The Sandstone Effect: How Causal Modeling Can Correct Methodological Failures in Evidence Based Medicine

By: Bennett Holman

Graduate Student in Logic and Philosophy of Science

University of California Irvine

Submitted in consideration of the Charles A. Lave Paper Prize for Creative Modeling in Social Sciences

Abstract

Expectancy effects are commonplace in medical treatment. To insure that a treatment is efficacious over and above an expectancy effect, medical research employs double-blind studies comparing an active treatment to a placebo. Past authors have identified specific cases in which the internal validity of a study was compromised and have proposed methodological corrections. This paper identifies the general conditions that can lead to expectancy effects within randomized clinical trials (dubbed "the sandstone effect") and proposes causal modeling as a statistical correction. The pressing nature of this confound is brought to the fore through a review of the research literature for antidepressant medication, where it is possible that accounting for the sandstone effect could eliminate the observed superiority of treatment over placebo. It is argued that the statistical correction is superior to the methodological correction, since the former achieves the same end and can be implemented without a drastic overhaul of the research paradigm. Indeed, since it can be carried out with the type of data already collected, it should be employed retroactively on studies already completed.

A central concern for evidence based medicine (EBM) is ensuring that expectancy effects are controlled for. A variety of motives underly this concern, but one salient issue is that of ensuring that a procedure offered by the medical community is more than *just* the result of positive expectations. One of the central tenets behind EBM is that the evaluation of medical treatments cannot rely exclusively on the perception of practitioners. Doctors only observe the *total effect* of a treatment--they cannot account for what portion of the witnessed improvement is due to the *specific effect* of the treatment over and above what improvement would have been achieved by a placebo. A long history of practitioners testifying to the efficacy of treatments without specific effects tempers the credence of claims derived from personal experience. This recognition is behind the methodological constraints imposed by the randomized clinical trials (RCTs). The central assumption is that if researchers randomly assign patients into a treatment group and a placebo group, then they can measure the specific effect. Calculating the specific effect becomes a straightforward subtraction of the placebo group from the treatment group.

This paper will argue that the methodological constraints employed in RCTs conjoined with comparisons of group means do not license the inference that the active component of the substance being tested cause the observed superiority to the placebo control. First, I will consider a fictitious example so that the methodological failure can be illustrated clearly. I will specify additional conditions that must be satisfied in order to ensure the integrity of RCTs and identify situations in which the internal validity of RCTs are compromised. With these considerations in mind I will outline how causal modeling can provide a more robust assessment of efficacy of specific effects. Second, I will apply this model to an actual RCT of antidepressants. I will then show that the causal modeling proposed above can yield different conclusions regarding the specific effect of a drug than those reached by traditional analysis. Finally, I will consider the broader importance of this analysis for the practice of evidence based medicine; particularly, for how we conceive of placebos and what it means for a treatment to be effective.

The Crystal Palace

To highlight potential methodological issues in RCTs, consider the following thought experiment. Suppose we lived in a society in which crystal healing was the standard form of treatment, but we were inclined to doubt the efficacy of crystals. Suppose the healer were to bring us a large number of patients and doctors with tales of miraculous cures. The healer notes that in addition to its proven track record, crystal healing is based on the well-established sciences of biology and geology. To suggest that it doesn't work would fly in the face of good sense – how could crystal healing gain such widespread acceptance if it were not effective?

Surely we could bring out exactly the points raised in the introduction to this paper: that these observations were made in situations that were not carefully controlled, that there is a long and storied history of doctors and patients firmly believing in treatments whose only active component is confidence, etc. We claim that although these treatments work better than nothing at all, this does not truly mean that they work *simpliciter*. For a treatment to be deemed efficacious, we require that it must meet our more stringent criteria.

In this case, let us suppose that crystal healers (and thus *ex hypothesi*, the population at large) believe that sandstone is the appropriate treatment for some disorder. To address our critique the crystal healer performs the following experiment. Two hundred ailing patients are gathered and informed that they will be randomly placed into two groups, only one of whom will receive actual crystal healing. In the first group, sandstone is used to treat the patients. In the second group, amethyst (which has no expected effect) is used instead. Suppose the amethyst and the sandstone are enclosed in a device which allows for the appropriate skin contact while preventing the patient and the healer from seeing which is being used. In follow-up evaluations, both groups improve considerably, and although the difference between them is small, it is shown that the sandstone group has significantly better outcomes. Suppose further that this finding is repeatedly confirmed. The healer concludes that there is extensive scientific evidence to support the claims widely accepted claims of efficacy.

In this circumstance, a hard-minded EBM adherent is not yet obligated to accept the amazing healing power of crystals. Suppose I hypothesized that patients were not really kept in the dark about what group they were in. To support this, I produce evidence that patients, doctors, and independent evaluators do significantly better than chance at identifying which group the patients were in. There are two possible explanations for this. It could be the case that subjects can discriminate which group they are in because they believe that sandstone works and sandstone does in fact work. Subjects who recover reason that if they recovered, then they must have been treated. Since they recovered, they must have been in the sandstone group. Subjects who don't recover reason the same way (*mutatis mutandis)* concerning amethyst. Alternatively, it could be the case that after the groups were randomized, amethyst did not mimic the "non-therapeutic" aspects of sandstone effectively and thus subjects could discriminate between the two based on properties that had no relation to the supposed healing properties of sandstone.

In line with the second hypothesis, suppose that when sandstone is rubbed on the body, it causes abrasions, whereas amethyst does not. Further, I show that abrasions and not improvement pred ict which group patients believe themselves to be in. Finally, suppose we find the following: When you control for which group patients believe themselves to be in, there is no added benefit to receiving the sandstone treatment. In light of these facts, I believe we could conclude that sandstone was not an effective treatment despite its apparent superiority in the trial. The effect of sandstone was to cause abrasions. The presence of abrasions signaled to subjects they were receiving the culturally accepted medical practice, which, in turn, resulted in increased symptomatic improvement via an expectancy effect. Let us call this "the sandstone effect."

We can expect the sandstone effect to occur in double blind trials if two conditions are satisfied. The first is that there is a detectable difference between the non-therapeutic aspects of the treatment being tested and the placebo. The second condition is that patients who believe they are receiving treatment do better than patients who believe they are receiving placebo. Thus, two factors will

contribute to the strength of the sandstone effect: the degree to which patients and doctors can discriminate which group they are in and the size of the expectancy effect.

