Why Most Sugar Pills Are Not Placebos

Bennett Holman*†

The standard philosophical definition of placebos offered by Grünbaum is incompatible with Cartwright's conception of randomized clinical trials. I offer a modified account of placebos that respects this role and clarifies why many current medical trials fail to warrant the conclusions they are typically seen as yielding. I then consider recent changes to guidelines for reporting medical trials and show that pessimism over parsing out the cause of "unblinding" is premature. Specifically, using a trial of antidepressants, I show how more sophisticated statistical analyses can parse out the source of such effects and serve as an alternative to placebo control.

1. Introduction. It seems rather obvious that when our experiments are potentially confounded, the prudent response is to take better precautions. Yet given the mounting empirical evidence that randomized clinical trials (RCTs) designed to be "double-blind" fail to remain so, the CONSORT 2010 guidelines have dropped their recommendation to report whether blinding was successful. They reason that if a treatment has a significant therapeutic effect, maintenance of the blind may be impossible. I will argue that this change represents a step backward and ultimately results from an inadequate definition of "placebo."

While Cartwright's (2010) analysis has illuminated the experimental logic undergirding ideal RCTs, worldly vagaries put this logic in tension with the standard philosophical account of placebos offered by Grünbaum (1986). Accordingly, the first tasks of this article will be explicating Grünbaum's conception of a placebo and substantiating the claim that, defined as such, placebos are incompatible with Cartwright's ideal. Next, I will provide a

*To contact the author, please write to: Veritas B Room 419 (UIC Office), Yonsei University, 85 Songdogwahakro, Yeonsu-gu, Incheon 406-840, Korea; email: bholman@yonsei.ac.kr.

†In memory and appreciation of my father, who encouraged my pursuit of philosophy even when he didn't understand it and who read no fewer than seven versions of this paper, never failing to find a way to make it better. Any remaining errors are his fault.

Philosophy of Science, 82 (December 2015) pp. 1330–1343. 0031-8248/2015/8205-0050$10.00

https://doi.org/10.1086/683817 Published online by Cambridge University Press

methodological definition of a placebo that resolves the incompatibility. Finally, I will explore two ways in which RCTs could be altered to bring practice in line with this resolution. Regardless of how this incompatibility is resolved, clarifying the nature of the tension is salutary. It both identifies the situations where medical trials fail to warrant the conclusions they are standardly taken to demonstrate and suggests how they might be amended so as to yield more reliable results.

2. Grünbaum's Placebo. Behind Grünbaum's (1986) work on placebos was an internecine debate in the mental health community over the efficacy of various schools of psychotherapy. With the example of psychotherapy in the foreground, it was clear that a placebo is not just any treatment that relies on psychological effects. It is similarly flawed to identify placebos with inert substances. Water, salt, and sugar are paragons of therapeutically inert substances. But while inert generically, they can be effective treatments for dehydration, hyponatremia, and hypoglycemia, respectively. Such modest examples elucidate Grünbaum's central insight: that a treatment *t* is a placebo with respect to some disease *D*.

More specifically, Grünbaum divides a treatment as follows: a therapy (*t*) is explained by a theory ψ as being composed of (*a*) the characteristic factors *F* that are purportedly responsible for improvement and (*b*) the incidental factors *C* that are not (see fig. 1). For example, suppose that a therapeutic theory holds that the chemical fluoxetine is the proper treatment for depression. Since Prozac and Sarafem are chemically identical (they are both fluoxetine), they have identical characteristic factors. Yet while they share some incidental factors (e.g., they are both manufactured by Eli Lilly), there are others they do not share (e.g., a pill of Sarafem is pink and purple, while Prozac is green and white; Sarafem is roughly 15 times more expensive).

Next, Grünbaum divides the effects of the treatment on the patient (he calls the patient a "victim") into (i) the intended effects on the target disorder *D* and (ii) other aspects of the patient's life (i.e., side effects). Staying with the example above, Prozac might cause improvement in mood, as intended, but unintentionally cause sexual dysfunction.

