



**The Right Patients for the Drug:  
Rating Scales and the Quantification of Experience**

Andrew Lakoff  
UC San Diego

ASA Montreal, August 2006  
DRAFT: Please do not cite or quote without permission

## Introduction

The category of “depression” simultaneously names an individual condition of suffering, a collective mode of identification, and an expert classification. It is an example of what Ian Hacking (2002) calls a “human kind” – that is, a way of classifying persons that dynamically interacts with the experience of those who are being classified. Human kinds are notoriously difficult to stabilize as scientific objects. Nonetheless, the legitimacy of depression as a biomedical illness – and its power as a form of self-identity – hinges on the ability to clearly delineate the boundary between normality and pathology.

In this talk I will describe a device that seeks to stabilize the category by making it measurable: the symptom rating scale. The talk is based on ethnographic work among drug developers, whose task is to demonstrate the efficacy of an experimental compound on a group of depressed subjects. My interest is in the challenges they face in quantifying the subjective effects produced by medication and what these challenges imply about “human kinds.” As I will show, depression and its transformations are notably recalcitrant to technical

measurement – which forces drug developers to find other, non-quantitative means to classify and measure experimental subjects.

## Rating Scales

Randomized, controlled trials involve testing an experimental compound against a placebo in a population of subjects with a common illness. In the case of depression, the key device for assembling such populations and measuring their response is the symptom rating scale. Here is the “gold standard” for antidepressant trials, the Hamilton Depression Rating Scale (HAM-D) – developed in 1960. The scale has seventeen items, which are scored in order of intensity: these include symptoms such as “depressed mood,” “feelings of guilt,” “agitation,” and “anxiety.” A score of seventeen or higher (out of fifty four) indicates mild depression, and twenty-five or higher points to moderate or severe depression.

The goal of using such scales is to eliminate reliance on the subjective judgment of experts, and to provide a stable reference point for testing whether a given compound in fact improves the patient’s condition; to turn amorphous,

heterogeneous experience into a statistically calculable problem. Thus, like other commensuration techniques, the rating scale “standardizes relations between disparate things and reduces the relevance of context”, as Espeland and Stevens (1998) put it. However, such standardization requires agreement as to the salient characteristics of the condition to be treated. This is a particular struggle in the case of illnesses such as depression, which do not lend themselves to physiological measurement (Lakoff 2005).

## Signal Detection

The use of rating scales points to a general assumption of the regulatory process: that the effect of a medication is uncertain until it has been demonstrated on a population with a shared biomedical condition. What is interesting is that antidepressant developers don’t actually hold this model of the relation between illness and intervention. Consider their use of the term “signal detection” to describe the goal of a trial: here the drug is already presumed to have efficacy—that is, a signal to transmit—and the question is how to pick it up. If measuring devices, such as rating

scales, record the signal, we can see that patients' role is to transmit it—they are the drug's medium.

So for antidepressant developers it is the drug, rather than the depressed patient, that serves as a stable reference point. In fact, they are quite skeptical about the capacity of rating scales to produce a consistent patient population for testing. From painful experience, they have learned that patients admitted under these criteria vary tremendously in their response to drugs and to placebos—and also, that the scales are applied inconsistently by raters at trial sites. Attempts to standardize the application of the scales—such as video training sessions and site audits—seem not to have improved trial success rates.

### **Uncertain Placebo**

From the vantage of drug developers, when trials fail, which happens disturbingly often, it is not that the drug doesn't work but that noise has crept into the signal detection process. And the most pernicious and obstinate source of noise is the placebo effect. The placebo effect is unpredictable and seemingly unmanageable, and costs drug companies hundreds of millions of dollars in failed

trials and delayed or shelved compounds. Since the placebo response rate in depression is typically at least 30%, and the response to antidepressants is often not much higher—perhaps 40%—it can even seem to impugn the efficacy of established drugs, used as active comparators in trials of novel compounds. And worse, it seems that the placebo response rate has actually been increasing in recent years, for unknown reasons. Drug developers have tried many things to reduce placebo response without at the same time reducing treatment response, but have been frustrated by its intransigence.

In a discussion with me, a biostatistician at Merck expressed his frustration at the challenges of depression trials. On the one hand, he said, you want to standardize raters' behavior as much as possible in order to glean consistent data—but then you might “dampen the signal” by failing to note clinical signs not measured by the rating scales. Yet if you focused too much on close clinical observation, you might actually create more placebo response because of the attachment that would then form between the rater and the subject. “There are so many different problems in this area,” he said, “it's like taking a balloon and trying to

squeeze it in a certain place—the air just gets pushed elsewhere.”

While advocates of alternative medicine have begun to see the placebo effect as a possible source of new forms of therapy (Harrington 2002, Kaptchuk, 1998) for drug developers it is an impediment to proving efficacy and bringing new drugs to market. For them, understanding the operations of the placebo effect – or at least managing it – is a question of corporate necessity. As a part of these efforts, they have located a number of possible candidates for the locus of placebo response, including the patient, the healer, the measuring device, and the illness itself.