The medical community uses double blind studies to ensure that improvement observed by clinicians is not the result of the placebo effect. I argue that the same motivations should lead researchers to guard against the sandstone effect with approaches that are more sophisticated than current double blind studies. Both the placebo effect and the sandstone effect are expectancy effects and should be accounted for when determining whether a treatment is legitimate. Further, both threaten the advance of medicine as science and practice. The placebo effect can convince doctors that they have an effective treatment when in fact they do not. The sandstone effect has the potential to lead researchers astray into thinking the drug's active ingredients cause specific therapeutic effects and encourages the development and refinement of treatments whose promise may be illusory.

Thus, if in addition to the conditions identified in the previous paragraph, there is evidence that in some cases the sandstone effect could account for the majority or totality of the difference between the drug and placebo group, then it is incumbent on researchers to include more stringent controls. Accordingly, we may be especially concerned when the difference between the treatment and control group is small, when placebo effects constitute a relatively large percentage of the total effect, and/or when the ability of patients to guess which group they are in far exceeds chance rates.

Two Solutions

The methodological solution[1] for this problem is conceptually simple: Use an "active placebo". An active placebo is a substance which mimics the side effect profile of the experimental substance, but is not expected to have any clinical benefit. If the only difference between the experimental group and the placebo group is the presence of side effects, then the incorporation of a active placebo will prevent patients from detecting which group they are in. If the incorporation of an active placebo serves its purpose, the first condition necessary to produce the sandstone effect (that there is a

1 This solution was proposed at least as early as Thompson (1982).

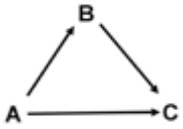
detectable difference between the non-therapeutic aspects of the treatment being tested and the placebo) would not be met.

While this solution is not conceptually complicated, neither is it a perfect solution. First, ethical concerns have been raised concerning the use of active placebos. Institutional review boards are in charge of protecting the interests and safety of research participants. Conceptually, review boards are opposed to withholding potential treatment for sick patients (which is of course what is required in a double blind trial); however, it may be well nigh impossible to pass a project that proposes not only to withhold treatment, but to give a patient a substance with the expressed purpose of causing negative effects.

Perhaps the largest problem is that implementing trials with active placebos would require a large scale change in RCT methodology. Guidelines would have to be established for when use of active placebos is warranted. Standards would have to be set concerning how to establish that two substances have similar enough side effect profiles that one could be used as an active placebo for the other. This does not even consider the likely resistance from the pharmaceutical industry.

What has been unappreciated is that given more recent sophisticated statistical approaches, there is a solution that can be carried out with the type of data already collected. Instead of testing group differences, regression analyses can be used to test a proposed model of symptomatic improvement. Specifically, statistical analysis can be used to ascertain whether one variable acts as a mediator between two other variables (Holmbeck, 1997; MacKinnon, Lockwood, & Hoffman, 2002). In the remainder of the paper, I will make clear the logic behind this method (formalized by Pearl, 2000), and offer a preliminary analysis using the TDCRP data.

First, suppose that we have three variables A, B, and C that are each highly correlated and we wish to rule out the case where A is a common cause of B and C, but B has no independent effect on C. This results in a model structure with three paths as in the figure below.



The arrows represent possible effects in the model. We have evidence of the sandstone effect if four conditions hold: 1) A is a significant predictor of C without regard to B; 2) A is a significant predictor of B; 3) B is a significant predictor of C after controlling for effect of A on C; and 4) A is a significantly poorer predictor of C when B is controlled for.[2] To get a feel for the model, consider three cases: a case where all four conditions are satisfied: the crystal palace thought experiment; a case where one of the conditions presumably fails – cancer treatment; and a case where it is unclear – antidepressants.

To assess crystal healing, let A be the amount of sandstone applied, let B be the patient's belief concerning group assignment, and let C equal the degree of symptomatic improvement. In the supposed situation, sandstone causes an expectancy effect, so there will be a relation between A and C. Further, since patients used abrasions to discriminate between "real crystal healing" and placebo crystals, there will be a relationship between A and B. Further, all of the effect variance is explained by people's confidence that they are receiving "real crystal healing," so there will be an effect of B on C (since crystals do not actually heal, controlling for their contribution to healing will not change the effect of B on C). Finally, the effect of crystals on healing disappears when we control for belief, so A is a significantly worse predictor of C when B is controlled for. Thus, all four conditions are met.

As a second example, let A be whether radiotherapy for cancer is directed at cancerous region, let B be patient's belief concerning group assignment, and let C be the remission rates. The first condition holds because radiotherapy is an effective treatment for cancer. Likewise, since patients may have knowledge of the approximate location of the cancerous region, A will predict B. The third condition will hold if expectancy effects contribute to remission. The fourth condition will presumably

---

2 A case where A is not a significant predictor of C when B is controlled for would indicate that A does not have a direct impact on C.

fail, as it stands to reason that radiotherapy is effective over and above any effect of positive expectation for treatment. In this case, we would have evidence that radiotherapy works when patients' belief about group assignment is controlled for.

The Pharmaceutical Kingdom

Pharmacological treatments for depression arose near the birth of modern psychiatry. In a search for less toxic versions of psychiatric medications, Roland Kuhn tested the imino-dibenzyl analogue of chlorpromazine (Thorazine)[3] for treatment of schizophrenics with disastrous consequences. Though it was a sedative, it caused paradoxically manic effects. Kuhn reasoned that if it caused mania in patients with a normal affect, it might cause those with depressed affect to achieve a state of normalcy. His first trials of Imipramine (Tofranil) were published in 1959 and came into use two years later. Despite its significant side-effect profile, imipramine and other tricyclic antidepressants (TCAs) would remain the gold standard for pharmacological treatments until the advent of Fluoxetine (Prozac) and other serotonin specific reuptake inhibitors (SSRIs).