A treatment *t* is "an intentional placebo" (a placebo intended for treatment) for victim *V* suffering from disorder *D* treated by practitioner *P* iff (1) none of the characteristic factors *F* positively affect *D*, (2) *P* believes that the first condition is true, (3) *P* believes that some incidental treatment factors *C* will positively affect *D* for *V*, and (4) *P* generally allows *V* to believe that *t* has remedial efficacy for *D* because of *F*. More plainly, condition 1 entails that a sugar pill is a placebo as long as the patient was not helped by the sugar. Conditions 2–4 specify that the patient, but not the doctor, believes that the treatment is effective because of some characteristic factor. By these conditions, most sugar pills are placebos.

https://doi.org/10.1086/683817 Published online by Cambridge University Press

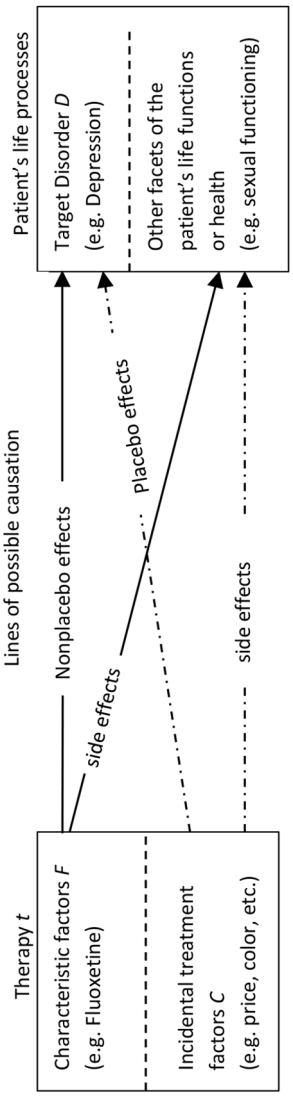

Figure 1. An example of Grünbaum's model used to specify drug/placebo effects.

https://doi.org/10.1086/683817 Published online by Cambridge University Press

3. The Crystal Palace. To highlight potential problems with Grünbaum's conception of a placebo, consider the following thought experiment. Suppose that crystal healing is the standard form of treatment, but we were inclined to doubt the efficacy of crystals. Crystal healers (and thus, ex hypothesi, the population at large) believe that sandstone is the appropriate treatment for some disorder. To assuage our concern, a crystal healer performs an RCT. In the experimental group, sandstone is used to treat the patients. In the control group, amethyst (which has no expected effect) is used instead. The amethyst and the sandstone are enclosed in a device that allows for the appropriate skin contact while preventing the patient and the healer from seeing which is being used. In follow-up evaluations, while both groups have improved considerably, the treatment group has significantly better outcomes. Suppose further that this finding is repeatedly confirmed. Healers conclude that this constitutes scientific evidence to support the widely accepted claims of efficacy.

Such trials would meet the general standards of evidence-based medicine and even the requirements of the Food and Drug Administration. Standard practice allows that the difference in outcomes between patients who received sandstone and patients who received amethyst can be attributed to the effectiveness of the sandstone treatment. Amethyst also meets Grünbaum's (1986) definition of a placebo control (a placebo intended for a trial) iff (1) none of its characteristic factors positively affect D, (2) P believes that amethyst is harmless for V with D, and (3) P wishes to know whether any of the observed improvement can be attributed to the characteristic factors.

Yet we can grant each of these conditions while maintaining that the test was not properly controlled. Suppose I hypothesized that it is possible to determine patients' group assignment during the trial. To support this, I produce evidence that patients, doctors, and independent evaluators are all able to make this discrimination at levels above chance. There are two possible explanations for this. It could be because they correctly believe that sandstone works. For example, subjects who recover reason that since they recovered, they must have been treated with sandstone. 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 "nontherapeutic" aspects (e.g., side effects) of sandstone effectively, allowing subjects to discriminate between treatments based on properties incidental to treatment.

