

# Prevention & Treatment

*Prevention & Treatment*, Volume 5, Article 28, posted July 15, 2002  
[Copyright 2002 by the American Psychological Association](#)

---

Commentary on [The Emperor's New Drugs: An Analysis of Antidepressant Medication Data Submitted to the U.S. Food and Drug Administration](#)

## The Emperor's New Drugs: Effect Size and Moderation Effects

Steven D. Hollon  
Vanderbilt University

Robert J. DeRubeis  
University of Pennsylvania

Richard C. Shelton and Bahr Weiss  
Vanderbilt University

---

### ABSTRACT

In a review of efficacy data submitted to the U. S. Food and Drug Administration (FDA), [I. Kirsch, T. J. Moore, A. Scoboria, & S. S. Nicholls \(2002\)](#) found that mean posttreatment differences in symptom levels between drug and placebo were modest at best. This led the authors to suggest that either medication effects are trivial or that new designs are needed that do not assume additive effects. We suggest that not all patients necessarily respond to a given medication and that effect sizes based on the "average" patient may underestimate drug–placebo differences for those who do. Data submitted to the FDA can also underestimate how a drug will perform in clinic practice, as studies sometimes are designed as much for marketing purposes as they are to estimate the magnitude of a medication's effects. Finally, even when drug–placebo differences are small, antidepressant medication remains a potent treatment that typically matches or exceeds the efficacy of alternative interventions.

Key Words: medication, placebo, depression, moderation

---

Preparation of this commentary was supported by National Institute of Mental Health (NIMH) Grant MH55875 and Independent Scientist Award MH01697 to Steven D. Hollon and by NIMH Grant MH01741 to

Richard C. Shelton.

Correspondence concerning this article should be addressed to Steven D. Hollon, Department of Psychology, Vanderbilt University, 306 Wilson Hall, Nashville, Tennessee 37203.

E-mail: <http://journals.apa.org/prevention/volume5/steven.d.hollon@vanderbilt.edu>

---

In a recent quantitative review of efficacy data submitted to the U. S. Food and Drug Administration (FDA), [Kirsch, Moore, Scoboria, & Nicholls \(2002\)](#) found that approximately 80% of the total response to six of the most widely prescribed medications approved over the last decade was duplicated by pill–placebo controls. The mean difference between drug and placebo was approximately two points on the Hamilton Rating Scale for Depression (HAM-D), a very modest effect for samples that typically started treatment in the mid-20s and showed average reductions on the order of 8 to 10 points. Kirsch and his colleagues concluded that drug effects are so small that they could be due to nothing more than breaking of the blind, or the assumption of additive effects that underlies the use of pill–placebo controls is not correct, so that novel designs are needed to detect "true" drug effects. If the former, then what the field needs is a better placebo (at least one without side effects). If the latter, then we run the risk of dismissing potentially active medications through an undue reliance on placebo controls.

## Moderation and the Magnitude of Effect

We think there is a third option that [Kirsch et al. \(2002\)](#) did not consider, namely, that different patients respond to different medications and that "average" effects underestimate the benefits derived by those patients who do respond to a given medication. Many have long been unimpressed by the magnitude of the differences observed between treatments and controls, what some of our colleagues refer to as the "dirty little secret" in the pharmaceutical literature. Mean differences on the order of 2 to 4 points are the norm in many studies, and Kirsch and his colleagues have done a service by drawing attention to the magnitude of that effect. At the same time, categorical differences in the number of patients who respond to treatment typically have been considerably more impressive. For example, an early quantitative review conducted for the Agency for Health Care Policy and Research (AHCPR) found a response rate of 50% among patients treated with medications, relative to a rate of 30% among patients treated with placebo—a difference of 20% ([Depression Guideline Panel, 1993](#)). A subsequent update that covered many of the newer medications reviewed by Kirsch et al. found virtually identical rates of 50% versus 32% ([Mulrow et al., 1999](#)). A 2-point difference on the HAM-D seems relatively unimpressive, whereas a 20% increment in rates of response is considerably more so.

What we think may be going on is that not all patients are truly pharmacologically responsive to a given medication. Indices based on "average" effects presume that each member of a population receives an equal amount of benefit from each constituent component of the intervention. If only some members of a population benefit from a given component, then "average" effects that appear trivial could underestimate specific effects that are clinically meaningful for some groups of individuals. If that is the case, then differences between "active" medications and pill–placebo controls should be larger for more responsive patients and minimal or nonexistent for less responsive patients (an "average" effects model would predict constant differences of trivial magnitude across the whole range of patients).

