transcultural psychiatry

March 2009

ARTICLE

Trussed in Evidence? Ambiguities at the Interface between Clinical Evidence and Clinical Practice

DAVID HEALY
Cardiff University

Abstract This article considers the dominance that randomized controlled trials (RCTs) of psychotropic agents currently have in relation to the practice of psychiatry in mental health and primary care settings. In contemporary psychiatry, data of marginal significance based on rating scale measures are privileged as evidence that treatments are effective, while judgments of drug effects based on clinical practice are downgraded. The dominance of RCTs has also led to an increasing promotion of rating scales in clinical practice, described here as 'rating scale mongering.' The logical consequence of current interpretations of RCT data is that clinicians should adhere to guidelines which are based on a systematic assembly of such data, but the selective publication of trial data and ghostwriting of publications, lays the basis for guideline capture, and a corresponding capture of evidence-based clinical practice by pharmaceutical companies.

Key words guideline-capture • informational reductionism • treatment effects • rating scale mongering

BACKGROUND

The discovery of breakthrough drugs for psychosis and mood disorders went hand in hand with a development that many thought would be even more valuable: the discovery of clinical trials. Regulatory developments in

Vol 46(1): 16–37 DOI: 10.1177/1363461509102285 www.sagepublications.com

the 1960s forced pharmaceutical companies to prove that their drugs worked by adopting clinical trials. Companies would peddle no more snake oil; everything licensed would have been proven to work.

The new drugs and trials entered a medical world in which clinical judgments were based on medical experience and were not readily questioned. The primary care or private practitioner rarely read medical journals, was slow to prescribe the latest drug, unless like penicillin it was clearly life-saving, but knew the patient and his family and community. If a new drug for 'nerves' was prescribed, and the patient came back with an unusual side effect, treatment was stopped on the basis of the physician's common sense; there was no evidence base to consult. There was also significant variation in medical care. For example, some clinicians prescribed antibiotics for ulcers, on the basis that while this treatment did not appear in the books, patients receiving it did not return, and the alternative was major surgery (Marshall, 2002). Such variations in care are now seen as problematic.

Today's clinician is brought up in an evidence-based world, reads the latest journals, adheres to guidelines, and typically prescribes the newest drugs. These clinicians have been trained to communicate, and to detect depression, social phobia, obsessive-compulsive disorder, generalized anxiety disorder, panic disorder or post-traumatic stress disorders, rather than to manage 'nerves.' When clinicians now prescribe the latest psychotropic drug it may be with reassurances about correcting chemical imbalances. If a patient wonders whether a new problem could be linked to treatment, the clinician will consult a computer to access what is known from controlled trials about this problem. If there is no record of this drug causing this problem, the patient will be told this, and probably advised to persist with treatment.

The new world of clinical trials produced benefits by transforming patients into informed consumers. Prior to the advent of clinical trials, there was no counterweight to medical opinion. While clinical trials have put evidence into the public domain where doctors can access it, they have also enabled patients to question the authority of their doctors. The status of physicians' authority has been tempered, so that they are now viewed as experts who know more about the data than the patient, and who ideally use this knowledge to guide their patients, but are constrained by the fact that the patient had some basis to contest issues.

While the evidence base of medicine clearly benefited from these developments, there is nevertheless a growing sense of crisis in healthcare. While many factors may contribute to this sense of crisis, as Porter argues in his 1996 book on the use of measurement in social disputes, the social response to uncertainty is increasingly a turn to the pursuit of measurable objectivity rather than to professional discretion. In medicine, such a

dynamic appears to underpin many pleas for an evidence-based clinical practice, if necessary, a practice constrained within guidelines, and an increasing advocacy of the use of rating scales in clinical practice.

However, clinical trial data are increasingly linked to pharmaceutical companies and this data appears shot through with problems stemming from the non-reporting of trials or ghostwriting of those that are published. As this article will indicate, because of these ambiguities, it is not inconceivable that an ever-closer adherence to what may appear to the best evidence could lead to a deterioration in the health of patients.

Follow the Evidence

It is now widely assumed that randomized controlled trials (RCTs) provide evidence for whether a treatment works. But far from being a method to prove that treatments worked, RCTs were initially designed to weed out treatments that did not work. For treatments that unquestionably do work, such as penicillin for tertiary syphilis or bacterial endocarditis, RCTs are not needed. However, when the outcome of a set of trials makes it neither possible to say that this agent does nothing nor that it works (in the sense of restoring a significant number of people who take it to full health), we are in much less certain waters than is generally realized.

Several meta-analyses of antidepressant trials (Kirsch, Moore, Scoboria, & Nicholls, 2002; Kirsch & Sapirstein, 1998; Stone & Jones, 2006) suggest that roughly 50% of patients entering published antidepressant trials have a response as measured on a rating scale like the Hamilton Rating Scale for Depression as compared with 40% of those who are given the placebo. The difference between the active drug and placebo is represented in Figure 1.

A difference between active drug and placebo that is statistically significant is taken to indicate that the drug 'works.' Regulators approve such drugs, drug companies market them as effective, and clinicians prescribe them. But if the trials are sufficiently large, even a minor difference of one or two rating scale points can be made statistically significant. As a result of this, a drug, which is a little bit sedating or tranquilizing, will show up as 'working for depression' if the rating scale includes sleep or anxiety items.

On this basis it would be possible to prove nicotine, benzodiazepines, anti-histamines, methylphenidate, or other treatments for ADHD, and most of the antipsychotics, and a number of anticonvulsants, to be 'antidepressants.' Indeed, many of these diverse agents have RCT evidence of benefit in depression (see Robertson & Trimble, 1982).