In line with the latter hypothesis, suppose that when sandstone is rubbed on the body, it causes abrasions, whereas amethyst does not. Further, suppose I show that patients guessed their treatment assignment on the basis of abrasions, not improvement. Finally, suppose we find the following: when you statistically control for which group patients believed they were

https://doi.org/10.1086/683817 Published online by Cambridge University Press

in, patients are no better off if they were actually exposed to sandstone. In light of these facts, I believe that a reasonable conclusion is that sandstone was not an effective treatment despite its apparent superiority. Its only independent effect was to cause abrasions. The presence of abrasions signaled to subjects that they were receiving the culturally accepted medical practice, which in turn resulted in improvement via an expectancy effect. Because the abrasions were a side effect (not a characteristic or incidental factor), such an expectancy effect does not realize Grünbaum's conception for either a placebo effect or a nonplacebo effect (see fig. 1). Let us call this "the sandstone effect."[1]

4. The Placebo Control Revisited. The methodological definition of a placebo is determined by the underlying logic of ideal RCTs. As articulated by Cartwright and Hardie (2012), the ideal RCT is a manifestation of Mill's method of difference. Patients are randomized into two groups. The control group receives precisely the same care as the treatment group except for x, where x is what is being evaluated for efficacy. Given that patients receiving treatment t improve and the only difference between t and t^* is x, x must be the cause of the improvement.

Faced with the situation above, it seems that we must either give up the standard definition of placebo or conclude that an RCT is insufficient for establishing efficacy. In the remainder of the article I will parse out the consequences for the former. In place of the traditional definition, I will propose the methodological definition of a placebo. In essence, the methodological definition turns "placebo" into a success term. As a first pass, a treatment t^* is a placebo in an RCT testing t for D iff it plays the methodological role required to determine whether t is a treatment for D. As an example, let's consider arthroscopic knee surgery for osteoarthritis of the knee.

Moseley et al. (2002) randomized patients into one of three groups. In the incision group, a patient was locally anesthetized and several incisions were made, but nothing more. The second group also had their joint lavaged (washed out), and the final group had their knee lavaged and debrided (surgeons removed damaged tissue). If the actual procedure was not performed, the surgeon still went through the motions of asking for the instrument, manipulating the patient's leg, even splashing saline to simulate the sounds of lavage. In this procedure a number of factors are clearly incidental: the fact

1. To the extent that doctors alter their practice or their assessment of patients based on which group the doctor believes the patient is in, there will be a corresponding sandstone effect for doctors (since they may see through their side of the double-blind). For simplicity, I will focus on the effects on patients, but this account applies equally well to doctors.

https://doi.org/10.1086/683817 Published online by Cambridge University Press

that Moseley was a celebrated surgeon, the sound of splashing water as the joint is rinsed, and so on. Other factors are less clearly categorized.

Surely, the group we consider to be the placebo group will depend on what our theory ψ determines to be the characteristic factors of t, but much can hang on such designations. Surgeons diverge in whether they attribute the therapeutic effect to the debridement, the lavage, or both (all agree that the incision is incidental). According to ψ_1, the debridement is the characteristic factor and the lavage is merely an incidental prelude. In contrast, ψ_2 supposes that the lavage is the characteristic factor of the procedure and the debridement is both incidental and otiose. Finally, ψ_3 posits that both the lavage and the debridement are characteristic factors, each of which makes a contribution to the therapeutic effect. The correct theory of arthroscopic knee surgery classifies as characteristic factors all and only the causally relevant factors of the treatment that exert a therapeutic effect independent of the patient's expectations. The choice is not conventional; the other theories are false.

The definition of a characteristic factor above implies that any factor that does not independently improve therapeutic outcomes is an incidental factor. In the sandstone parable, the amethyst group improved (though to a lesser extent than the sandstone group); however, according to the society's prevailing theory, mere exposure to amethyst outside of the healing ritual would not have resulted in any therapeutic effect. Accordingly, any therapeutic effect that did occur must be the result of the patient's expectations.

