes that this process means that, in the case of the antidepressants. all we have is the signal of a treatment effect rather than a demonstration of effectiveness.

With the current system, it would seem entirely possible to put alcohol, nicotine, diazepam or dexamphetamine through the system and get approval as an antidepressant. Giving any of these for 6-8 weeks would probably not cause significant clinical problems and indeed might cause fewer problems than the SSRIs have caused in these trials.

Because of their familiarity with alcohol, and with the reputation of these other drugs, most people would know that taking these "antidepressants" regularly beyond 6 weeks might not be a great idea. But most patients and many doctors are disarmed by the current testing processes, which transform prescription-only drugs (on prescription-precisely because they may turn out to be risky) into risk-free drugs. What is needed is a metric that assumes *ab initio* that novel agents will come with problems such as dependence and other consequences if used in the longer term.

We have stepped back from viewing new drugs as poisons to be treated warily on the basis that their risks have not yet been demonstrated and in practice we increasingly assume that the lack of evidence means these new drugs pose no risks. It would be more in keeping with traditional clinical practice during the early life of a drug to assume a range of hazards are likely to happen, with an appropriate adjustment made later, only if it transpires a drug is safer than average.

References

Angel M (1997). *Science on Trial.* Norton, New York.

Cates CJ (2002). Simpson's paradox and calculation of number needed to treat from meta-analysis. *BMC Medical Research Methodology* wwwbiomedcentral.com/1471-2288/2/1

Davies EB (1962). Proceedings of the symposium held at Cambridge, 22-26 September 1959. Cambridge University Press.

Greenland S, Robins JM (1988). Identifiability, exchangeability, and epidemiological confounding. *International J of Epidemiology* 15, 412-418.

Healy D (2004). *Let them eat Prozac.* New York U Press, New York.

Healy D (2006). The Antidepressant Tale: Figures Signifying Nothing? *Advances in Psychiatric Treatment* 12, 320-328.

Healy D (2012). *Pharmageddon.* U California Press, Berkeley.

Lanes SF (1999). Biological interpretation of relative risk. *Drug Safety* 21, 75-79.

Le Noury J, Khan A, Harris M, Wong W, Williams D, Tranter R, Healy D (2008). The incidence and prevalence of diabetes in patients with serious mental illness in North West Wales: Two cohorts 1875-1924 and 1994-2006 compared. *BMC Psychiatry* 8: 67. doi:10.1186/1471-244X-8-67.

Montgomery DB, Roberts A, Green M, Bullock T, Baldwin D, Montgomery SA (1994). Lack of efficacy of fluoxetine in recurrent brief depression and suicidal attempts. *Eur Arch Psych ClinNeurosci*, 244, 211-215.

Rothman KJ, Greenland S, Lash T (2008). *Modern Epidemiology*. Lippincott, Williams & Wilkins, Phila Pa.

VerkesRJ, Van der Mast RC, Hengeveld MW, Tuyl JP, Zwindermann AH, Van Kempen GM (1998). Reduction by paroxetine of suicidal behavior in patients with repeated suicide attempts but not major depression. *American J Psychiatry* 155, 543-547.

David Healy, Bangor University.

Email: david.healy54@googlemail.com