For example, in an as yet unpublished multisite placebo-controlled trial involving paroxetine, we found that the least responsive patients within each condition hardly differed with respect to change on the HAM-D, whereas more responsive patients showed evidence of greater differential change. What we did was to rank order patients from least to most responsive within each treatment condition and compared HAM-D change between patients who were of comparable rank. Although mean differences between the two conditions were comparable to those described by Kirsch and his colleagues (2.33 points on the 17-item HAM-D), the greater the rank within condition, the greater the differential change between conditions. Drug-placebo differences by ascending quartile were 0.17 points on the HAM-D in the first quartile, 2.29 in the second, 3.50 in the third, and 3.39 in the fourth. Moreover, there were differences between the sites in the proportion of drug-responsive patients that further obscured differential patterns of individual response. When the respective conditions were ranked separately within site, only about a quarter of the ranked pairs showed a drug-placebo difference of two points or greater at one site, whereas that was true for over three quarters of the ranked pairs at the other. Over a third of the ranked patient pairs showed drug-placebo differences of four points or greater (37%), and a few showed differences of six points or more (9%). This suggests that there exists a subset of patients who derive considerable differential benefit from medications relative to placebo.

Summary statistics based on the arithmetic mean may be prone to being misinterpreted when there is variability in differential response. Under such circumstances, the shape of the distributions will be different for different conditions, and effect sizes based on categorical response (probability) may be more informative than those based on average response (magnitude). Our sense is that such moderation is likely a consequence of individual differences between patients; however, it could also reflect variation in treatment implementation. Regardless of its source, such moderation needs to be taken into consideration when determining whether a given medication has a "true" pharmacological effect.

We were unable to determine how often this occurs, since most studies do not publish distributions of individual response. Some studies do provide standard deviations for the respective conditions, which could be informative (the more skewed the distribution, the greater the standard deviation), but that presumes that placebo response is normally distributed. We did examine the standard deviations for a number of published trials and found that variances were often larger for patients treated with active medications, particularly in more heterogeneous samples. However, to address this question would require a systematic study of response distributions that goes beyond the scope of this commentary.

Although we do not want to go too far in interpreting the results of the single study just described, there are other indications that patients vary in their response to a given medication. Patients with atypical symptom patterns are more likely to respond to monoamine oxidase inhibitors than to other medications ([Liebowitz et al., 1988](#)), and gender predicts differential response to selective serotonin response inhibitors (SSRIs) versus tricyclic antidepressants (TCAs; [Kornstein et al., 2000](#)). Patients matched to the wrong medication can even do worse than when treated with placebo ([Stewart, Garfinkel, Nunes, & Klein, 1998](#)). Patients who do not respond to a given medication often respond to another, and augmenting with a second medication often helps enhance response to a first ([Thase & Rush, 1997](#)). These indications all point to moderation, and we think it is important to consider its possible effects. To the extent that different patients respond to different medications (in terms of showing a "true" drug effect), taking drugs can be likened to

playing the lottery; only some patients will receive specific benefit for pharmacological reasons, but most will receive some benefit from purely psychological effects. Identifying specific moderators can help improve initial choice of medications, and evidence of differential patterns of response can be used to guide the search for moderation. The bottom line is that effect sizes based on arithmetic means are fine when they fit the data, but they may be less informative than indices based on categorical response when moderation distorts the shapes of the respective distributions.

## FDA Studies and Clinical Practice

Similarly, it is not clear that studies submitted to the FDA necessarily provide the best basis for estimating actual drug effects. On the one hand, this collection of studies is relatively free from publication bias, since estimated effect sizes are unlikely to be inflated by the systematic exclusion of null findings. On the other hand, submitted studies sometimes are designed to serve other purposes beyond getting a medication approved. For example, when fluoxetine was first brought to market, its parent company was intent on demonstrating that the compound produced fewer side effects than did the rival tricyclics that dominated the market at that time. Dosage levels in those early trials were often constrained to minimize side effects while producing just enough response to establish efficacy. In other words, dosing strategies sometimes were driven more by marketing concerns than they were by a desire to maximize the medication's effect. The success of this strategy was evident; fluoxetine rapidly became the most widely prescribed antidepressant, not because it was the most effective agent but because it came to be seen as the least problematic ([Olfson & Klerman, 1993](#)).