Beyond their therapeutic effects, characteristic factors may also cause side effects (i.e., effects unrelated to the disorder). Side effects might be caused independently of patient belief, as when the incision during knee surgery results in soreness or, contra Grünbaum, as a result of expectancy effects (Kirsch 2010). For the sake of clarity I'll use "expectancy effects$_t$" for therapeutic effects and "expectancy effects$_{se}$" for side effects (see fig. 2).[2]

Returning to arthroscopic surgery and pace the subsequent controversy, we are now in a position to spell out the methodological role required to determine whether t is a treatment for D. With regard to ψ_1, only the lavage group serves as a suitable control. If the debridement group were compared with the incision group, a superior efficacy would not have established that the debridement alone was responsible for the improvement. According to ψ_1, the lavage group contains all of the incidental factors C of t and thus controls for any expectancy effects$_t$ caused by C.

2. It is worth noting that because side effects can be beneficial (e.g., better cardiovascular health is a side effect of treating depression with aerobic exercise), these are not necessarily "nocebo effects" (negative placebo effects). Moreover, since there can be negative expectancy effects on D, nocebo effects are not limited to side effects.

https://doi.org/10.1086/683817 Published online by Cambridge University Press

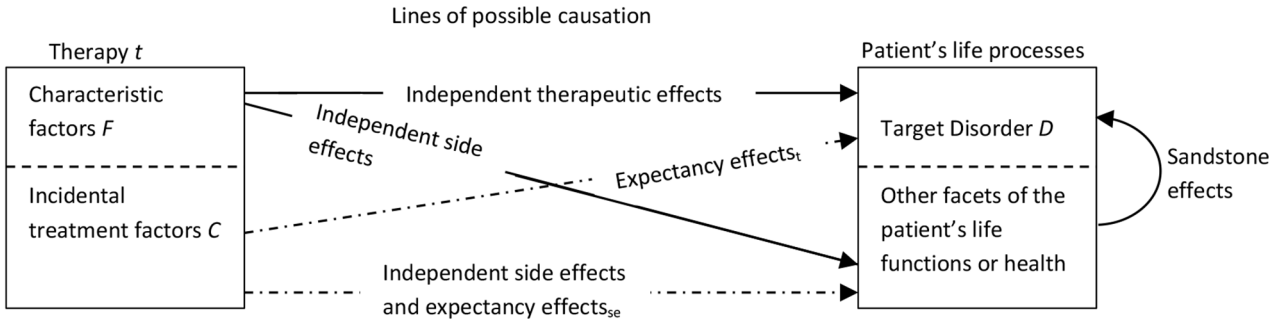

Figure 2. Grünbaum's basic model stays intact; however, expectancy effects are not limited to therapeutic effects, and side effects may impact *D*.

As it turns out, Moseley et al. (2002) found no difference between the lavage and the debridement groups. Nevertheless, this does not show that the procedure is ineffective; perhaps the lavage has an independent effect on osteoarthritis. If so, the lavage was not a placebo control but an effective therapy. Such a possibility can be assessed by examining the difference between the lavage and the incision group. This comparison is complicated by the fact that these groups had different levels of pain shortly after the surgery, allowing for the possibility of a sandstone effect. Fortunately, Moseley et al. assessed patients' beliefs about which group they were in, as well as their expectations for a successful recovery, and found no differences in either. Thus, the incision treatment serves as a placebo control for the lavage treatment. Furthermore, as Moseley found no difference between the incision and the lavage, the latter was also a placebo for debridement.[3]

In summary, the methodological definition of a placebo is as follows: *t** is a placebo in an RCT testing *t* for *D* iff (1) *t** methodologically controls for the expectancy effects$_t$ caused by *t*,[4] (2) none of the incidental factors of *t** have an independent therapeutic effect, and either (3*a*) *t** produces all of the same side effects as *t* or (3*b*) despite the failure of 3*a*, patients' beliefs about which treatment they are receiving are the same for both *t* and *t**. It is worth noting a number of nonstandard implications of this definition. First, whether *t** succeeds in serving as a placebo for *t* should be supported by evidence. Second, contra Grünbaum, it is not enough for the placebo to simply lack the characteristic factors of *t*; it must share the incidental factors. Moreover, a placebo can, and in many cases must, cause side effects independent of any expectancy effects. For example, if patients in the control group did not have to heal from the incision, the procedure would fail to serve as a placebo control by virtue of the fact that it caused no side effects. I think that one can, if one wishes, maintain that a "placebo effect" that is independent of a patient's expectation violates one's concept of a placebo (i.e., placebos should be inert); however, as alluded to above, I do not think that one can maintain that position and continue to have placebos serve the methodological role they currently occupy in medical research.