The key difference between this diverse group of drugs and the drugs that came to be thought of as antidepressants is that the 'antidepressants'

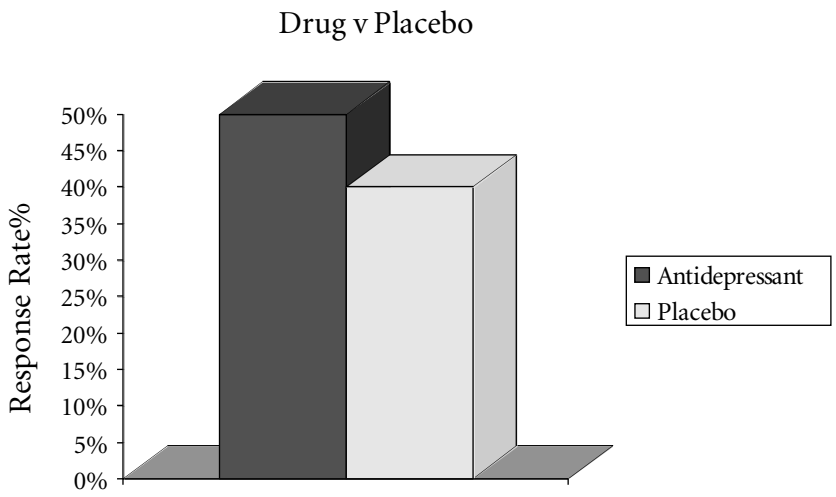

Figure 1 The difference between the active drug and placebo.
Source: The data for this figure stem from the FDA's review of antidepressants drugs (Stone & Jones, 2006).

were newly patented for treating 'depression,' while drugs like nicotine or the antihistamines were unpatentable for this purpose, and other agents such as the antipsychotics and anticonvulsants were targeted for other markets. In contrast, the 'antidepressants' were heavily marketed as treatments for depression, although they had little more than these other drugs to recommend them for primary care nervous problems, and in the process they replaced other treatments (Healy, 2004).

Such a claim sits uneasily with the supposed chemical imbalance that antidepressants fix. No one claims that nicotine, methylphenidate, benzodiazepines or antipsychotics fix this imbalance. While this claim may have helped generate the perception that RCTs provide evidence that 'antidepressants' 'work' for 'depression' rather than evidence that these drugs do something, there is little more than marketing myth to this chemical imbalance (Healy, 1997).

A Cautionary Tale

The ambiguous meaning of the word 'works' in this context is brought out by one of the first controlled trials in medicine, which compared reserpine to placebo in a group of anxious depressives (Davies & Shepherd, 1955). Reserpine worked, in the sense that it caused improvements on rating scales that were greater to a statistically significant degree than did placebo, and it did so to the same extent that fluoxetine later did. Reserpine was

not penicillin, but it was something more than snake oil. However, Shepherd's trial of reserpine, published in the *Lancet*, had almost no impact. His later view as to why the message sank without trace was that medicine was still dominated by physicians who believed the evidence of their own eyes or received their information from clinical articles describing cases in detail – 'anecdotes' (Shepherd, 1998).

On the basis of just such detailed descriptions, instead of becoming an antidepressant, reserpine became viewed as a drug that caused depression and triggered suicides. Reserpine was also used to treat hypertension, and when treated with it, many hypertensive patients felt better than well (Healy & Savage, 1998). But while it suited many people, it did not suit all. In the same issue of the *Lancet* in which Shepherd's trial was reported, the two preceding articles reported hypertensive patients becoming suicidal on reserpine (Smirk & McQueen, 1955; Wallace, 1955). Reserpine can induce akathisia, a then-unknown complication of treatment (Healy & Savage, 1998). The case reports of this new hazard were compelling and the endpoints are so clear-cut that clinical trials are not needed to demonstrate the phenomenon.

In the 1950s, the benefits or hazards of drugs were discovered by clinicians giving these new drugs to patients and paying close heed to what happened; discoveries did not happen in clinical trials. Lithium was discovered by giving it to 10 people who had mania (Cade, 1949). Imipramine and all its effects were outlined after 40 patients had been given the drug (Kuhn, 1958). The first articles on chlorpromazine outlined its effects on a series of 38 patients (Delay & Deniker, 1952). When a treatment has a substantial effect, either beneficial or hazardous, there is no reason to think that case descriptions of this sort are not an appropriate mode of discovery.

In the case of reserpine, Smirk and McQueen and Wallace's case reports trumped Shepherd's RCT. However, the key point is that even though their results were superficially contradictory, there is no reason to think that either the case report or RCT findings were wrong. Indeed, there is a close parallel here with early trials showing that fluoxetine 'worked' and subsequent clinical reports of suicidality induced by its use (Healy, 2004). Had Shepherd's trial registered more widely, the view that depression involves a lowering of serotonin that antidepressants normalize would have been less likely to become dominant, for the simple reason that reserpine lowers serotonin but still has antidepressant effects and can leave many people feeling better than well.

Since 1955, the pendulum has swung away from clinical reports and in favor of RCTs. Many leading journals, such as the *British Journal of Psychiatry*, are now reluctant to publish case reports. The Cade, Kuhn and Delay articles might not now be published in a major journal. A key factor

in this change has been the emergence of the view that RCT evidence trumps all other evidence, especially that from case reports. This is the case even when clinical reports outline effects that follow a challenge-dechallenge-rechallenge protocol and when these effects are explicable in terms of known pathological mechanisms, as was the case for fluoxetine and suicide (Teicher, Glod, & Cole, 1990). This view of the relative merits of RCTs was reduced to absurdity in an article proposing that parachutes should not be used as clinical trials had not proven them to be beneficial (Smith & Pell, 2003).

Follow the Evidence: An Alternate Scenario

In contrast with trials from some other areas of medicine, psychotropic trials do not show evidence of lives saved or people returned to work. The data offer evidence that the drugs have an effect which might be consistent with them 'working' in some people. This use of the term 'treatment effect' stems from FDA approval hearings for antidepressants. Such a definition seems reasonable, given that only one in two antidepressant trials shows a positive result, that regulators have described these trials as assay systems rather than demonstrations of what is likely to happen in the real world, and that some beneficial effects may co-occur with a worsening in the patient's condition. In short, treatment efficacy and effectiveness for these drugs remains to be demonstrated.

Another way to read the data is that these trials allow us to quantify the contribution a drug makes in the treatment of a group of people. The placebo response provides a useful example. It is known that the natural history of depression means that many people will improve within a few weeks whether treated or not. It is also widely thought that sensible advice from a clinician on matters of diet, lifestyle, alcohol intake, and work and relationship problem-solving may make a difference. It is suspected that patient perceptions that they are being seen and cared for by a medical expert may make a difference, and this effect may be enhanced by being given a substance they think will restore some chemical balance to normal – even if that imbalance is mythical and the substance is a placebo. The fact of presenting for treatment may make a difference. All of these factors are reflected in the placebo response, but it is not possible to quantify the distinct contribution of these components, how much for example the natural history of the disorder contributes compared to advice about lifestyle.