There are now more than fourteen drugs in different drug classes that have been approved by the FDA. The number of antidepressant RCTs in the pharmacological literature is enormous. To form clinical guidelines the Agency of Healthcare Research (in the U.S. Department of Health and Human Services) commissioned a meta-analysis of trials for antidepressants. Drawing from research published between 1980 and 1998, Williams et. al. (2000) identified 315 trials which met their search criteria (from a total of 1,277) to determine the effect of newer antidepressants, such as Prozac.

A previous systematic review sponsored by the U.S. Department of Health and Human Services (Depression Guideline Panel, 1993) had established the efficacy of first and second generation TCAs for treating depression. Using an empirical Bayes random effects estimator method, Williams et. al. (2000) showed that newer antidepressants (including SSRIs) were equally effective to older

---

3 Trade names of pharmaceuticals will follow their chemical designation in parentheses.

antidepressants. Defining improvement as at least a 50% reduction in symptoms as assessed by a depression symptom rating scale or a much improved score on a global assessment measure, they found that newer antidepressants lead to improvement in 54% of patients compared to 54% of patients using older antidepressants and 32% of those receiving placebo. Relapse rates at 24 weeks after treatment were 10% for active medication and 35% for placebo. They found no significant differences within the class of newer antidepressants or between the new and the older generations. They concluded that primary care physicians should discuss various potential side-effects with their patients to determine which side-effect profile was preferred.

In light of hundreds if not thousands of controlled studies, it seems that antidepressants are well-researched and conclusively shown to be an effective treatment for depression. However, as was discussed the hypothetical case above, RCTs may not fully control for expectancy effects. Recall, the two conditions for the emergence of the sandstone effect are (1) a significant placebo effect and (2) the presence of a detectable non-therapeutic difference between the treatment and the placebo groups. Further, (3) when the difference between the drug and the placebo is small and/or the ability of participants to determine which group they are in is significantly above chance, there is a threat that the apparent efficacy may be accounted for by the sandstone effect. I shall deal with each of these points in turn.

Evidence for a significant placebo effect

The presence of a significant expectancy effect is widely acknowledged though underappreciated. As noted above 32% of patients in the placebo group experienced a significant improvement in depressive symptoms (Williams et al., 2000). There are two plausible alternatives that account for the symptomatic improvement of patients who received placebo. The first is the placebo effect. The second is that improvement represents the natural course of the illness. The latter is not tested for explicitly by RCTs; however, in clinical trials for psychotherapy, patients who remain on waitlists achieve little if any improvement (Kirsh & Sapirstein, 1998).

Critics might note that this is only indirect evidence for the presence of expectancy effects (e.g. Klein, 1998a) and they would be right. However, the widely accepted consensus is that there are expectancy effects in treatments of depression. Given the *prima facie* plausibility and indirect evidence in support of a placebo effect, it seems that the burden of proof is on the critic to show that all or most of the improvement seen in placebo groups can be accounted for by spontaneous recovery. Though this is not ironclad evidence of a placebo effect, the same plausibility that motivates double blind trials in the first place should allow that a placebo effect is present unless we have good evidence to the contrary.

Evidence for a detectable difference between experimental and control groups

The major assumption of double blind trials is that patients do not know to which group they have been assigned. If the only difference between the drug group and the placebo group is the active component of the drug, then the comparison of group means yields the effect size of the active ingredient. Yet, this crucial assumption is assessed in less than 5% of the hundreds of drug trials included in the Snow, *et al.* (2000) meta-analysis.[4] The central assumption, which must be true in order to apply the statistical analysis used in evaluating the efficacy of treatment is rarely assessed.

Some attempts have been made to determine whether blinding is effective and if it is not, why it fails. Fisher and Greenberg (1994) examined assessments of the double blind in trials of psychotropic drugs (viz. not only antidepressants) and found that of 26 trials which assessed whether randomization was effectively preserved, only 3 studies showed that trial patients guessed which group they were in at levels that would be expected by chance alone. Yet the ability of patients to identify which group they are in is not an inherent threat to the integrity of the trial. Rabkin *et al.* (1986; cf. Bystrisky & Whiaker, 1994) suggested two mechanisms for a failure to maintain the integrity of the study: (1) Patients might guess based on whether they were experiencing side-effects; (2) Patients might guess based on whether they had improved. Only the former would be a threat to the internal validity of RCTs, "since this clue

4 How much less than 5% is unspecified.

is the clinical response, it cannot alter *assessment* of clinical response" (p. 76).

Rabkin *et al.* (1986), found that patients could discriminate which group they were in at levels far above chance and that patients in the drug groups[5] were more accurate in their guesses than patients receiving placebo, (between 87% and 100% versus 59% and 60% correct identification). Further they found an interaction between clinical response and guessing in the expected direction (a positive clinical response increased guessing accuracy in the drug group and worsened accuracy in the placebo group). They did not find that side-effects during week six contributed to guessing accuracy; however, they note that when controlling for clinical improvement, patients and doctors still guess far above chance levels. In discussing their results, they them as follows: "Most limiting of all is the absence of weekly side effect assessments, controlled for baseline manifestations, to permit evaluation of timing of onset and patterns." In a more recent consideration (Quitkin *et al.,* 2000), they note that though there may be violation of the double-blind, given that medication is two to three times more effective than placebo, the failure is unlikely to threaten the demonstrated efficacy of antidepressants.

Regardless of how patients determine which group they are in, it is clear that even after accounting for improvement, participants can do so. Thus, it is sufficient to note for now that the second criterion for the sandstone effect obtains. However, if Quitkin *et al.* (2000) are correct concerning the clear superiority of antidepressants, nothing short of an extremely large sandstone effect would impugn the efficacy of antidepressants. It is to the third criterion (small relative difference between drug and placebo) that I now turn.

Reassessing the superiority of antidepressants

It has recently been realized that the case for antidepressants is significantly overstated. For instance, Kirsch and Saperstein (1998) sought to measure the extent of the placebo effect in trials of antidepressants. While they expected to find a placebo effect, they did not expect that they would be

5 This study involved two treatments for depression and one placebo and was replicated yielding four drug groups and two placebo groups.

unable to identify a drug effect. They noted, first, that for both TCAs and SSRIs, the placebo achieved 75% of the drug effect but, secondly, that medications that were not considered antidepressants achieved equal efficacy to TCAs and SSRIs. Assuming the result is not a statistical anomaly, this allows three interpretations: 1) the other medications affected depression indirectly (e.g., by ameliorating anxiety); 2) the other medications are actually antidepressants; 3) the other medications and antidepressants achieve efficacy as the result of a sandstone effect. In other words, the presence or absence of side effects alters patients' belief about whether they are receiving treatment. It is not surprising that these conclusions stirred controversy.

