When Positive Isn't Positive: The Hopes and Disappointments of Clinical Trials

Michael J. Ostacher, MD, MPH, MMSc

We are inclined in medicine to be enamored with positive results in clinical research, be they from large, well-designed clinical trials or from small case series. It is much more often heard that a negative result would somehow become positive if only the sample size were increased rather than that a positive result could become negative for the same reason. Unfortunately, it is more likely that a larger sample will lead an early positive result to become negative than the other way around, since movement toward the null and away from positive findings seems so inexorable in medicine that discussion of it has made its way into the popular press.[1] Nevertheless, we want our trials to be positive, we want treatments to offer to our patients, we want to be able to offer hope, and, perhaps most of all, we want to feel good about our field and ourselves.

The introductions to the reports of nearly all published clinical trials for bipolar depression emphasize the need for more validated and effective treatments, and this need has not changed in 2 decades. In spite of years of research, advances in the treatment of bipolar depression have been few; there are 2 established monotherapy treatments—quetiapine for depression in bipolar I and II disorder and lurasidone for depression in bipolar I—and a single combination treatment—olanzapine and fluoxetine combination for depression in bipolar I disorder. The other atypical antipsychotics have failed when studied; lamotrigine has not definitely proven itself as an acute treatment in multiple trials; and antidepressants . . . well, let's just leave it that their benefit in bipolar disorder has not been established.

All of this brings us to this ambitious and well-designed study by Calabrese et al[2] of the wakefulness drug armodafinil (the R-enantiomer of the indirect dopamine receptor agonist and dopamine reuptake inhibitor modafinil) in depressed subjects with bipolar I disorder who were already receiving maintenance medications for bipolar disorder.

Both modafinil and armodafinil were promising compounds for the treatment of bipolar depression. As small randomized trials of modafinil suggested a rapid and robust antidepressant effect for the drug in bipolar depression, it was hoped that the same effect would be found for the R-enantiomer, leading Teva Pharmaceuticals to initiate trials that might result in a US Food and Drug Administration (FDA)–labeled indication for armodafinil in bipolar disorder. While the trial published in this issue found a statistically significant but modest effect in favor of armodafinil 150 mg after the seventh week of an 8-week trial, the company's 2 other trials (currently unpublished) failed to find a statistically significant effect—and Teva has abandoned its efforts to gain FDA approval for the bipolar depression indication. How then, should we interpret the results of this single study with regard to our clinical practice?

Perhaps the biggest difficulty in interpreting studies is in understanding the relationship between the statistical power inherent in a given study and whether the results of that study can be trusted. We have come to rely on P values as the determination of whether treatments are effective or not: $P < .05$ and the result is positive; $P > .05$, it's not.[3] No matter how often statisticians publish about this oversimplification and its dangers, this sense of a magic threshold persists. It is the power of the study, however, rather than the P value itself that determines whether the finding is likely to be a true finding. That is to say, an underpowered study is likely to result in both a negative finding ($P > .05$) that is actually true (false-negative) and a positive finding ($P < .05$) that is actually false (false-positive). Most findings in neuroscience are likely to be false-positive findings, as Button et al[4] have pointed out, yet positive findings are the ones that are most likely to be published.

It is rare that the a priori power calculations for a study are reported in publications.[5] Instead, what tends to be reported are assumptions about power that were used to design the study and justify sample size—the predicted effect, its standard deviation, and the calculated number of subjects required to have 80% or 90% power—although in the case of this study, it is not reported at all. There is some controversy about the importance of post hoc power analyses, with some arguing that, for positive trials, they are not important at all, but the work of John Ioannidis[6] and others[4] has quite elegantly demonstrated that there is a strong relationship between lack of power and false-positive results. The post hoc power of this study, calculated from the reported results, appears to be somewhere between 40% and 68% and suggests that the study is, in fact, underpowered. This would mean that there is a high probability that the authors are falsely rejecting the null hypothesis, even when their P value is < .05.

P values do not measure the size of the effect, whether the effect is large enough to be of importance to patients and treatment providers.[3] For that, we rely on effect size or number needed to treat (NNT). The authors in this study calculated the effect size of the treatment and found a Cohen d therapeutic effect size of 0.28 in favor of modafinil 150 mg. This is, at best, a modest effect size, one that would

Submitted: August 4, 2014; accepted August 4, 2014.