5. An Alternative to Placebos. The view rehearsed above cuts against current trends in medical research. Recently, in light of the fact that blinding consistently fails in practice, the CONSORT 2010 guidelines retracted their previous recommendation for trials to test to see whether blinding was successful, claiming, "We now believe the interpretation of a test of its

3. Technically, it could be the case that the incision had some independent effect on osteoarthritis. Although nothing rules this out, I know of no one who takes this position.

4. Essentially, *t** contains all and only the incidental factors *C* of *t*. This condition is similar to Howick's (2012, 82) requirement for a "legitimate placebo."

https://doi.org/10.1086/683817 Published online by Cambridge University Press

[blinding] success to be dubious, if not impossible" (Schulz et al. 2010, 1145). They claim that because blinding might fail owing to patient improvement, such a test may simply reflect the efficacy of the drug instead of the methodological failure of the trial. Recall that the logic of RCTs is that to attribute the efficacy to the drug, the only difference must be characteristic factors. Thus, the CONSORT 2010 guidelines essentially confront a possible failure of the central assumption required by the statistical analysis used to assess efficacy with the unsubstantiated hope that everything will work out.

Fundamentally, this problem is an outgrowth of the fact that inert substances often fail to serve as adequate placebos. If RCTs used a placebo in the sense I propose, no such ambiguity would arise, as the therapeutic effect would be the only possible cause of unblinding. However, since most drugs have side effects, so would most placebos. Nevertheless, for either practical or ethical reasons, it may not be possible to give patients a substance that causes all of the negative effects known to be caused by treatment, but without any of the characteristic factors hypothesized to be beneficial. An alternative to the methodological control is to administer an inert substance (i.e., not a placebo) and attempt to control for these effects statistically.

Consider a trial on antidepressants carried out by the National Institute of Mental Health that used an inert substance in the control group.[5] Given the methodological definition of a placebo, whether the trial was in fact placebo controlled must be substantiated. Since the control group was given a lactose pill that was identical in appearance to the antidepressant, the design of the trial provides prima facie evidence that the placebo group controlled for all of the incidental factors of antidepressant treatment without containing any independent therapeutic effects. However, while the first two conditions of a placebo were met, patients in the antidepressant group experienced far more side effects. Moreover, patients were able to guess which group they were in at levels far above chance. In short, the trial was not a placebo-controlled experiment, and the possibility that the observed superiority is due to a sandstone effect cannot be ruled out.

The ability of participants to correctly identify which group they were in does not entail that the drug was ineffective, but the standard practice of comparing group means is not justified. Instead, a more complicated procedure should be employed to achieve statistically what a placebo is intended to achieve methodologically. The null hypothesis is that there is no therapeutic effect of the drug. Accordingly, if patients were able to identify which group they were in on the basis of therapeutic improvement, then the

5. In what follows I will describe my own results in less than full detail. However, a fuller analysis can be found in the technical appendix, and greater detail can be provided upon request.

https://doi.org/10.1086/683817 Published online by Cambridge University Press

null hypothesis is false (the treatment is effective) and the logic of the RCT is not threatened (Howick 2012). In line with such reasoning, proper analysis must ascertain the cause of patients' beliefs about group assignments. If, as in the antidepressant trial, patients' guesses were the result of differences in side effects, a sandstone effect cannot be ruled out and we cannot reject the null hypothesis.

In such cases there is a solution that can be carried out with the type of data already collected in many trials. Instead of testing group differences, structural equation modeling (SEM) or regression analysis 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 et al. 2002).

First, suppose that we have three variables *A* (treatment), *B* (belief about group assignment), and *C* (symptomatic improvement) that are each highly correlated, and we wish to determine whether *A* has an independent effect on *C*. This results in a model structure with three paths as in figure 3. 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*, (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.