These findings was not refuted by conflicting evidence, but by noting methodological flaws in Kirsch and Saperstein's (1998) analysis. Rehm (1998) noted that the manner by which the study equated the efficacy of antidepressants and other medications was suspect. Klein (1998) excoriated the study, noting several severe methodological flaws. The meta-analysis used studies which were not representative, combined outcome measures (some of which were "insensitive") to calculate effect sizes, averaged the effects of multiple dosages within studies (some of which were sub-clinical) to calculate a single drug effect, failed to include unpublished studies, and used completer samples.[6] As a result of these biases against finding specific drug effects, Klein concluded that a 75% placebo effect was a "grossly flawed underestimate" of the efficacy of antidepressants (pg. 2). A series of responses and rejoinders followed in which Klein continued to identify flaws in Kirsch's reasoning (1998b) and Kirsch (1998a, 1998b) often failed to address Klein's most cogent critiques.

However, Kirsch, *et. al.*, (2002) rectified these flaws by conducting a second meta-analysis that: 1) was representative; 2) made use of a single "sensitive" outcome measure; 3) separated out drug doses; 4) included unpublished studies; and 5) included data for both "completer data" and LOCF. His

6 If patients with poor outcomes drop out of the study and this occurs more frequently in placebo groups, differences between groups will be artificially small. The alternative is using a last outcome carried forward (LOCF) analysis in which dropping out is considered a treatment failure and the last measurement is carried forward and analyzed with the rest of the completion data.

technique for identifying studies was especially important; he used the Freedom of Information Act to gain access to the studies which had been submitted to the FDA to gain approval for the antidepressant. It should be noted that, from an epistemic view, these are phenomenal data. In this instance, the research community had access to the full slate of studies conducted on antidepressant medication at a certain critical juncture in the research cycle.

The data are also phenomenal from a rhetorical point of view. A proponent of pharmaceutical treatments who finds fault with the meta-analysis has only won a pyrrhic victory; any criticism of the studies used in the meta-analysis is *ipso facto* a criticism of the data used to approve the treatment in the first place. However, Kirsch, et. al., (2002) did have to admit that their previous estimation of the placebo effect (75% of the drug response) was inaccurate; when all the data was included, the placebo effect actually accounted for over 80% of the drug response. Using updated FDA data, Kirsch et al. (2008) stratified patients by initial severity of depressive symptoms and showed that though the difference between the treatment and placebo groups was statistically significant for all groups, only patients that had the highest initial severity exhibited a difference that (just barely) passed the threshold for clinical significance.[7]

Though the 80% estimate is higher than the percentage generated by published literature, Turner, et. al. (2008), showed that of the data submitted to the FDA, only 69% of studies were published. Of the 74 RCTs in the FDA database, 37 of the 38 studies that the FDA viewed as positive were published. In contrast, of the 36 studies that the FDA considered to have negative or questionable results, 22 were unpublished, 11 were published in a way that Turner, et al., judged to portray a positive outcome, leaving 3 that were published and conveyed the outcome of the study as negative. Thus, published studies, the basis for meta-analysis is nonrandom.

Given the reliance of the scientific community on pharmaceutical companies to publish data and

---

7 It has long been noted that very small differences can be *statistically* significant if the group sizes are large enough. *Clinical* significance is supposed to capture a qualitative difference in improvement.

evidence that published studies are a nonrandom selection of completed studies, there is a plausible argument to be made that the FDA data set is the most reliable data set available.[8] If this argument is granted, then even a small sandstone effect would cause the difference between treatment and placebo group to drop below clinical significance. If there is a significant sandstone effect, it could account for the entire observed difference. It is to an example of employing causal modeling that I now turn.

Methods

Sample: This data set was obtained by contacting researchers who worked on clinical trials. Ultimately, a data set was provided by Dr. Elkin, the principal investigator of the Treatment of Depression Collaborative Research Project, a multi-site study funded by the National Institute of Mental Health.

Participants: The trial randomly assigned 118 patients to receive either imipramine and clinical management, (IMI-CM) or placebo and clinical management (PLA-CM). The first two analyses will use the end point 204 sample to make use of all the available data in establishing general phenomena (PLA-CM, n = 35; IMI-CM, n =46). The third and fourth analysis will be restricted to the 71 patients who completed the study (PLA-CM, n = 35; IMI-CM, n =36) to assess previous findings. Further details on the sample can be found here (Elkin, *et al.,* 1989).

Measures

Patients were assessed at the beginning of the trial for the presence of side effects and every week thereafter. For each side effect, the rater included a measure of severity from "not present" to "severe". Two measures of side effects were constructed. The first simply recorded whether a participant experienced a given side effect over the first three weeks of the trial. The second took into account

---

8 However, even the FDA data set may be biased towards inflated efficacy of antidepressants. Funnel plots can be used in meta-analytic studies to detect publication bias. Kirsch, et. al., (2008) noted that funnel plots conducted on the "complete" data set sent to the FDA suggest what I suppose can only be referred to as "unpublication bias". This could be construed as evidence that perhaps not all data had been submitted to the FDA. As this is against FDA guidelines, other explanations of the effect were offered. One explanation not offered is that drug companies often have the ability to review data midway through a study and halt studies with unpromising results. If true, the difference between placebo and drug groups may be even smaller.

severity and duration. Absence of a side effect was coded as zero, a mild side effect as one, a moderate side effect as two, and a severe side effect as three. In line with previous recommendations, scores reflected the difference from intake. Thus, a participant who had a mild headache at intake and a moderate headache for the next three weeks would obtain a score of three. Conceptually, this measure tries to capture physical changes that occur after treatment begins. Both measures are specifically targeted at the early weeks to determine if there are immediate differences in side effects between groups. Patients completed the Hopkins Symptom Checklist-90 (HSCL-90) at intake, eight and sixteen weeks as measure of overall symptoms.[9] Improvement scores reflect change from intake. In the eighth and sixteenth week of the trial, raters indicated if the participant spontaneously indicated which group they believed they were in during the assessment and recorded their own opinion.