There are other problems with the typical industry-funded trial. Many of these studies are conducted on a contractual basis that favors speed over accuracy. Subjects are not always carefully screened, and symptom scores at intake are sometimes inflated to fill sample quotas as rapidly as possible. Severity is known to be a particularly important moderator of drug effects; drug-placebo differences often are negligible among less severely depressed patients and larger among those who are more severely depressed ([Elkin et al., 1989](#)). Trials that pay "by the head" to deliver samples in an expeditious fashion may cut corners in ways that underestimate the true effects of the medications tested.

It also may be the case that some medications generate larger or more specific effects than do others. In their review, [Kirsch et al. \(2002\)](#) did not include estimates for three medications that failed to report mean scores for trials that had null findings. Excluding those medications was a reasonable thing to do (so as not to introduce bias due to missing data) but may have had the inadvertent consequence of excluding several of the more powerful current medications. Drugs that work on multiple neurotransmitter systems (e.g., venlafaxine and, possibly, paroxetine and sertraline) may be more effective than more purely serotonergic medications (e.g., fluoxetine) ([Entsuah, Huang, & Thase, 2001](#)). Of the medications reviewed by Kirsch and his colleagues, venlafaxine (which has both serotonergic and noradrenergic effects) is thought by some to produce the most robust effects and fluoxetine the least. Although we understand the authors' rationale for excluding medications, it is unfortunate that these included two of the more effective agents (paroxetine and possibly sertraline) and that this left only three medications on which to base conclusions (one of which was fluoxetine).

## Research Design and Policy Considerations

Finally, [Kirsch et al. \(2002\)](#) raise an interesting point in noting that placebo designs based on the assumption of additive effects may mask possible pharmacological effects. Different treatments may mobilize different mechanisms to varying degrees; to the extent that this is true, these mechanisms can be additive for only so long as they are truly independent and not constrained by "ceiling" effects. For example, if a given patient has the capacity to respond to either psychological or pharmacological mechanisms but can only get so much better, then the magnitude of the pharmacological effect detected may well be constrained by the quality of the psychological treatment. This may be why some patients (typically the most responsive) sometimes show a smaller "true" drug effect than do patients in the mid-range of response. Similarly, "warm" placebo controls that do a particularly good job of mobilizing expectations and providing support might produce greater change than do "cold" ones and, as a consequence, account for a greater proportion of a medication's effects ([Elkin, Pilkonis, Docherty, & Sotsky, 1988](#)). In this context, it is possible that placebo controls that do a more effective job of mobilizing psychological mechanisms may account for a larger proportion of a medication's effect without increasing its absolute clinical efficacy.

From an epistemological perspective, we like the notion advanced by [Kirsch et al. \(2002\)](#) of applying balanced placebo designs in the study of depression, as it might well detect times when psychological mechanisms mask pharmacological effects. Such designs cross the belief that one is receiving an active drug (yes vs. no) with the actual medication itself (active vs. inert) and have proven useful in exploring the pharmacological properties of substances like alcohol and caffeine. Nonetheless, we think it will be considerably more difficult to solve the pragmatic problems involved in implementing such designs for medications that require days or weeks to show a clinical effect than it was for faster acting agents like alcohol or caffeine. Moreover, it likely will prove difficult to mimic or mask anything other than anticholinergic side effects. But the idea is interesting nonetheless. We suspect that pharmacological and psychological mechanisms are not wholly independent (since each may mobilize the same final common pathways) and that, even if they were, they might compete for the same potential response. Combining drugs and psychotherapy rarely has produced increments in response as great as what would have been expected if their effects were truly independent ([Hollon & Shelton, 2001](#)).