To get a feel for the model, consider the crystal palace thought experiment. To assess crystal healing, let *A* be the actual treatment, let *B* be which treatment the patient believes they are receiving, and let *C* be 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 variance in improvement 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

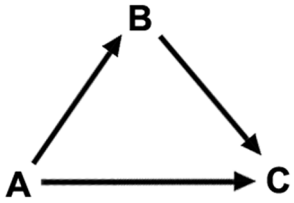

Figure 3. Possible causal model in which the treatment has both a direct effect on symptomatic improvement (the arrow between A and C) and an indirect effect via the treatment's effect on patients' beliefs concerning group assignment (the arrows from A to B and B to C).

https://doi.org/10.1086/683817 Published online by Cambridge University Press

[*A*] 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.

With this more sophisticated analysis, we can see that the pessimism of CONSORT 2010 is unfounded. With sufficient data, it is entirely possible to identify the cause of unblinding and subsequently tease apart the direct effect of the treatment from the expectancy effect caused by side effects. In the trial of antidepressants, it turns out that patients' guesses are driven by side effects and not improvement. Patients who received antidepressants improved more than the control group, and patients who believed they were receiving antidepressants improved more than patients who believed they were receiving placebos. Crucially, once what a patient believes is accounted for, there is no added benefit to actually receiving antidepressants. In short, this trial, once properly controlled for statistically, provided no evidence that antidepressants were an effective treatment. The apparent superiority of antidepressants in this case is caused by the sandstone effect. Given that the patients (and doctors) could determine which group the patients were in by the presence of side effects, the lactose pill does not meet the methodological criteria to be considered a placebo; however, we might reasonably call trials analyzed in this fashion *statistically controlled RCTs*.

6. Conclusion. The standard meaning of "placebo" stands in need of revision. All too often trials are called "double-blind" or "placebo controlled" on the basis of their design. There is now significant evidence that for any condition that is susceptible to expectancy effects treated by a drug that causes side effects, providing an inert substance to the control group will fail to warrant the assumptions that are standardly used to analyze the data. The same rationale that motivated the introduction of inert substances into trial design can be marshaled to argue for a truly adequate placebo. Nevertheless, ethical or practical concerns may militate against the use of such placebos. In such cases, statistical controls can be used to attenuate methodological shortcomings. In either event, trials must be described not by their intended design but by what conditions actually obtain, and reporting guidelines should be changed accordingly. If researchers accepted such methodological strictures, then it is overwhelmingly likely that they will conclude that most sugar pills are not placebos.

Technical Appendix

This data set was obtained by contacting Dr. Elkin, the principal investigator of the Treatment of Depression Collaborative Research Project, a multisite study funded by the National Institute of Mental Health. Results here com-

https://doi.org/10.1086/683817 Published online by Cambridge University Press

pare the groups receiving imipramine (IMI-CM) with those receiving the placebo (PLA-CM) and use the Hopkins Symptom Checklist-90 (HSCL-90) as the measure of improvement. The first two analyses 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 analyses will be restricted to the 71 patients who completed the study (PLA-CM, $n = 35$; IMI-CM, $n = 36$) to assess previous findings. One case was deleted from the fourth analysis on the basis of the regression diagnostics (it was overly influential on the regression). Further details on the sample can be found here (Elkin et al. 1989).

Statistical Analyses. The first 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 second 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 by the method described in Freedman and Schatzkin (1992) as per recommendations showing it to be the most reliable test of mediating effects (MacKinnon et al. 2002).

Results

Failure of Blinding Procedures. Of the patients receiving imipramine, 26/27 believed that they were receiving the drug. Patients receiving a placebo guessed roughly at chance (10/19). Combined, patients had correct beliefs 78% of the time ($p < .001$).

Side Effects Predict Patient Beliefs, Improvement Does Not. Side effects ($p < .001$), but not improvement ($p = .14$, NS), influence which group patients believe they are in. Side effects continue to significantly predict patients' beliefs after improvement has been removed from the model ($p < .001$). The model's fit cannot be rejected using a Hosmer–Lemeshow test for lack of fit ($\chi^2 = 7.83$, $df = 7$, $p = .351$). In cases where improvement was an independent predictor of patient belief, SEM might be used in place of the analysis below.