Statistical Analyses

The first analysis uses independent sample t-tests to determine if treatment groups experienced equal side effects and if failure to take into account duration and severity obscures group differences. The mean score for IMI-CM on the second side effect measure was tested using a one sample t-test against the prediction generated from multiplying the ratio of means for PLA-CM on both measures with the mean for IMI-CM on the first side effect measure. The second analysis uses a binomial distribution to assess whether participants and/or raters can identify group assignment above chance levels. Results are similar in the eighth and sixteenth weeks. The former are reported. The third analysis uses logistic regression to determine whether side effects and/or improvement influence beliefs about group assignment. Results are similar in both the eight and sixteenth weeks. The latter are reported. The final analysis employs a series of ordinary least squares (OLS) regressions to determine whether belief about group assignment mediates the effect of imipramine. The t-statistic is calculated as per recommendations showing it to be the most reliable test of mediating effects (MacKinnon *et al*. 2002).

9 Analyses from the Beck Depression Inventory and the Hamilton Rating Scale for Depression are not presented as the initial difference between groups was non-significant.

Results

Side effect measures obscure duration and severity

Participants receiving imipramine scored higher on both the binary measure (IMI-CL: M = 3.54±1.26, PLA-CM: M = 2.85±1.56; p = .018; 95% CI = (.046, 1.34)) and the measure taking into account duration and severity (IMI-CL: M = 7.14±3.5, PLA-CM: M = 3.13±2.4; p < .001, 95% CI = (2.67, 5.36)). If the second measure was simply a linear transformation of the first (that is, duration and severity were obscured equally in both groups by standard measures) then the expected mean score for the drug group on the second measure would be 3.89. A one sample t-test shows that it is not (p < .001, 95% CI = (2.14, 4.36)). The second measure is used in subsequent analyses.

Failure of blinding procedures

Both patients and independent raters can determine which group patients have been assigned to (p < .001; p < .001) (Table 1). It might be objected that, as the guesses were unsolicited, the result may be inflated because only people who were most confident in their guesses mentioned anything to the raters. However, two reasons speak against this. First, the results are in line with previous findings (Rabkin *et al.,* 1986). Second, even if we assume that every person who did not indicate their belief guessed at random, the number of correct identifications would remain significant (p = .007).

Side effects predict patient beliefs, improvement does not

Improvement does not significantly affect which group patients believe they are in (p = .14, n.s.; Table 2) or rater's beliefs about group assignment (p = .76, n.s.; Table 3). Side effects continue to significantly predict patients' beliefs after improvement has been removed from the model (Tables 3 and 4). Interpreting the model statistics, we can see that all models fit the data. The Hosmer & Lemeshow statistic tests this as a null hypothesis. The pseudo R^2 is a measure which attempts to capture the variance in the dependent variable explained by the independent variables. However, unlike R^2 in OLS regression, these statistics do not have the precise mathematical definition they do in

OLS regression and should be interpreted with caution. Table 4 contains a model prediction for which group the patient (or rater) would believe herself (or the patient) to be in. As can be seen by comparing the pseudo R^2 values and the model predictions, removing the improvement variable makes no significant difference to the models.

Earlier it was noted that two hypotheses had been offered for how participants in randomized trials are able to determine which group they are in: (1) Patients might guess based on whether they were experiencing side-effects; (2) Patients might guess based on whether they had improved. These hypotheses are both tested by the logistic regression. The logit is the natural log of the (probability of x) / 1 – (the probability of x). The β represents the regression coefficient for predictor and exponentiating β yields the odds ratio. When holding other predictors constant, the odds ratio indicates the change in the odds of guessing drug per unit change in the independent variable. For example, the odds ratio of the side effect variable was 2.66; if a patient was equally likely to believe they were in the drug group as the placebo group, a patient with one more side effect would be 2.66 times more likely to believe they were in the drug group. From these regressions, collectively we can see that increases on the side effect measure predict increases in the likelihood of believing that the patient is in the drug group, while improvement over the course of the trial is not a significant predictor of belief.

The effect of imipramine is completely mediated by patient belief

While receiving imipramine was a significant predictor of improvement, patients could also determine group assignment. To disentangle these effects, a mediator analysis was run. Let A be the treatment condition, B be the group the patient believes herself to be in, and C be symptomatic improvement. The first regression shows that there was a significant difference between treatment conditions ($p = .04$). Next, the paths between A → B and B→ C were assessed independently. Again, each path was statistically significant ($p < .001$ and $p = .004$, respectively). Finally both A and B were put into the model and a t-statistic for mediation was calculated. This analysis shows that the effect of imipramine is mediated by patients' belief about group assignment ($p = .04$). Because the effect of treatment group

is non-significant (p = .49. n.s.) when both treatment group and patient belief are included in the model, this can be a considered a case of complete mediation (Table 5). That is, after controlling for what group a patient believes they are in, there is no added benefit to actually taking imipramine.

Discussion and Limitations

One general limitation that applies to all four analyses is the ability to generalize these results. Each analysis relied on the detailed records kept by TDCRP, and there may not be sufficient data in other trials already conducted to reanalyze them as done here. While detecting results with so few cases can be an indication that the effect one is detecting is large, the small sample size also raises the worry that the results are idiosyncratic. Such analyses should be conducted with other data where possible. These general issues aside, I will now consider each individual result in more detail.

The first analysis noted that traditional measures of side effects do not take into account duration and severity. The results demonstrate that this causes the side effect profiles of the drug and placebo groups to appear similar. Due to the extremely large effect size and general plausibility considerations I believe that these results are robust. Further, the data were only entered through the third week; as the trial proceeded, the groups pulled farther apart. Given that patients who are in drug groups are far more likely to drop out for adverse side effects, and that patients in the placebo group are far more likely to drop out due to lack of effect, such results are probably typical.

The second analysis showed that patients can predict which group they are in. Given that this result replicates other results in the literature, it too is likely to be robust. Again, it bears mentioning that very few trials assess whether the blinding was successful, and thus this result may be difficult to replicate widely in the extant body of research. However, the fourth result underscores how important such measures are to ensuring proper analysis.