We think that the FDA is a laudable institution that provides a valuable public service; safety generally has been well protected and each new medication must meet a minimum standard before it can be approved for sale. We would prefer a more stringent standard; requiring only two positive trials without considering the number of null findings is truly minimal but still better than what is required in most other countries. Given the market forces involved and ethical concerns regarding the use of placebo controls, we think the FDA has done well to maintain such standards as are currently in place (see [Leber, 1991](#)). These are the same standards that typically have been used to determine whether psychotherapy has empirical support ([Chambless & Hollon, 1998](#)). To date, only two psychotherapies have met even this minimal standard with respect to the treatment of depression ([DeRubeis & Crits-Christoph, 1998](#)). Null findings are few with respect to minimal treatment controls (compared with about a third of all placebo-controlled medication trials), but such comparisons have rarely been attempted in truly clinical populations ([Hollon & Shelton, 2001](#)).

Certain targeted psychotherapies like interpersonal psychotherapy (IPT) or cognitive behavior therapy (CBT) do compare favorably with medications in the treatment of all but the most severely depressed patients ([American Psychiatric Association, 2000](#)). Nonetheless, no psychotherapy has ever been shown to be superior to medications in the

reduction of acute distress when the latter was adequately implemented, and even the empirically supported psychotherapies sometimes have struggled to exceed the efficacy of pill–placebo controls in comparison to medications ([Hollon & Shelton, 2001](#)). There may be other reasons to prefer psychotherapy (either alone or in combination); IPT appears to have a greater breadth of effect in terms of enhancing relationship quality, and CBT may have a more enduring effect that reduces subsequent risk ([Hollon & Shelton, 2001](#)). However, no other intervention has been shown to be better than medications in the reduction of acute distress.

What this suggests is that medication treatment is no less potent than other clinical alternatives. Psychological mechanisms may account for the bulk of its effects (on average), but it is at the least a very effective way of mobilizing those mechanisms. If there were no "true" pharmacological effect, then ethical questions could be raised about exposing patients to side effects and potential adverse events in the service of mobilizing purely psychological mechanisms. But the existence of what is at least a modest pharmacological effect for most patients (and a larger effect for some) would appear to justify the use of medications, even if psychological mechanisms contribute greatly to their effects.

Finally, there are indications that the effects produced pharmacologically may be more stable over time than the psychological mechanisms mobilized by pill–placebos, at least so long as one continues taking medications. In a pair of studies, Quitkin and colleagues found that patients receiving active medication showed considerably less fluctuation in response over time than did patients treated with placebo ([Quitkin, Rabkin, Markowitz, Stewart, McGrath, & Harrison, 1987](#); [Quitkin, Rabkin, Ross, & Stewart, 1984](#)). Examination of the cumulative proportion of patients showing persistent improvement suggested that drug and placebo began to separate during the 2nd week of treatment and that differences continued to grow over the next several weeks. Pattern analysis based on delayed onset and persistent response better differentiated the patients treated with drugs from those treated with placebo than did absolute response, largely because this pattern was largely limited to only a subset of the medication-treated patients.

This same pattern replicated across several different medications. This suggests that although response to the pharmacological properties of a given medication may not be universal, when it occurs it tends to be more stable than response to the more purely psychological mechanisms that underlie placebo response. Since different patients appear to respond to different medications, it is likely that medication treatment (broadly defined) will be more effective than might be apparent from examining the results of specific drug–placebo comparisons. Switching medications (or augmenting with another) should increase the probability of response across different types of patients; in the context of moderation, effect sizes based on average drug–placebo differences for a given medication will underestimate the ultimate efficacy of medication treatment and likely misconstrue its stability as well.

## Conclusions

Given that not all patients respond to all medications, estimates of effects based on the "average" patient may underestimate drug–placebo differences for those who do respond. When such moderation occurs, estimates based on categorical response may be more informative than are estimates based on mean effects. Moreover, data submitted to the FDA are not necessarily representative of how a drug will perform in actual clinic practice.

Nonetheless, drug–placebo differences can seem surprisingly small, due in part to the magnitude of the impact of the psychological mechanisms mobilized by placebo controls. However, there are indications that the pharmacological effects of medication are more stable over time than are the psychological mechanisms involved in placebo response. Regardless of the relative magnitude of the mechanisms involved, medication remains a potent treatment that typically matches or exceeds the efficacy of alternative interventions.

## References

American Psychiatric Association. (2000). Practice guideline for the treatment of patients with major depressive disorder (revision). *American Journal of Psychiatry*, *157* (Suppl. 4), 1-45.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting & Clinical Psychology*, *66*, 7-18.