An initial distinction they make is between *artifactual* and *real* placebo response. There are at least two kinds of artifactual placebo response. One has to do with the motives of raters at the trial site. If the site is under pressure to rapidly enroll subjects, raters may inflate their scores at the beginning of the trial. If those in the placebo group then show improvement, it may be due to more accurate later measurement. A second potential source of artifactual placebo response is statistical “regression to the mean”—when the subject has a rapidly fluctuating course of illness, and enrolls in

the trial when it is at its worst. Then, when the illness improves on its own over the course of the trial, one again sees what looks like, but is not really a placebo response.

“Real” placebo response can also be attributed to the characteristics of either the trial site or the subjects. One of its possible causes has to do with “investigator behavior.” Here researchers presume a certain understanding of what the placebo response is: it is based on hope, expectation, an attachment to the healer or to the treatment itself. Thus if the site-based investigators—contracted by the drug developers—perform what is termed “covert psychotherapy” or in some other way give those who have been assigned placebo the sense that they are being helped, that can induce unwanted placebo response. Stuart Montgomery, co-inventor of one well known rating scale, argues that such “non-specific supportive contact” can even include the filling out of forms, if these are meant to reassure patients about their participation in the trial. He advises sternly: “Patients who are overly sensitive to reassurance need to be identified and if possible excluded” (Montgomery 1999).

## The Right Patients for the Drug

This advice relates to a more general set of strategies based on the hypothesis that the source of excessive placebo response is the presence of a certain class of patients who are overly susceptible to placebo. Given the heterogeneity of the depressed patient population, antidepressant developers here shift the locus of potentiality in the trial from the drug to the patient. Instead of seeking to test the drug on an established category of patients, they seek to find the right patients for the drug. As the statistician I spoke with complained, “The biggest problem is getting the right patients.” Who are they? “No one knows, but there are a lot of different ideas.” He mentioned some of the possible clues to placebo-susceptibility: duration of illness, family history, age of onset. But “they don’t hang together from one trial to the other,” he said. “Things disappear as you look at them more closely.”

The most salient sub-populations to be delineated before the trial begins are drug responders and placebo responders. Here subjects are not considered depressed patients but rather potential transmitters of drug efficacy. As Montgomery writes, “samples selected for trials

should be able to deliver a predicted response to drug and not to placebo.” Unfortunately, he continues, standardized diagnostic criteria are “not up to the task of distinguishing between clear drug responsive patients and placebo responders.” However, seasoned investigators seem to intuitively be able to make this distinction and exclude the latter from trials. Such non-standardizable knowledge, linked to experience, might then explain why some trial sites do better than others in demonstrating drug efficacy.

Attempts to delineate the characteristics of placebo-responders have a long history, beginning in the years after World War II, when the double-blind, randomized controlled trial was accepted as the means to police fraudulent medications. In the rationale of the controlled trial, the placebo effect was both an epistemological necessity and a practical obstacle to showing true drug efficacy. If one could ascertain before the trial which subjects were likely to respond to placebo and thereby contaminate the results, one could ostensibly eliminate them from the trial beforehand and improve the chances for a successful trial.

In the 1950s, pioneer placebo researcher Louis Lasagna (1954) used personality

questionnaires and Rorschach tests to characterize the typical placebo-reactor in post-operative pain trials. When asked, “What sort of people do you like best?” placebo reactors were more likely to respond, “Oh, I like everyone.” They were more often active churchgoers than non-reactors, and had less formal education, Lasagna and his colleagues noted. As for the Rorschach results, they wrote: “In contrast to the non-reactors the reactors were... more anxious, more self-centered and preoccupied with internal bodily processes, and more emotionally labile... the reactors are in general individuals whose instinctual needs are greater and whose control over the social expression of these needs is less strongly defined and developed than the non-reactors.” In the 1970s, researchers found that placebo reactors scored higher on the “Social Acquiescence Scale” (SAS), based on agreement with proverbs such as “Obedience is the mother of success;” “seeing is believing;” and “One false friend can do more harm than 100 enemies” (McNair 1979).

This line of research linking placebo response to suggestibility lost its momentum in the ensuing years, and more recently, researchers have focused on other factors (Schatzberg and Kraemer 2000).

In the case of depression, they hypothesize that milder severity of illness, a more rapidly fluctuating course, or certain kinds of somatic complaints correlate with placebo response. One explanation for the increasing placebo response rate might then be that less severely ill patients are now being used more often given the shortage of subjects for clinical trials (Montgomery 1999). But efforts to operationalize these criteria bring other problems, such as limiting the potential indication of the approved drug or extending the length and expense of the trial in the search for better qualified patients.

### **The Mousetrap**

Given the difficulty in trying to root out the placebo effect by addressing its underlying causes, antidepressant researchers have turned to a more pragmatic approach that might be called the “mousetrap technique”—after the play within a play Hamlet staged in order to goad his father’s murderer into revealing his guilt. Here experimenters in effect stage the trial before it actually begins, giving all patients placebo for a week, and then eliminating those who respond from

the trial. This is called a “single-blind placebo run-in period,” since the investigators know that all of the subjects are receiving a placebo (Quitkin, et al 1998).