The third analysis showed that side effects and not patient improvement predicted patients' guesses as to which group they were in. Prior results (Rabkin *et al.,* 1986) claimed the contrary, but their analysis was restricted in three ways. First, they only used side effect assessments obtained at

week six. Second, they did not take into account severity. Third, their analysis was within groups. They showed that patients in the placebo group who believed themselves to be in the drug group did not have more side effects than patients who correctly believed themselves to be in placebo group. Similarly they showed that patients in the drug group, who thought they were in the drug group, did have more side effects than patients who correctly identified themselves to be in the drug group. But what is at issue is whether people who are initially blinded use side effects to determine which group they are in. This analysis shows they do.

A further question could be explored, which is whether patients who were in the placebo group make more use of improvement as an indicator in determining the group to which they belong. It seems plausible to think that patients in the drug group can discern their assignment by obvious somatic effects, while patients in the placebo group do not realize that a failure to get such a signal is in fact a signal that they are in the placebo group. In this case, they may look to early improvement to determine which group they are in. Though plausible, this was not borne out in the current analysis, though this could be due to lack of power to detect the effect of symptomatic improvement on patients' guesses, an effect which may nonetheless exist. Further it should be noted that side effects did not do nearly as well at predicting raters' beliefs. Further work should be done to improve the model.

Finally, the fourth analysis demonstrated that when the patients' beliefs about which group they are in is controlled for, there is no evidence that the drug is efficacious. Another way to say this is, being on imipramine causes side effects, and those side effects cause expectancy effects, which in turn result in better clinical outcomes. Here we see profound evidence of the sandstone effect. Because the fourth analysis is the most provocative, I will consider it in detail.

First, there are non-statistical considerations that make this result unsurprising. There are currently multiple classes of drugs that are approved for treating depression. Some act on serotonin, some act on dopamine, some act on norepinepherine, some act on serotonin and norepinepherine, some act on norepinepherine and dopamine, some act on serotonin, dopamine and norepinepherine; and *all*

are equally effective (Turner, *et al.,* 2008). Now three possibilities suggest themselves to account for this: (1) There are several different subtypes of depression that respond to different drug classes and these subtypes are equally prevalent in the population; (2) One neurochemical is primarily responsible for the ameliorative effect of antidepressants and other drugs affect this neurochemical indirectly in such a way that essentially the same result is achieved; and (3) they all share a common property which is responsible for the effect in each case. I think a case can be built for the third option, and some evidence has been provided here.

That said, there are several reasons to question these results. Though the second analysis rejects the hypothesis that improvement contributes to patients' predictions, this may not be the case. To the extent that patients are guessing based on improvement and not on side effects, then controlling for guessing is controlling for improvement. Thus, utilizing a variable that is beyond reproach would greatly strengthen the case presented above. Further work can be done to construct such a variable, building on the third analysis.

The second is that the strength of this result is dependent in an odd way on the soundness of patients' ability to determine which group they are in. MacKinnon *et al*. (2002) note that while the Freedman & Schatzkin (1992) is generally the best measure of mediation, if $A \rightarrow B = 0$, then the test has vastly inflated Type I (false positive) error rates. Thus, if it is not the case that patients can distinguish which group they are in, then the probability that this result is incorrect is dramatically increased.

Lastly, a case was deleted from the analysis on the basis of the regression diagnostics because it was overly influential on the regression. That is, this case was unrepresentative of the broader pattern in the data. Such practice is common in regression:

> The basic problem with influential cases is that they have an inordinate impact on the results of our regression analysis… Indeed, the best definition of an influential case, both statistically and intuitively, relies on this property. *An influential case is any case that significantly alters the value of a regression coefficient whenever it is deleted from*

> *an analysis.* If the deletion of particular cases in an analysis alters the parameters of the regression equation significantly, then these cases represent influential cases (Allen, 1997, pg. 171).

One case in particular (patient 372 at site 2) was identified as a significant outlier. This patient was in the placebo group, believed that he or she was in the drug group, was the third most severe case at the beginning of the trial (of 239), and the seventh most severe case at the end of the trial.[10] Including this single patient in the analysis increases the coefficient for the drug effect in the final model (β') by almost 250% (from .122 to .304), and reduces the coefficient for patient belief in that model by a factor of nearly 10 (from .406 to .048). This is an influential case indeed! Including it in the model obscures the effect that is representative of the majority of the data. However, a critic might be inclined to point out that this case is so influential precisely because it is a counterexample to the analysis proposed above. In response and in support of the deletion of this case, I assert that the purpose of regression analysis is to best describe the trend present in the majority of the data; the exclusion of this case is in line with that aim. On the other hand, this speaks to the need for replicating this result in other data sets. In a larger sample the impact of such an extreme case would have less influence on the data. This was the only case examined for deletion, I have noted it, and I will leave the reader to evaluate its propriety.

The Gold Standard

The importance of RCTs is based on the realization that when patients believe they are receiving treatment, they often improve. Indeed controlling for expectancy effects is the main motivation behind the elaborate blinding procedures employed in clinical trials. However, the science of medicine will be severely hampered if steps are not taken to ensure that the experimental manipulations employed are successful. If double-blind studies are as important as they are made out to be, then it must surely be important to confirm that people are truly blind.

Above, I presented data showing that neither patients nor raters were sufficiently blinded. But a

10 Recall that Kirsh *et al*. (2008) showed that expectancy effects are less effective for severe cases.

blindfold that one can see through is nothing but a mask. If these results are typical, then calling such trials 'double-blind' is nothing but the guise of rigor. Thus, perhaps the most important implication of these results is that the label 'double-blind' should be attached, not to the way the trial was designed, but to whether the design was successfully implemented. Were this standard to be followed, the current category of 'RCT' would be bifurcated into trials that included randomization, placebos and a failed attempt at blinding, and (truly) double-blind RCTs.