Depression Guideline Panel. (1993). *Depression in primary care: Vol. 2. Treatment of major depression* (Clinical Practice Guideline No. 5, AHCPR Publication No. 93-0551). Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research.

DeRubeis, R. J., & Crits-Christoph, P. (1998). Empirically supported individual and group psychological treatments for adult mental disorders. *Journal of Consulting and Clinical Psychology*, *66*, 37-52.

Elkin, I., Pilkonis, P. A., Docherty, J. P., & Sotsky, S. M. (1988). Conceptual and methodological issues in comparative studies of psychotherapy and pharmacotherapy: I. Active ingredients and mechanisms of change. *American Journal of Psychiatry*, *145*, 909-917.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., Glass, D. R., Pilkonis, P. A., Leber, W. R., Docherty, J. P., Fiester, S. J., & Parloff, M. B. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: General effectiveness of treatments. *Archives of General Psychiatry*, *46*, 971-982.

Entsuaeh, A. R., Huang, H., & Thase, M. E. (2001). Response and remission rates in different subpopulations with major depressive disorder administered venlafaxine, selective serotonin reuptake inhibitors, or placebo. *Journal of Clinical Psychiatry*, *62*, 869-877.

Hollon, S. D., & Shelton, R. C. (2001). Treatment guidelines for major depressive disorder. *Behavior Therapy*, *32*, 235-258.

Kirsch, I., Moore, T. J., Scoboria, A., & Nicholls, S. S. (2002). The emperor's new drugs: An analysis of antidepressant medication data submitted to the U.S. Food and Drug Administration. *Prevention & Treatment*, *5*, Article 23. Available on the World Wide Web: <http://www.journals.apa.org/prevention/volume5/pre0050023a.html>

Kornstein, S. G., Schatzberg, A. F., Thase, M. E., Yonkers, K. A., McCullough, J. P., Keitner, G. I., Gelenberg, A. J., Davis, S. M., Harrison, W. M., & Keller, M. B. (2000). Gender differences in treatment response to sertraline versus imipramine in chronic

depression. *American Journal of Psychiatry*, 157, 1445-1452.

Leber, P. (1991). Is there an alternative to the randomized controlled trial? *Psychopharmacology Bulletin*, 27, 3-7.

Liebowitz, M. R., Quitkin, F. M., Stewart, J. W., McGrath, P. J., Harrison, W. M., Markowitz, J. S., Rabkin, J. G., Tricamo, E., Goetz, D. M., & Klein, D. F. (1988). Antidepressant specificity in atypical depression. *Archives of General Psychiatry*, 45, 129-137.

Mulrow, C. D., Williams, J. W. Jr., Trivedi, M., Chiquette, E., Aguilar, C., Cornell, J. E., Badgett, R., Noel, P. H., Lawrence, V., Lee, S., Luther, M., Ramirez, G., Richardson, W. S., & Stamm, K. (1999). *Treatment of depression: Newer pharmacotherapies. Evidence Report/Technology Assessment No. 7* (AHCPR Publication No. 99-E014; prepared by the San Antonio Evidence-based Practice Center based at the University of Texas Health Science Center at San Antonio under Contract 290-97-0012). Rockville, MD: Agency for Health Care Policy and Research.

Olfson, M., & Klerman, G. L. (1993). Trends in the prescription of antidepressants by office-based psychiatrists. *American Journal of Psychiatry*, 150, 571-577.

Quitkin, F. M., Rabkin, J. D., Markowitz, J. M., Stewart, J. W., McGrath, P. J., & Harrison, W. (1987). Use of pattern analysis to identify true drug response: A replication. *Archives of General Psychiatry*, 44, 259-264.

Quitkin, F. M., Rabkin, J. D., Ross, D., & Stewart, J. W. (1984). Identification of true drug response to antidepressants: Use of pattern analysis. *Archives of General Psychiatry*, 41, 782-786.

Stewart, J. W., Garfinkel, R., Nunes, E. V., & Klein, D. F. (1998). Atypical depression and treatment response in the NIMH Treatment of Depression Collaborative Research Program. *Journal of Clinical Psychopharmacology*, 18, 429-434.

Thase, M. E., & Rush, A. J. (1997). When at first you don't succeed...sequential strategies for antidepressant nonresponders. *Journal of Clinical Psychiatry*, 58(Suppl. 13), 23-29.