With this approach, it doesn’t matter why subjects respond to placebo, nor does the knowledge or technique of the investigator matter. One simply needs to know which patients have responded in order to eliminate them. However these efforts have also proven disappointing—placebo response rates during these run-in periods tend to be low. Then in the actual trial other subjects continue to respond to the placebo, drowning out the signal of drug efficacy and undermining the trial.

As a result, some trial consultants have called for the abandonment of the run-in period. But researchers at Ely Lilly recently noticed some interesting things from looking at the post-trial data from trials with placebo run-ins (Faries, et al, n.d.). They saw that the treatment response curve tended to be flat until the point of randomization to drug or placebo—at week two or three. Sharp improvement then set in, even in subjects who remained on placebo at that point. So perhaps there was

something about moving from the run-in period to the actual trial that increased placebo response.

The Lilly researchers then designed an experimental study to see whether it was the double-blind of the post-randomization phase that was causing this rise in response. They built in a “double-blind placebo run-in period,” in which neither the site-based evaluators nor the subjects knew about the run-in period. And this time they got a much larger placebo response rate—28%, as opposed to 5% in the single-blind run-in period. By eliminating these placebo responders from the final trial results, they improved evidence of drug efficacy significantly. Unlike the partial simulation of the single blind run-in, this time the experimenters staged a full simulation of the trial before the actual trial, in order to make visible the heretofore hidden population of potential placebo responders.

## Conclusion

What do these practices tell us about symbolic boundaries and the quantification of identity? At one register, they are simply tactics for manipulating clinical trial methodology in the service of bringing drugs to market. But at the same time,

they challenge a central premise of antidepressant trials, and of the biomedical model of depression more generally: that rating scales define a coherent group of subjects with a shared illness (Rosenberg 1992).

The story indicates the intransigence of a certain set of experiences to quantification. Trial designers are frustrated by the limits of rating scales, and thus turn to a performative technique – the generation of “non-specific response” through simulation – in order to re-classify subjects: no longer as depressed, but as “placebo-responders.” At this point, of course, these newly classified subjects re-enter the calculative process – insofar as they are eliminated in order to statistically demonstrate efficacy among the remaining population. Some form of rational calculation can proceed, in other words, even when the symbolic boundaries of depression prove too fluid for reliable quantification.

## References

Espeland, Wendy Nelson and Mitchell Stevens. “Commensuration as a Social Process,” *Annual Review of Sociology* 24, 1998.

Faries, Douglas E., Heiligenstein, John H., Tollefson, Gary D., Potter, William Z. “The Double Blind Variable Placebo Lead-in Period: Results from Two Antidepressant Clinical Trials. *Journal of Clinical Psychopharmacology* Dec: 21:6 (2001).

Harrington, Anne, “‘Seeing’ the placebo effect: Historical Legacies and present opportunities.” In: *The Science of the Placebo: Toward an Interdisciplinary Research Agenda*. 2002.

Hacking, Ian. *Historical Ontology*. Cambridge: Harvard University Press, 2002.

Kaptchuk, Ted J. “Intentional Ignorance: A History of Blind Assessment and Placebo Controls in

Medicine. *Bulletin of the History of Medicine* 72:3 (1998).

Lakoff, Andrew. *Pharmaceutical Reason: Knowledge and Value in Global Psychiatry*. Cambridge: Cambridge University Press, 2005.

Lasagna, Louis, Mosteller, Frederick, von Felsinger, John M., and Beecher, Henry K. "A Study of the Placebo Response." *American Journal of Medicine*, June 1954.

McNair, D.M., Gardos, G. Haskell, D.S., and Fisher, S. "Placebo Response, Placebo Effect, and Two Attributes." *Psychopharmacology* 63 (1979).

Montgomery, S.A. "The failure of placebo-controlled studies." *European Neuropsychopharmacology* 9 (1999), 271-276.

Quitkin FM, McGrath PJ, Stewart JW, Ocepek-Welikson K, Taylor BP, Nunes E, Delivannides D, Agosti V, Donovan SJ, Ross D, Petkova E, Klein DF.

"Placebo run-in period in studies of depressive disorders. Clinical, heuristic and research implications." *British Journal of Psychiatry*. 1998 Sep; 173:242-8.

Quitkin, Frederic. "Placebos, Drug Effects, and Study Design: A Clinician's Guide." *American Journal of Psychiatry* 156(6), June 1999.

Rosenberg, Charles. *Explaining Epidemics, and Other Studies in the History of Medicine*. Cambridge: Cambridge University Press, 1992.

Schatzberg, Alan F. and Kraemer, Helena C. "Review of Placebo Control Groups in Evaluating Efficacy of Treatment of Unipolar Major Depression." *Biological Psychiatry* 47 (2000).

Trivedi MH, Rush H. "Does a placebo run-in or a placebo treatment cell affect the efficacy of antidepressant medications?" *Neuropsychopharmacology* 11:1 (1994).