Likewise we might also consider what is required for a substance to be a placebo. Traditionally research employs a material definition of a placebo: it is an inactive substance, used in RCTs, likely a sugar pill. In contrast, suppose that what a placebo is is an empirical question. If a placebo is defined as a substance which serves as an adequate control for the non-specific components of proposed treatment, then a placebo is some substance which has no clinical effect yet prevents detection of group assignment. With this definition, what a placebo is turns out to be something one must discover. If one adheres to a methodological definition of what a placebo is, then one can in turn ask whether a sugar pill is a placebo. In the case at hand we could suggest that it is not, because it is ineffective at preventing participants from determining which group they are in. If one insists that what a placebo is is a matter of definition, then I assert that our definition of 'effective' is insufficient for reasons which should now be obvious. If the definition of a placebo is an empirical question, I assert that a sugar pill is not a suitable placebo for imipramine. Either way something must give.

Though the use of active placebos has been suggested to address these issues, what has been unappreciated is that, given more sophisticated statistical approaches, there is a solution that can be carried out with the type of data already collected, as has been demonstrated here. The need for such rigor is fueled by the growing appreciation that antidepressants are not as effective as previously believed. This analysis should add to those concerns. First, it shows the importance of considering duration and severity when considering adverse effects. Second, it shows the importance of empirical verification of methodological design. Actual ignorance of group assignment must be assessed, not

only to ensure the integrity of the trial, but because standard statistical analysis relies on it. As demonstrated here, when the blind fails, a drug can appear superior to placebo when it does nothing more than signal to patients that they are receiving an active substance, creating an expectancy effect. Proponents commonly reply that antidepressants still have clear superiority in severely depressed patients (Fournier, *et al*, 2010). Yet the superiority of antidepressants for this population is not accounted for by antidepressants working better, but by placebos working worse (Kirsch *et al*, 2008). A plausible account for this is that the severely depressed are less likely to believe they are receiving medication unless they receive a definite signal such as side effects.

While this analysis only considers an older antidepressant, it should be recalled that newer antidepressants do not have superior outcomes (Snow, Lascher, & Mottur-Pilson, 2000; Turner, Matthews, Linardatos, Tell, & Rosenthal, 2008). The ability to generalize these findings is strengthened by the fact that expectancy effects are likely to persist in any trial that uses an inactive placebo. Further, there are non-statistical considerations that make this result unsurprising. All antidepressants are equally effective despite drastically different neurological effects (Turner et al. 2008). Three possibilities suggest themselves to account for this: (1) There are several different subtypes of depression that respond to different drug classes and these subtypes are equally prevalent in the population; (2) One neurochemical is primarily responsible for the ameliorative effect of antidepressants and other drugs affect this neurochemical indirectly in such a way that essentially the same result is achieved; and (3) the equivalent efficacy is due to what they have in common – side effects.

It may seem that the failure of imipramine to outperform placebo in a single trial could not possibly call into question the amassed research showing its efficacy. Yet to focus on the particulars of the present analysis is to miss what is truly at stake. I do not claim to have shown that antidepressants are ineffective. Though that was the outcome in this particular trial and broader considerations suggest that it possible that antidepressants are ineffective, this evidence alone is hardly sufficient to support

such a sweeping claim. What I have argued for, is the claim that current methods of analyzing data in RCTs are insufficient in licensing the causal claims of a specific effect; the empirical data merely serves as an example of the way in which trials must be analyzed to be considered legitimate. It is not the outcome of this analysis that is crucial, but what the analysis shows us about the evidence used to support treatment efficacy more generally. It begins with a result previously published as evidence of a specific effect and shows how causal modeling can be used to disentangle expectancy effects in ways that traditional analysis leaves undifferentiated. The example demonstrates that it is possible for a significant result to disappear when the data is so analyzed. Thus, it casts doubt on the reliability of the entire body of evidence that has heretofore been brought to bear on the matter. In so doing this does not show that antidepressants don't work; it shows that we have no reason to think that they do.

References

Allen, M. (1997). *Understanding Regression Analysis*. New York: Plenum.

Bystritsky, A., & Waiker, S. (1994). Inert placebo versus active medication: Patient blind ability in clinical pharmacological trials. *Journal of Nervous and Mental Disorders, 182,* 485-487.

Elkin, R., Shea, M., Watkins, J., Imber, S. Sotsky, S., Collins, J., et. al. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: General effectiveness of treatments. *Archives of General Psychiatry, 46,* 971-982.

Fournier, J., DeRubeis, R., Hollon, S., et al. Antidepressant drug effects and depression severity. JAMA 2010;303:47-53.

Freedman, LS., & Schatzkin, A. (1992). Sample size for studying intermediate endpoints within intervention trials or observational studies. *American Journal of Epidemiology,* 136(9), 1148-.

Greenberg, R., Bornstein, R., Zboroski, M., Fisher, S., & Greenberg, M. (1994). A meta-analysis of fluoxetine outcome in the treatment of depression. *Journal of Nervous and Mental Disorders, 182,* 547-551.

Holmbeck, G., (1997). Toward terminological, conceptual, and statistical clarity in the study of mediators and moderators: Examples from the child-clinical and pediatric psychological literatures. *Journal of Consulting and Clinical Psychology, 65,* 599-610.

Kirsch, I., Deacon, B., Huedo-Medina, T., Scoboria, A., Moore, T., & Johnson, B. (2008). Initial Severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. *Public Library of Science, 5*(2), e45.

Kirsch, I., Moore, T., Scoboria, A., & Nicholls, 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, 5,* article 23.

Kirsch, I., & Sapirstein, G. (1998). Listening to Prozac but hearing placebo: A meta-analysis of antidepressant medication. *Prevention and Treatment, 1* (2), 00.

Klein, D. (1998a). Listening to meta-analysis but hearing bias. *Prevention and Treatment, 1,* article 0006c.

Klein, D. (1998b). Reply to Kirsch's rejoinder regarding antidepressant meta-analysis. *Prevention and Treatment, 1,* article 0008r.

MacKinnon, D., Lockwood, C., & Hoffman, J., (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83- 104.

Moerman, D. (2002). "Loaves and Fishes": a comment on "The emperor's new drugs: An analysis of antidepressant medication data submitted to the U.S. Food and Drug Administration. *Prevention and Treatment, 5,* article 29.

Pearl, J. (2000). *Causality: Models, reasoning and inference.* Cambridge: Cambridge University Press.