These factors also contribute to the therapeutic response for those on an active drug, but in contrast to the difficulties in quantifying the components of the placebo response, RCTs allow us to quantify the contribution made by the drug. The specific drug effect in Figure 1 is 50%

– 40% = 10%. That is, four out of five, or 80%, of responders would have improved had they received the placebo. Of responders, 20% have a specific response to the drug. The number of patients needed to treat (NNT) to produce one specific drug response therefore is 1/10% = 10. Far from proving that the drug in question works, the data presented in Figure 1 is not inconsistent with the possibility that the drug on balance does more harm than good.

In contrast, the NNT for placebo is 1/40% = 2.5. That is two in every five treated with a placebo show a response. When the data for response is recast as in Figure 2, it becomes clear that if we are to follow the evidence we should ensure as good a placebo response as we can by trying to amplify the therapist's magic, by trying to ensure patients get the very best lifestyle advice and reasonable problem-solving input, and we should have a greater resort to judicious waiting. While most clinicians probably appreciate this point in the abstract, in practice few resort to judicious waiting or caution patients that an apparent response may not be drug induced.

The New Anecdotes

Physicians and others using or advocating psychotropic drugs behave as though the trial data for these drugs look like the data from trials of penicillin for fulminating pneumonia might look. This reflects a failure to

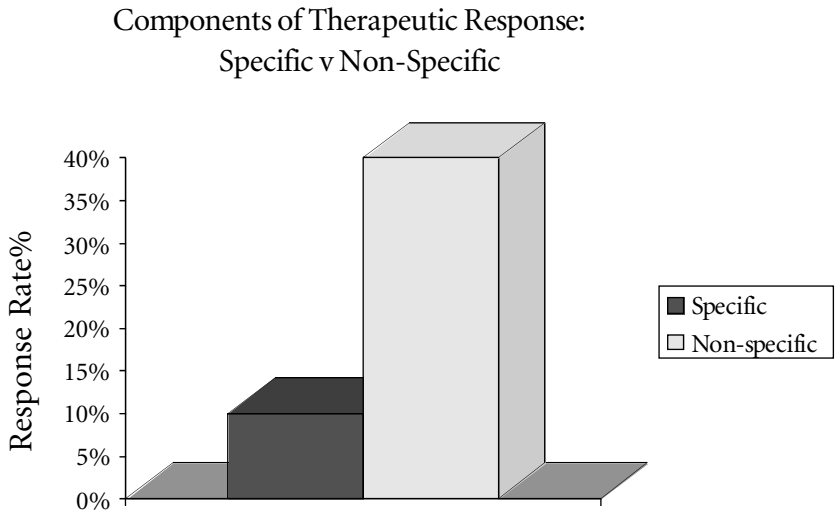

Figure 2 Components of therapeutic response: specific vs. non-specific.
Source: The data for this figure stem from the FDA's review of antidepressants drugs (Stone & Jones, 2006).

recognize that the outcomes for a treatment like penicillin are not based on rating scale scores but on the numbers of dead bodies or significantly scarred lungs in the active treatment group compared with the placebo group. There are more deaths and residual disability on placebo, which is the reverse of what is found in psychotropic trials. (The role that guidelines may play in generating such misperceptions is developed later.)

If we return to the data in Figure 1, we see that a further 50% of patients do not respond to the specific treatment. In preferentially accepting RCTs, as they are currently framed and interpreted, over case reports journals risk privileging the experiences of the single specific drug responder over the nine-fold larger pool of other responders or non-responders. Preferentially publishing such clinical trials over case reports therefore appears to elevate a new form of anecdotalism. The difficulties with the evidence base are compounded by a number of additional factors: many of these trials are ghostwritten (Healy & Cattell, 2003); only selected data are reported from selected trials and even these data may be misreported (Healy, 2006b); and the significance of the data reported is generally misinterpreted (Healy, 2006a).

These developments have consequences for all parties to therapy. For journals, there remains the fact that the first discoveries of a new benefit or hazard are more likely to come in the form of a series of cases than in an RCT. For clinicians, they are the shift from a world in which they were slow to use new drugs and conservative when they did so, to one in which clinicians rapidly take up the newest treatments, driven by evidence that is less generalizable than commonly thought. Faced with a patient experiencing an unusual effect, they consult the RCT evidence base, which often will not list such effects. In the case of patients who became suicidal while taking selective serotonin reuptake inhibitors (SSRIs), until very recently, the evidence base would not have listed this effect, or might have listed it under a code such as 'emotional lability.' Failing to see evidence of a hazard, the clinician may double the dose of the new agent.

Current data suggests that patients offered an 'antidepressant' have a 50% chance of receiving a drug that will do nothing for them, and a 50% chance of receiving a treatment that will do little more than nicotine, an antihistamine or methylphenidate would do. Antidepressants are not magic bullets, the equivalent of insulin or penicillin, but almost all the information available within the therapeutic arena is likely to obscure this fact. More generally, if the minor benefit of antidepressants had been more clearly recognized, it would have been apparent what the next scientific step should be: determining who would benefit from the type of manipulation offered by methylphenidate or a serotonin reuptake inhibiting antihistamine. Instead, huge efforts were made to get patients off of benzodiazepines, on which they may have been doing quite well, or off

of methylphenidate, which was once widely advertised for depression, and onto an SSRI.

Because evidence that a drug 'works' for depression or schizophrenia is all that is required to legally market it as an antidepressant or an antipsychotic, companies have no incentive to subsequently research out which depressives or which psychotics their drugs work best for. While virtually the entire literature on the effects of antipsychotics in schizophrenia made it clear that although some clinical presentations had an over 80% response rate to antipsychotics, others have a less than 10% response (see Delay, Deniker, & Ropert 1955; Fish, 1964), this is not recognized by modern clinical practice. Instead, the RCT evidence that antipsychotics 'work' for schizophrenia makes it difficult not to give these drugs, and to give them in ever-larger doses, precisely to those patients who fail to respond.

One criticism in response to this argument might be that undermining a patient's belief in a treatment is not a good thing. This argument brings out the journey biological psychiatry has traveled in recent years. Formerly the magic was in the therapist; he or she might also give pills, but these were an extension of his or her impact on us. Now the magic has passed into the capsule and the physician is little more than a conduit for medication. Therapists have forgotten how to manipulate their impact on patients. With the focus both doctor and patient now have on the pill, neither heed the context in which the patient has become distressed. Neither appears to see how small a contribution this chemical manipulation is likely to make, or to see the potential for a chemical manipulation to make things worse.