The Effect of Imipramine Is Completely Mediated by Patient Belief. Let A be the treatment condition, B be the group the patient believes himself or herself to be in, and C be symptomatic improvement. First, $A \rightarrow C$ is assessed to determine whether treatment group affects improvement. An

https://doi.org/10.1086/683817 Published online by Cambridge University Press

TABLE A1. Mediator Analysis of Patient Predictions

Predictor	β	SE β	t	df	p	R^2
$A \rightarrow C$:						
Treatment group	.318	.146	2.181	35	.036	.123
$A \rightarrow B$:						
Treatment group	.486	.126	3.87	35	<.001	.306
$B \rightarrow C$:						
Believed group	.48	.157	3.058	35	.004	.216
$(A \ \& \ B) \rightarrow C$:						
Treatment group	.122	.167	.784	34	.486	NS
Believed group	.406	.19	2.125	34	.041	
Model						.228

OLS regression shows that there is a significant difference between treatment conditions (see table A1). Next, the paths between $A \rightarrow B$ and $B \rightarrow C$ are assessed independently. Again, each path is statistically significant. Finally, both A and B are put into the model. The t-statistic for mediation is calculated with the equation (Freedman and Schatzkin 1992) $t_{N-2} = \tau - \tau'/(\sigma_\tau^2 + \sigma_{\tau'}^2 - 2\sigma_\tau\sigma_{\tau'}1 - \rho^2)^{1/2}$, 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 ($r = .553$). Entering the observed values into the equation, we get $t(34) = 2.11$, $p = .042$. Thus, the null hypothesis that $\tau - \tau' = 0$ (i.e., the effect of treatment group is equivalent when side effects are added into the model) is rejected.

MacKinnon et al. (2002) differentiate between complete and partial mediation. Complete mediation occurs when β' is not significantly different from 0; a partial mediation occurs when the test for mediation is significant but β' remains significant. The hypothesis that $\beta' = 0$ is evaluated in the last model, and the hypothesis cannot be rejected ($p = .486$). Thus, this presents a case of complete mediation. That is, the benefit of the antidepressant disappears once we take into account which group patients believed themselves to be in.

REFERENCES

Cartwright, Nancy. 2010. "What Are Randomized Controlled Trials Good For?" *Philosophical Studies* 147:59–70.

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy.* Oxford: Oxford University Press.

Elkin, Irene, et al. 1989. "National Institute of Mental Health Treatment of Depression Collaborative Research Program: General Effectiveness of Treatments." *Archives of General Psychiatry* 46:971–82.

https://doi.org/10.1086/683817 Published online by Cambridge University Press

Freedman, Laurence, and Arthur Schatzkin. 1992. "Sample Size for Studying Intermediate Endpoints within Intervention Trials or Observational Studies." *American Journal of Epidemiology* 136:1148–59.

Grünbaum, Adolf. 1986. "The Placebo Concept in Medicine and Psychiatry." *Psychological Medicine* 16:19–38.

Holmbeck, Grayson. 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.

Howick, Jeremy. 2012. *The Philosophy of Evidenced-Based Medicine.* West Sussex: British Medical Journal Books.

Kirsch, Iving. 2010. *The Emperor's New Drugs.* New York: Basic.

MacKinnon, David, Chondra Lockwood, Jeanne Hoffman, Stephen West, and Virgil Sheets. 2002. "A Comparison of Methods to Test Mediation and Other Intervening Variable Effects." *Psychological Methods* 7:83–104.

Moseley, J. Bruce, Kimberly O'Malley, Nancy Petersen, Terri Menke, Baruch Brody, David Kuykendall, John C. Hollingsworth, Carol M. Ashton, and Nelda Wray. 2002. "A Controlled Trial of Arthroscopic Surgery for Osteoarthritis of the Knee." *New England Journal of Medicine* 347:81–88.

Schulz, Kenneth, Douglas Altman, David Moher, and Dean Fergusson. 2010. "CONSORT 2010 Changes and Testing Blindness in RCTs." *Lancet* 375:1144–46.

https://doi.org/10.1086/683817 Published online by Cambridge University Press