Rabkin JG, et al. (1986). How blind is blind? Assessment of patient and doctor medication guesses in a placebo-controlled trial of imipramine and phenelzine. *Psychiatry Research, 19,* 75–86.

Rehm, L. (1998). Listening to Prozac and hearing noise: A commentary on Kirsch and Sapirstein's "Listening to Prozac and hearing placebo." *Prevention and Treatement, 1,* 1522.

Snow, V., Lascher, S., and Mottur-Pilson, C. (2000). Pharmacologic Treatment of Acute Major Depression and Dysthymia. *Annals of Internal Medicine, 132,* 738-742.

Thompson, R. (1982). Side effects and placebo amplification. *The British Journal of Psychiatry*, *140,* 64-68.

Turner, E., Matthews, A., Linardatos, E., Tell, R., & Rosenthal, R. (2008). Selective publications of antidepressant trials and its influence on apparent efficacy. *The New England Journal of Medicine, 358,* 252-260.

Quitkin, FM., Rabkin, JG., Gerald, J., Davis, JM., and Klein, D. (2000). Validity of Clinical Trials of Antidepressants. *The American Journal of Psychiatry. 157* (3), 327.

Williams, J., Mulrow, C., Chiquette, E., Hitchcock, P., Aguilar, C., & Cornell, J. (2000). A systematic review of newer pharmacotherapies for depression in adults: Evidence report summary. *Annals of Internal Medicine, 132,* 743-756.

Table 1.— Patient and Rater Predictions week 8

Patient Predictions				*Rater Predictions*			
	Predicted				Predicted		
Treatment Group	Drug	Placebo	% Correct	Treatment Group	Drug	Placebo	% Correct
Drug	26	1	96%	Drug	36	10	78%
Placebo	9	10	53%	Placebo	6	29	82%
Overall % correct			78%	Overall % correct			80%

Table 2.—Logistic Regression Analysis of Patient Predictions

Predictor	β	*SE* β	Wald's χ^2	*df*	*p*	e^{β} (odds ratio)
Constant	3.24	1.46	5.13	1	.024	27.78
Improvement	-1.53	1.03	2.21	1	.137	0.22
Side effects	0.88	0.33	7.28	1	.007	0.41

Test	Pseudo R^2	$\chi 2$	*df*	*p*
Overall model evaluation				
Cox & Snell	.387			
Nagelkerke	.560			
Goodness-of-fit test				
Hosmer & Lemeshow		5.607	8	.691

Logistic Regression Analysis of Predictions (w/o improvement scores)

Predictor	β	*SE* β	Wald's χ^2	*df*	*p*	e^{β} (odds ratio)
Constant	1.78	0.93	3.68	1	.055	5.93
Side effects	-0.85	0.32	7.00	1	.008	0.43

Test	Pseudo R^2	$\chi 2$	*df*	*p*
Overall model evaluation				
Cox & Snell	.349			
Nagelkerke	.505			
Goodness-of-fit test				
Hosmer & Lemeshow		2.348	7	.938

Table 3.—Logistic Regression Analysis of Rater Predictions

Predictor	β	*SE* β	Wald's χ^2	*df*	*p*	e^{β} (odds ratio)
Constant	2.33	0.93	6.32	1	.012	10.28
Improvement	-0.21	0.66	0.10	1	.756	0.81
Side effects	-0.65	0.17	13.39	1	.000	0.52

Test	Pseudo R^2	$\chi 2$	*df*	*p*
Overall model evaluation				
Cox & Snell	.352			
Nagelkerke	.477			
Goodness-of-fit test				
Hosmer & Lemeshow		10.949	8	.205

Logistic Regression Analysis of Predictions (w/o improvement scores)

Predictor	β	*SE* β	Wald's χ^2	*df*	*p*	e^{β} (odds ratio)
Constant	2.02	0.64	10.01	1	.002	7.53
Side effects	-0.59	0.16	13.99	1	.000	0.55

Test	Pseudo R^2	$\chi 2$	*df*	*p*
Overall model evaluation				
Cox & Snell	.335			
Nagelkerke	.454			
Goodness-of-fit test				
Hosmer & Lemeshow		7.205	6	.302

Table 4.—The Observed and the Predicted Frequencies for Treatment group by Logistic Regression with the Cutoff of 0.50

Regression Model for Patient Beliefs (model predictions including improvement)

Observed Belief	Predicted Belief: Drug	Predicted Belief: Placebo	% Correct	
Drug	29 (29)	3 (3)	91%	Sensitivity = 91%. Specificity = 66%.
Placebo	4 (4)	8 (8)	66%	False positive = 12%. False negative = 27%.
Overall % correct			84%	

Regression Model for Rater Beliefs (model predictions including improvement)[*]

Observed Belief	Predicted Belief: Drug	Predicted Belief: Placebo	% Correct	
Drug	29 (30)	12 (11)	71%	Sensitivity = 71%. Specificity = 71%.
Placebo	8 (7)	20 (20)	71%	False positive = 22%. False negative = 38%.
Overall % correct			71%	

*Note: There is one participant that did not have an improvement score

Table 5.—Moderator analysis of 36 Patient Predictions

Predictor	β	*SE* β	t	*df*	*p*	R^2
			A → C			
Treatment Group	.318	.146	2.181	35	.036	.123
			A → B			
Treatment Group	.486	.126	3.870	35	<.001	.306
			B → C			
Believed Group	.480	.157	3.058	35	.004	.216
			A → B → C			
Treatment Group	.122	.167	0.784	34	.486 (n.s.)	
Believed Group	.406	.190	2.125	34	.041	
Model						.228

Key: A- Treatment group; B-Side effects; C- Improvement
Note: The t-statistic for the mediation is calculated using

$$t_{N-2} = \frac{\tau - \tau'}{\sqrt{\sigma_{\tau}^2 + \sigma_{\tau'}^2 - 2\sigma_{\tau}\sigma_{\tau'}\sqrt{1 - \rho_{XI}^2}}}$$

Where τ is the β from the regression with just the treatment group, τ' is β from the regression with both variables in the model, and ρ is the correlation between treatment group and believed group (r = .553). Given the observed values, the null hypothesis that τ - τ' = 0 (viz. the effect of treatment is equivalent when belief is added into the model) is rejected (t (34) = 2.12, p = .042)