From Anecdotes to Data

One reason that antidepressants have been so commercially successful is that their lack of generalizable efficacy and their hazards are not apparent in journal articles. Clinical trial results are reported as 'significantly better than placebo' and safety results are reported as 'not statistically different from placebo.' These statements are simultaneously true and misleading (Healy, 2006a). It is quite possible to engineer findings that are significantly better than placebo for agents that should not routinely be used or findings that are not significantly different from placebo in the case of undoubted adverse effects.

An alternative might be to report benefits in terms of odds ratios and confidence intervals to quantify the magnitude of an effect, instead of reporting benefit in terms of the dichotomous classification of treatments that either work or do not work, which significance tests create. While such reporting is currently practiced to some extent, its use might be enhanced.

Given the growing recognition that selective reporting of trials represents a new form of anecdotalism, and the agreement that all clinical trials should be registered, it should be possible in the future to have the confidence interval for an individual trial presented alongside a revised odds ratio and confidence interval for all trials carried out for that agent in that condition. Where trials are registered but the results are unavailable, such results could be incorporated by assigning an odds ratio of 1.0, with an appropriate confidence interval.

Recent FDA reviews of all antidepressant studies show that the confidence interval for all depression trials in children and adolescents straddles 1.0. The odds ratio for a benefit over placebo in 18–25 year olds is 1.54 (95% C.I., 1.34, 1.76), for 25–64 year olds is 1.84 (95% C.I., 1.77, 1.93) and for 65 and over is 1.39 (95% C.I., 1.24, 1.57) (Stone & Jones, 2006). This data drawn from close to 100,000 subjects means that many trials will have had an odds ratio of 1.5 with a confidence interval that included 1.0. This result would have indicated, not that the findings are not significant and should have been disregarded, but rather that the treatment has benefits and that further scientific input was needed to specify the characteristics of responders and non-responders. Findings presented in this way would also have offered scientific support for a presentation of case reports that, with appropriate controls such as challenge and dechallenge, might have made it clear that new drugs, even when effective for some might trigger clinical deterioration and an outcome like suicide in others.

The Midas Touch

As health advances to the fore of our existential concerns and healthcare becomes an arena of increasing competition, managers, clinicians, patients and other interested parties are faced with increasing complexity. Trumpeted as providing gold standard evidence, the lure of RCTs as a solution to these complexities increases, and everything associated with RCTs seems to take on a similar validity. This includes the abstractions from clinical practice we call rating scales.

Rating scales are increasingly being imported into clinical practice, based on the argument that they will reduce variability in the clinical encounter and make that encounter more scientific. Healthcare practitioners are encouraged to administer depression or other behavioral rating scales when seeing patients. Thus guidelines such as the British National Institute for Clinical Excellence (NICE) guidelines on antenatal care advocate using the Hospital Anxiety and Depression scale for all pregnant women (NICE, 2007). As a result pharmaceutical companies now run symposia at major professional meetings aimed solely at introducing clinicians to rating scales.

Rating scale mongering has succeeded disease mongering as the promotional instrument de jour. For example, at the 2007 American Psychiatric Association meeting Pfizer supported the symposium 'From Clinical Skills to Clinical Scales: Practical Tools in the Management of Patients with Schizophrenia.' The practical tools discussed were rating scales, the use of which would draw attention to how the company's drug was superior to others in the field.

The hazards of taking measurement technologies like these out of the clinical trial context are rarely acknowledged. In the first place, a majority of rating scales within the behavioral domain are simply checklists and thus are information poor. The main advantage likely to accrue from their use is to ensure that a number of possibly irrelevant questions are checked off as asked. In time-limited clinical exchanges, if these questions are asked other possibly more important questions are likely to be sacrificed. As Porter (1996) notes measurement often offers solutions across a range of contested situations, but such solutions typically sacrifice profundity for the sake of superficial agreement. When applied to healthcare this dynamic means that the clinical perspective risks being captured by those whose interests are served by the measurement technology.

Second, while rating scales do indeed generate data, exclusive reliance on such data leads to an informational reductionism, which may in turn do more to dehumanize clinical exchanges than the frequently criticized biological reductionism. If specific measurements lead clinicians to overlook dimensions of an individual's functioning or situation that are not open to measurement or which are simply not being measured, we risk being pseudoscientific.

Third, the abstraction, or informational reductionism, of rating scales has a double-edged potential. Having figures for weight can allow us to plot norms for healthy weight, and the feedback from such figures can offer potent feedback in a weight-reduction programme. However, these figures can also seduce both patient and clinician. In the absence of figures from other areas of a person's life, against which the figures for weight can be put in context, there is a risk for the patient that the figures for weight will come to dominate their concerns, establishing a neurosis. The risk for the clinician is that she will also treat the figures rather than the person, although we typically do not pathologize a clinician's figure-centeredness.

An older generation of clinicians would have readily made the case that even in the treatment of eating disorders, weighing scales should rarely if ever be introduced. Whereas in the 1970s and 1980s, the treatment of anorexia differed notably from the treatment of any other condition in psychiatry by virtue of a new centrality accorded to measurement technologies, today this management style is rapidly becoming the norm. Indeed, many clinicians might be alarmed at the prospect of

encountering a patient without a battery of such technologies. There is a good case for getting back to seeing patients without such technologies, even weighing scales, but at present this would involve an all but impossible return to professional discretion.

Evidence-Based Tramlines & Fault-lines

Given the current premium on RCT evidence, it is logical for healthcare organizations to ensure that clinicians adhere to guidelines compiled from a synthesis of available evidence. As a result clinicians worldwide are increasingly faced with managers who enquire as to their compliance with guidelines. It is not clear what the consequences of a failure to comply might be. Medico-legal opinion, while stating that clinicians do not always need to adhere to guidelines, does suggest that any deviation from guidelines needs to be justified (Colbrook, 2005). Because these things are not clear, guidelines risk becoming tramlines within which clinical practice gets constrained, even though in most instances they are not meant to be prescriptive.

By 2003, a series of bodies issuing guidelines, such as the Texas Medication Algorithm Project (TMAP) (Healy, 2006d), and the British Association for Psychopharmacology (Healy & Nutt, 1997), had endorsed the use of SSRIs for treating childhood depression. The same year, NICE was poised to issue guidelines on childhood depression, but the crisis surrounding the pediatric use of SSRIs made it clear that serious flaws marked the literature on which NICE had initially depended, and which other guidelines had appealed to. Not only did a majority of trials in this domain remain unpublished, but also those trials that were published, overemphasized the benefits and concealed the hazards of treatment. The differences between what is now known about the data for this treatment area and what was claimed in the then-published literature, is the greatest known divide of this sort in medicine (Healy, 2006d).