Corresponding author: Michael J. Ostacher, MD, MPH, MMSC, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, 3801 Miranda Ave, Palo Alto, CA 94304 (ostacher@Stanford.edu).

J Clin Psychiatry 2014;75(10):e1186–e1187 (doi:10.4088/JCP.14com09427).
© Copyright 2014 Physicians Postgraduate Press, Inc.

generally be considered small. In spite of this, an effect size of 0.28 would be of clinical significance because of the lack of effective treatments for the problem at question, bipolar depression—if the results of the study could be trusted. Effect sizes and NNT are agnostic as to statistical significance, so, even if the calculated effect size were quite large, it would not be possible to know if that effect was merely due to chance without knowing whether the study itself was reliable enough to answer the question—and it is not clear that this study is reliable enough to answer the question. The small effect size coupled with 2 negative trials makes it difficult to accept that the statistical significance of this study alone is enough to have confidence in the results, especially if one ascribes to a Bayesian rather than a Frequentist approach to statistics in clinical trials.[7] The failure to replicate suggests that the reported effect size of this positive trial may be overestimated and that the actual effect size is even smaller than 0.28.

So where does this leave us? We have an ostensibly positive clinical trial of modafinil for major depressive episodes in bipolar disorder, one with a relatively small effect size but of a relatively safe drug in a therapeutic area in desperate need of additional treatments. On its own, perhaps, the study could be cause for optimism that another treatment for bipolar depression has been found, even if the benefit that was found came unexpectedly late in the course of the study of a drug expected to have rapid onset of effect. Realistically, however, this is a finding that was not replicated in 2 additional trials, and replication is the standard by which we should most rigorously judge our treatments. One can reply that studies are difficult to do (they are), that many methodological factors make it difficult for efficacy trials to find an effect (which is true), that the inclusion of maintenance treatments in this study blunted the effect of a likely effective treatment (impossible to know and, in any case, representative of how it would likely be deployed clinically), and that therefore a positive trial should be accepted as evidence of benefit. I'm not so sure. There are times when we have to accept that our hopes for a drug do not hold up to scientific scrutiny, and, for this drug, this is perhaps one of those times, especially when considering that we are exposing our patients to at least some risk if we prescribe it to them. Many people already use armodafinil and modafinil off-label for depression and consider it safe and effective as an additional tool in the mood disorders toolbox, so this opinion may not be widely shared. Yet, it is essential that we not insist that a benefit exists because it is too disappointing not to accept it. We must continue to pursue research in new treatments (whatever the financial and professional risks), to remember that psychosocial interventions (such as exercise, psychotherapy, family treatment, and addiction treatment) must be considered an essential part of psychiatric treatment and be integrated into care, and to always remain present for our patients and their suffering, even when the options for pharmacologic treatment remain few.

Drug names: armodafinil (Nuvigil and others), fluoxetine (Prozac and others), lamotrigine (Lamictal and others), lurasidone (Latuda), modafinil (Provigil and others), olanzapine (Zyprexa and others), quetiapine (Seroquel and others).

Author affiliations: VA Palo Alto Health Care System, Palo Alto, California and Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, California.

Potential conflicts of interest: None reported.
Funding/support: None reported.

REFERENCES

1. Lehrer J. The truth wears off: is there something wrong with the scientific method? *The New Yorker*. December 13, 2010. http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off. Accessed August 8, 2014.
2. Calabrese JR, Frye MA, Yang R, et al. Efficacy and safety of adjunctive armodafinil in adults with major depressive episodes associated with bipolar I disorder: a randomized, double-blind, placebo-controlled, multicenter trial. *J Clin Psychiatry.* 2014;75(10):1054–1061.
3. Goodman SN. Toward evidence-based medical statistics, 1: the *P* value fallacy. *Ann Intern Med.* 1999;130(12):995–1004.
4. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci.* 2013;14(5):365–376.
5. Chan AW, Hróbjartsson A, Jørgensen KJ, et al. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *BMJ*. 2008;337:a2299.
6. Ioannidis JP. Why most published research findings are false. *PLoS Med.* 2005;2(8):e124.
7. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. West Sussex, England: John Wiley & Sons; 2004:221–223.