There is no reason to believe that the practices that gave rise to this divide have been confined to the treatment of pediatric depression. Indeed, these practices, which include the control of clinical trials by Clinical Research Organizations (CROs), and the ghostwriting of trial results, are endemic to psychiatry and to a great deal of medicine. Recent treatment controversies reveal that hazards are now systematically downplayed in medical publications, while benefits are consistently oversold. However, these published articles, rather than the raw data from RCTs, provide the only material on which experts can base their guidelines.

In the case of pediatric depression, when all clinical data became available, NICE went on to issue guidelines that cautioned against the use of antidepressants in this age group. However, faced with uncertainties about

what was dependable data and what was not, at one point NICE had considered giving all pharmaceutical company trials a lower ranking in the hierarchy of evidence used to elaborate guidelines (Healy, 2006d). Such a downgrading may never have been politically feasible and has not in fact occurred, but without a mechanism to take into account the distortions outlined earlier, the guideline process risks 'capture.'

Guideline Capture

The phenomenon of regulatory agency capture occurs when experts linked to companies sit on regulatory panels assessing the efficacy and safety of drugs, and when regulators depend on company summaries of clinical trials (Abraham, 2002; Abraham & Reed, 2002). Bodies that issue guidelines are perhaps even more vulnerable to capture of this sort than regulatory agencies, because these bodies have no access to raw trial data and cannot access unpublished trials. For companies, receiving a prestigious guideline to endorse a treatment option offers the most effective marketing possible.

Against this background, consider the recent NICE Guideline on Bipolar Disorder (NICE, 2006). This particular guideline has been chosen to illustrate the process of guideline capture, because NICE is widely regarded as being independent of the pharmaceutical industry. The bipolar guideline contains a number of sensible suggestions that should be part of standard clinical practice, such as monitoring the physical health of patients with bipolar disorder. Aside from such suggestions, one of the agents NICE recommends for bipolar disorder is olanzapine, whose makers have at the time of writing settled legal action for over $1.2 billion for treatment linked physical disorders, with further actions pending (Berenson, 2007).

When it comes to evidence based suggestions, this guideline appears to have all the problems that the NICE guideline for pediatric depression would likely have included, had the issues with the validity of the literature not surfaced. Among others, the guideline makes the following six recommendations.

First, it emphasizes the use of recent on-patent antipsychotics, such as olanzapine, risperidone and quetiapine, and does not mention the older antipsychotics that have been the mainstay of the management of manic-depressive illness since 1952. The reason these older agents are not endorsed is instructive. Classic bipolar disorder leading to hospitalization is relatively infrequent, and when present is typically so severe as to make recruiting patients to an appropriate clinical trial very difficult. Most early papers on chlorpromazine concerned its utility for manic and confusional states, pointing at the same time to its relative inefficacy for schizophrenia

(Delay & Deniker, 1952; Delay et al., 1955). This lack of RCT data for older agents for mania opened up an opportunity for companies to seek an indication for newer, probably no more effective, and potentially more hazardous agents in the management of this illness. To achieve this end, companies have recruited patients with conditions of lesser severity and perhaps less certain diagnoses to short-term trials using crude outcome measures that may reflect little more than the effects of sedation rather than convincing evidence of efficacy. Generating a treatment effect in trials by this means, however, still enables companies to gain a license for the treatment of the condition. As a result the only agents supported by RCT evidence for the treatment of mania or bipolar disorder are newer antipsychotics or anticonvulsants, and these are endorsed by NICE over older possibly better treatments.

There are further complexities. The guideline suggests using risperidone for acute mania. The key trial underpinning this recommendation was conducted in India (Khanna, Vieta, Lyons, Eerdekens, & Kramer, 2005). The correspondence in the *British Journal of Psychiatry* on the ethics and validity of this study (Srinivasan, Pai, Bhan, Jesani, & Thomas, 2006) are more extensive than for any other study the journal has published. Issues of validity aside, this study is a good illustration of a set of processes that center on the control of clinical trials by CROs, which as of 2000 ran more than two-thirds of the clinical trials undertaken by industry.

Privatized research of the sort run by CROs is profoundly different to previous clinical research. These organizations have transformed human subjects research, restructured controls of disclosure and confidentiality, and made RCT data more clearly proprietary than it was when a federation of academic centers conducted trials. CROs provide a privatized IRB system that grants ethical approval to company studies, when university centers might not (Lemmens & Freedman, 2000), and they have made it possible to move trials on drugs for North American and European markets into Asia or Africa, in a way that university departments could not have done (Petryna, 2006). Whether this move has been prompted by concerns to avoid regulatory oversight, or cost considerations is less clear. Even in trials conducted in western settings, it is now clear that CRO run psychotropic trials have included bogus patients (Healy, 2004). Company trials carried out in Asia or Africa seem even more likely to be written up by separate agencies, thus producing all of the problems NICE faced in formulating guidelines for pediatric depression.

In this newly globalizing world of clinical trials, everyone faces a future in which the bulk of the evidence that dictates the local practice of psychiatry will come from settings that are very different from those in which the treatment is given. There are likely to be many consequences for

clinical practice, not least from the fact that different population groups have markedly different responses in terms of both efficacy and side effects to psychotropic agents.

Second, in the case of the prophylactic management of bipolar disorder, NICE recommends the use of one agent, olanzapine, on the basis of data from one trial (Tohen et al., 2006). The pattern of deterioration in subjects in this trial, randomized from olanzapine to placebo, where there is abrupt deterioration after olanzapine is halted, can be interpreted as stemming from a drug induced physical dependence and a withdrawal syndrome rather than a treatment benefit (Ghaemi, 2005; Healy, 2006c).

Third, NICE recommends stopping treatment with antidepressants in favor of 'mood stabilizers' after an acute depressive episode has resolved, stating that there is no evidence that continuing antidepressant treatment reduces relapse rates. This recommendation is unsupported by any evidence and the idea of not giving antidepressant drugs to patients who are depressed is clearly appealing to the marketing departments of companies pushing 'mood stabilizers.'

Fourth, NICE recommends using valproate for prophylaxis, even though this agent has not received a license for this purpose, because there is no supporting evidence for such a claim. The reason for valproate's inclusion by NICE may lie in little more than a set of semantics. Abbott Laboratories christened valproate a mood stabilizer when they launched it in 1995 (Healy, 2006c). This term has no precise clinical or neuroscientific meaning, because of which its use does not risk being illegal. However, the term generates expectations of a prophylactic effect. It is legal for Abbott to claim valproate is a mood stabilizer whereas it would be illegal to claim that it is prophylactic. Of course, there is no need for Abbott to break the law, if a prestigious guideline recommends the use of the drug for this purpose.

Fifth, NICE recommends a series of treatment combinations for patients with frequent relapses or ongoing functional impairment. These combined treatment regimens are not supported by RCT data. The distinguishing feature of the recommendations is that they involve agents that have been more recently been placed under patent. Advocating such options, while failing to mention combinations involving older agents supported by decades of clinical experience, appears to endorse a set of current fashions rather than treatments that have been proven to advance clinical care. Treatment combinations bring out another ambiguity in the current evidence base. Having treatment resistant patients on four or five drugs that 'work' might seem a good option. However it is important to remember that the trials of many drugs primarily show only that it is not possible to say that the drugs do nothing – rather than providing evidence that they 'work.' In this case, while taking a risk with one drug may seem

reasonable, combining four or five such drugs seems like a recipe for unforeseen consequences.

Sixth, in a section on children and adolescents, the guideline considers the possibility of treating bipolar disorders in childhood. There is no mention of the fact that hitherto unanimous clinical opinion, outside North America, has held that bipolar disorders do not start in childhood (Healy & LeNoury, 2007). This outcome has resulted because NICE must necessarily consider clinical trial data rather than other sources of evidence and a series of trials of newer antipsychotics have recently been undertaken in preschoolers and preteens designated as bipolar (Healy & LeNoury, 2007). By considering the treatment for bipolar disorders in childhood, NICE envisages children being given a set of drugs with potent metabolic effects without any evidence for benefits in the long-term. The power of guideline capture can perhaps best be seen in this instance, because a company does not need to seek an indication for treatment in children if influential guidelines tacitly endorse such treatment. This point needs to be read against a background of vigorous efforts taking place in recent years in the USA to convert childhood difficulties into diseases like bipolar disorder to be managed by pharmacotherapeutic means (Harris, 2005; Healy & LeNoury, 2007).

Finally, the guideline does not include any recommendation to monitor treated patients for signs of suicidality, even though current clinical trial data for the drugs it otherwise recommends have been shown these double the risk of suicidal acts compared to placebo (Healy, 2006d; Storosum et al., 2005).

Within psychiatry, however misleadingly certain academic papers may be written, with the possible exception of clozapine for treatment-resistant schizophrenia, no body of studies allows claims for the comparative superiority of one pharmacotherapeutic agent over another. The clearest evidence for this lack of superiority lies in the fact that the regulatory authorities have not permitted any company to make claims for comparative efficacy. The studies on which claims are made are all placebo-controlled trials, and the limited superiority of these active agents compared to placebo should make it clear that no treatment options currently reach the evidential threshold that would mandate their use in preference to other available agents.

In the absence of compelling evidence, the erection of guidelines that advocate one set of agents over another, however well meaning, leaves guideline makers open to being captured. Through a combination of apparently novel indications and publication strategies, companies can make diseases fashionable, engineer the appearances of comparative efficacy and enlist academic advocates for particular treatment options. It now appears that by these means they can also capture guidelines.

Deteriorating Outcomes

A series of recent studies have demonstrated that clinicians fail to adhere to guidelines or that there is, at present, no evidence that outcomes improve with adherence to guidelines (Croudace et al., 2003; Tyrer, King, & Fluxman, 2003).

There are darker aspects to the current situation. Debate about the topic has made it clear that a large number of clinicians are worried about the coercive aspects of guidelines. Cynics may think that clinicians can be expected to worry if their autonomy is being curtailed, however for reasons of even more compelling self-interest few clinicians are likely to want to prescribe treatments that have been demonstrated to be ineffective, or to fail to prescribe treatments that are clearly better than other options. An element of coercion emerges if we consider primary care, where in many settings reimbursement is increasingly tied to guideline adherence. This interpretation of guidelines as coercive becomes even more compelling if one considers that evidence is currently framed within settings in which pharmaceutical companies set up patient groups that lobby for new treatments even when there is no evidence suggesting that they are any better than older treatments.

At a Conference on the Evaluation of Psychotropic Drugs convened in 1956, Ed Evarts from the National Institute of Mental Health (NIMH) put it to his colleagues that but for an accident of history they could now be discussing the use of the new tranquilizing agents for the treatment of *dementia paralytica* rather than *dementia praecox* (Evarts, 1959). None of the rating scales, clinical trial methods or animal models being proposed to move the field forward would have helped researchers to work out that penicillin rather than chlorpromazine or psychotherapy was the right answer to the problem. He predicted that the proposed scaffolding of clinical trials, although eminently sensible, would create an academic and industrial complex inimical to progress in therapeutics.

Fifty years later, in North Wales, compulsory detentions into mental illness units have risen three-fold, admissions for serious mental illness have risen seven-fold, admissions overall have risen fifteen-fold (Healy et al., 2001), suicide rates in schizophrenia are twenty-fold higher than they were previously (Healy et al., 2006) and general mortality for serious mental illness has risen substantially (Harris, 2005). It is unlikely that these changes are local findings: studies of mortality in patients on antipsychotics have indicated that mortality seems correlated with the number of psychotropic drugs given (Joukamaa et al., 2006). In the US, there is an increasing divergence between the life expectancy of patients with serious mental illness and that of the rest of the population (Colton & Manderscheid, 2006). Such findings have been replicated for other countries (Ösby, Correia, Brandt, Ekbom, & Sparén, 2000).

While changing social expectations and other social factors cannot but play some part in these outcomes, this profile is inconceivable against the background of current rhetoric that endorses the practice of evidence based medicine with the latest and most effective treatments. What we are seeing now is not what happens when treatments work; it is not what happened to the *dementia paralytica* services after the discovery of penicillin.

Ways Forward?

After such an unremittingly bleak analysis, it is necessary to point out that many of the sources of difficulty outlined above also provide opportunities for progress. First, clinical trials can answer questions not amenable to individual judgment. Current confusions may stem in part from the harnessing of these methods to solve regulatory difficulties. Few psychotropic trials are designed to inform clinical practice, but recent independent trials of antipsychotics have shown that trials can still play this role (Jones et al., 2006; Lieberman et al., 2005).

Second, many of the problems of guidelines outlined above stem from efforts to endorse particular practices on the basis of limited data, and an assumption that experts can simply compile the outcomes of studies without doing further research. These problems do not apply to guidelines based on studies that point to the inefficacy of treatment options, perhaps because this is what RCTs were designed to do. For example, a series of trials has uniformly indicated that debriefing is not at present an appropriate treatment for post-traumatic stress disorder, a finding which has lead to NICE guideline recommendations against this treatment option (Bisson, Jenkins, Alexander, & Bannister, 1997; NICE, 2005; Raphael, Meldrum, & McFarlane, 1995). In other words, it is possible to create evidence-based guidelines that will be immune from capture by interested parties.

Even the use of rating scales can be helpful if such technologies are embedded in a framework of clinical discretion. As Porter (1996) has been at pains to stress in his work on these issues, the 'hardest' and most objective of the sciences, such as physics, leave greater scope for individual discretion and indeterminacy than many human sciences leave today. It may be time to accept that embracing the notion that human encounters contain an irreducible variability may be a step toward objectivity rather than a retreat to subjectivity.

References

Abraham, J. (2002). The pharmaceutical industry as political player. *Lancet, 360*(9344), 1498–1502.

Abraham, J., & Reed, T. (2002). Progress, innovation and regulatory science in

drug development: The politics of international standard setting. *Social Studies of Science*, *32*(2), 1–33.

Berenson, A. (2007, January 5). Lilly settles with 18,000 over Zyprexa. *New York Times*.

Bisson, J. L., Jenkins, P. L., Alexander, J., & Bannister, C. (1997). Randomised controlled trial of psychological debriefing for victims of acute burn trauma. *British Journal of Psychiatry*, *171*, 78–81.

Cade, J. F. J. (1949). Lithium salts in the treatment of psychotic excitement. *Medical Journal of Australia*, *2*(10), 349–353.

Colbrook, P. (2005, April 9). Can you ignore guidelines? *British Medical Journal Careers*, 143–144.

Colton, C. W., & Manderscheid, R. W. (2006). Congruencies in increased mortality rates, years of potential life lost, and causes of death among public mental health clients in eight states. *Preventing Chronic Disease*, *3*(2). Retrieved from http://www.cdc.gov/pcd/issues/2006/apr/05_0180.htm.

Croudace, T., Evans, J., Harrison, G., Sharp, D. J., Wilkinson, E., McCann, G., Spence, M., Crilly, C., & Brindle, L. (2003). Impact of the ICD-10 primary health care (PHC) diagnostic and management guidelines for mental disorders on detection and outcome in primary care: Cluster randomized controlled trial. *British Journal of Psychiatry*, *182*, 20–30.

Davies, D. L., & Shepherd, M. (1955). Reserpine in the treatment of anxious and depressed patients. *Lancet*, *269*(6881), 117–121.

Delay, J., & Deniker, P. (1952). 38 cas de psychoses traitées par la cure prolongée et continue de 4560 RP. *Compte Rendu du Congrès Aliénistes et Neurologistes de Langue Francaise*, *50*, 497–502.

Delay, J., Deniker, P., & Ropert, R. (1955). Étude de 300 dossiers de maladies psychotiques traits par la chlorpromazine en service fermé depuis 1952. *L'Encéphale*, 528–535.

Evarts, E. (1959). A discussion of the relevance of effects of drugs on animal behavior to the possible effects of drugs on psychopathological processes in man. In J. O. Cole & R. W. Gerard (Eds.), *Psychopharmacology: Problems in evaluation* (pp. 284–306). Washington, DC: National Academy of Sciences/ National Research Council Publication 583.

Fish, F. (1964). The influence of the tranquilizers on the Leonhard schizophrenic syndromes. *L'Encéphale*, *53*, 245–249.

Ghaemi, N. S. (2005). Uses and abuses of evidence based medicine in psychiatry. In Evidence based psychiatry: What it is and what it is not. American Psychiatric Association meeting, Atlanta Symposium 37A, May. Retrieved from http://www.psych.org/edu/other_res/lib_archives/archives/meetings/2005saps.cfm; http://www.mobiltape.com.

Harris, J. (2005). The increased diagnosis of juvenile 'bipolar disorder,' what are we treating? *Psychiatric Services 56*, 529–531.

Healy, D. (1997). *The antidepressant era*. Cambridge, MA: Harvard University Press.

Healy, D. (2004). *Let them eat prozac*. New York: New York University Press.

Healy, D. (2006a). The antidepressant tale: Figures signifying nothing? *Advances in Psychiatric Treatment, 12*, 320–328.

Healy, D. (2006b). Did regulators fail over selective serotonin reuptake inhibitors. *British Medical Journal*, *333*, 92–95.

Healy, D. (2006c). The latest mania. Selling bipolar disorder. *PloS Medicine, 3*(4). Retrieved from http://dx.doi.org/10.1371/journal.pmed.0030185.

Healy, D. (2006d). Manufacturing consensus. *Culture, Medicine and Psychiatry, 30*(2), 135–156. DOI: 10.1007/s11013–006–9013–3.

Healy, D., & Cattell, D. (2003). The Interface between authorship, industry and science in the domain of therapeutics. *British Journal of Psychiatry, 182*, 22–27.

Healy, D., Harris, M., Tranter, R., Gutting, P., Austin, R., Jones-Edwards, G., & Roberts, A. P. (2006). Lifetime suicide rates in treated schizophrenia: 1875–1924 and 1994–1998 cohorts compared. *British Journal of Psychiatry*, *188*, 223–228.

Healy, D., & Le Noury, J. (2007). Pediatric bipolar disorder. An object of study in the creation of an illness. *The International Journal of Risk & Safety in Medicine*, *19*, 209–221.

Healy, D., & Nutt, D. (1997). British Association for Psychopharmacology consensus on statement on childhood and learning disabilities psychopharmacology. *Journal of Psychopharmacology*, *11*, 291–294.

Healy, D., & Savage, M. (1998). Reserpine exhumed. *British Journal of Psychiatry*, *172*, 376–378.

Healy, D., Savage, M., Michael, P., Harris, M., Hirst, D., Carter, M., Cattell, D., McMonagle, T., Sohler, N., & Susser, E. (2001). Psychiatric bed utilisation: 1896 and 1996 compared. *Psychological Medicine*, *31*(5), 779–790.

Jones, P. B., Barnes, T. R., Davies, L., Dunn, G., Lloyd, H., Hayhurst, K. P., Murray, R. M., Marwick, A., & Lewis, S. W. (2006). Randomized controlled trial of the effect on quality of life of second vs first generation antipsychotic drugs in schizophrenia. *Archives of General Psychiatry*, *63*, 1079–1087.

Joukamaa, M. J., Heliovaara, M., Knekt, P., Aromaa, A., Raitasalo, R., & Lehtinen, V. (2006). Schizophrenia, neuroleptic medication and mortality. *British Journal of Psychiatry*, *188*, 122–127.

Khanna, S., Vieta, E., Lyons, B., Eerdekens, M., & Kramer, M. (2005). Risperidone in the treatment of acute mania: Double-blind, placebo-controlled study. *British Journal of Psychiatry*, *187*, 229–234.

Kirsch, I., Moore, T. J., Scoboria, A., & Nicholls, S. S. (2002). The Emperor's new drugs: An analysis of antidepressant medication data submitted to the U.S. Food and Drug Administration. Prevention and Treatment. Retrieved from http://www.journals.apa.org/prevention/volume5/pre0050023a.html.

Kirsch, I., & Sapirstein, G. (1998). Listening to Prozac but hearing placebo: A meta-analysis of antidepressant medication. *Prevention and Treatment*, 1. Retrieved from http://www.journals.apa.org/prevention/volume1/pre0010002a.html.

Kuhn, R. (1958). The treatment of depressive states with G22355 (imipramine hydrochloride). *American Journal of Psychiatry*, *115*(5), 455–464.

Lemmens, T., & Freedman, B. (2000). Ethics review for sale? Conflict of interest and commercial research review boards. *The Milbank Quarterly*, *78*(4), 547–584.

Lieberman, J. A., Stroup, T. S., McEvoy, J. P., Swartz, M. S., Rosenheck, R. A., Perkins, D. O., Keefe, R. S. E., Davis, S. M., Davis, C. E., Lebowitz, B. D., Severe, J., & Hsiao, J. K. (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England Journal of Medicine*, *353*, 1209–1223.

Marshall, B. (2002). *Helicobacter pioneers. Firsthand accounts from the scientists who discovered helicobacters, 1892–1982*. Victoria, Australia: Blackwell Science.

National Institute for Clinical Excellence. (NICE). (2005). Post-traumatic stress disorder. *Clinical Guideline, 26*. Retrieved from http://www.nice.org.uk.

National Institute for Health & Clinical Excellence. (NICE). (2006). Bipolar disorder. *Clinical Guideline, 38*. Retrieved from http://www.nice.org.uk.

National Institute for Health & Clinical Excellence. (NICE). (2007). Antenatal and postnatal mental health. *Clinical Guideline*, *45*. Retrieved from http://www.nice.org.uk.

Ösby, U., Correia, N., Brandt, L., Ekbom, A., & Sparén, P. (2000). Time trends in schizophrenia mortality in Stockholm County, Sweden: A cohort study. *British Medical Journal, 321*, 483–484.

Petryna, A. (2006). Globalizing human subjects' research. In A. Petryna, A. Lakoff, & A. Kleinman (Eds.), *Global pharmaceuticals: Ethics, markets, practices* (pp. 33–60). Durham, NC: Duke University Press.

Porter, T. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press.

Raphael, B., Meldrum, L., & McFarlane, A. C. (1995). Does debriefing after psychological trauma work? Time for randomised controlled trials. *British Medical Journal*, *310*, 1479–1480.

Robertson, M. M., & Trimble, M. (1982). Major tranquilizers used as anti-depressants. *Journal of Affective Disorders*, *4*, 173–193.

Shepherd, M. (1998). Psychopharmacology: Specific and non-specific. In D. Healy (Ed.), *The psychopharmacologists, Volume 2* (pp. 237–257). London: Arnold.

Smirk, F. H., & McQueen, E. G. (1955). Comparison of rescinamine and reserpine as hypotensive agents. *Lancet*, 115–116.

Smith, G. C., & Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge. *British Medical Journal*, *327*, 1459–1461.

Srinivasan, S., Pai, S. A., Bhan, A., Jesani, A., & Thomas, G. (2006). Trial of risperidone in India – concerns. *British Journal of Psychiatry*, *188*, 489–492.

Stone, M., & Jones, L. (2006). Clinical review: Relationship between anti-depressant drugs and adult suicidality. Retrieved from http://www.fda.gov/ohrms/dockets/ac/06/briefing/2006–4272b1-index.htm, p 31.

Storosum, J. G., Wohlfarth, T., Gispen-de Wied, C. C., Linszen, D. H., Gersons, B. P. R., van Zwieten, B. J., & van den Brink, W. (2005). Suicide-risk in placebo controlled trials of treatment for acute manic episode and prevention of manic-depressive episode. *American Journal of Psychiatry*, *162*, 799–802.

Teicher, M. H., Glod, C., & Cole, J. O. (1990). Emergence of intense suicidal pre-occupation during fluoxetine treatment. *American Journal of Psychiatry*, *147*, 207–210.

Tohen, M., Calabrese, J. R., Sachs, G. S., Banov, M. D., Detke, H. C., Risser, R.,

Baker, R. W., Chou, J. C.-Y., & Bowden, C. L. (2006). Randomized, placebo-controlled trial of olanzapine as maintenance therapy in patients with bipolar I disorder responding to acute treatment with olanzapine. *American Journal of Psychiatry*, *163*, 247–256.

Tyrer, P., King, M., & Fluxman, J. (2003). Treatment of common mental disorders in general practice: Are current guidelines useless? *British Journal of Psychiatry*, *183*, 78.

Wallace, D. C. (1955). Treatment of hypertension: Hypotensive drugs and mental changes. *Lancet, 269*(6881): 116–117.

David Healy, MD, FRCPsych, is a Professor of Psychiatry at Cardiff University. He studied medicine in University College Dublin, Ireland, and at Cambridge University. He is a former Secretary of the British Association for Psychopharmacology, and author of over 150 peer reviewed articles, 200 other pieces and 15 books, including *The Antidepressant Era*, and *The Creation of Psychopharmacology* from Harvard University Press, *The Psychopharmacologists* Volumes 1–3, *Let Them Eat Prozac* from New York University Press, and *Mania* from John Hopkins University Press. His main areas of research are clinical trials in psychopharmacology, and the history of psychopharmacology. *Address:* Cardiff University, Department of Psychological Medicine, Hergest Unit, Ysbyty Gwynedd, Bangor, Wales LL57 2PW, UK. [E-mail: healy_hergest@compuserve.com]