See article beginning on page 1344.

Another Point of View: Superiority, Noninferiority, and the Role of Active Comparators

Helena Chmura Kraemer, PhD

ABSTRACT

Despite substantial agreement with points made by Andrew C. Leon, PhD, in his article, I am not in complete agreement in a few areas. The definition of *noninferiority* proposed by Leon allows drugs somewhat less effective than placebo to be characterized as noninferior to placebo, and 2 active drugs may each be simultaneously noninferior to the other. Moreover, including a placebo arm in comparing 2 active drugs is of no use in deciding whether the study is well designed or not, since a significant difference between one of the active arms and the placebo may be due to chance or to a bias in the design. An alternative view of the situation is presented.

J Clin Psychiatry 2011;72(10):1350–1352

Submitted: *September 30, 2010, accepted September 30, 2010 (doi:10.4088/JCP.10com06607whi).*

Corresponding author: *Helena Chmura Kraemer, PhD, 1116 Forest Ave, Palo Alto, CA 94301 (hckhome@pacbell.net).*

The topics addressed by Andrew C. Leon, PhD,[1] are crucially important to clinical decision-making, not only in psychiatry but in all fields of medicine (particularly commercial drug development and US Food and Drug Administration decision-making).[2] On the major issues, Leon and I strongly concur: the advocacy for comparative effectiveness clinical trials, the importance accorded active comparators so necessary to evidence-based medicine, and the emphasis on effect size (ES) rather than *P* values. However, there also remain some areas of disagreement and confusing issues.

Basic to the discussion is ES, a measure comparing the clinical impact on the patients in the population sampled, of the investigational drug (*I*) versus a control/comparison treatment (*C*), which may be placebo (*P*) or an active comparator (*A*). Leon uses standardized mean difference as the ES. Such a population ES is never known exactly, but it is estimated in randomized controlled trials (RCTs) with a certain margin of error conveyed by its confidence interval.

An ES would be zero if there were absolutely no difference between *I* and *C*, but such an occurrence is only theoretically possible, since by the time there are a theoretical rationale and an empirical justification for an ethical RCT comparing *I* versus *C*, there is little chance that the difference between them will be *absolutely* zero.[3,4] However, the *I* versus *C* difference may well be too small to warrant any clinical preference for one intervention over the other, in which case *I* and *C* are *clinically equivalent*. Only if the ES is greater than some value d* is a strong clinical recommendation of one treatment over the other warranted. The top rows of Figure 1 represent traditional views of clinical superiority, inferiority, and equivalence; the bottom 2 rows represent Leon's understanding of inferiority and *noninferiority*.

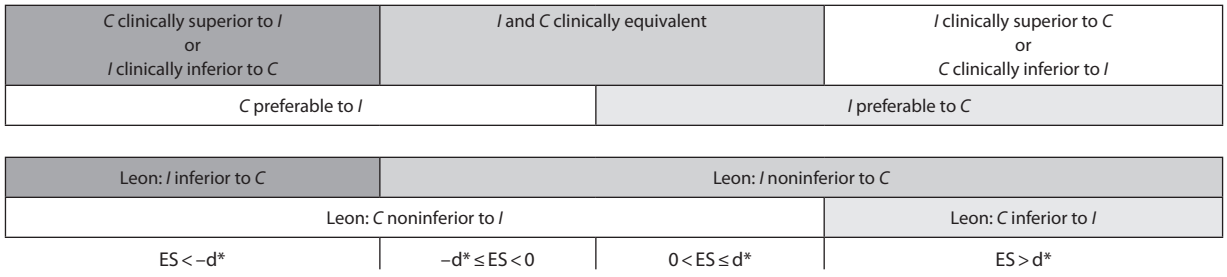
The traditional *valid* 2-tailed hypothesis test at the 5% level of significance comparing *I* versus *C* typically requires that the chance of a statistically significant result, if ES were indeed 0, be less than 5%. An *adequately powered* valid hypothesis test at the 5% level of significance also requires that the chance of a statistically significant result be greater than, for example, 80%, whenever the ES is greater than d*. The parameter d* is typically called the *critical value* or the *threshold of clinical significance*.[5,6] Leon calls his d* the *threshold of clinical meaningfulness*, which seems awkward, because +d* and –d* each lie within a noninferiority region (Figure 1).

Thus, on the traditional view, either *I* is preferable to *C*, or *C* is preferable to *I*, but not both. If *I* is preferable to *C*, *I* is either superior to or equivalent to *C*, and if *C* is preferable to *I*, *C* is either superior to or equivalent to *I*. In Leon's view, if *P* is preferable to *I*, but the difference is not enough to be clinically significant ($-d^* < ES < 0$), *I* is considered *noninferior* to *P*. If *I* were approved on the basis of its noninferiority to *P*, drugs less effective than placebo might be approved. Also, in the comparison of 2 active drugs, *I* and *A*, *I* might be found noninferior to *A*, and *A* noninferior to *I*, perhaps even in the same data set. This situation is bound to confuse.

However, the issue of whether to use the traditional 2-tailed hypothesis test at the 5% level of significance or the noninferiority test becomes irrelevant, if we agree that the goal of an RCT is not to show statistical significance but to estimate the ES comparing *I* and *C*.[7]

© Copyright 2011 Physicians Postgraduate Press, Inc.

Figure 1. Relationship of Investigational Drug (*I*) and Control/Comparison Treatment (*C*) Based on the Effect Size (ES), With a Critical Value d*



Symbol: d* = threshold of clinical significance.

Leon is correct in pointing out how deficient we have been in setting the value of d*, which depends on variables such as the seriousness of the indication, the danger of the consequences of inadequate treatment, and the costs and risks of the treatments. This deficiency has long been a problem, resulting in a proliferation of underpowered and often misleading RCTs.[8,9] Generally, if *C* is a placebo, d* would be set nearer .8, while if *C* is an active comparator, d* would be set nearer .2.[5,6] Since the sample size necessary for adequate power increases as d* decreases, the sample size for an *I* versus *A* comparison must typically be much larger than that for an *I* (or *A*) versus *P* difference. Thus, the adequacy of the design to detect an *I* (or *A*) versus *P* difference is no indication of the adequacy of the design to detect a difference in an *I* versus *A* comparison. Moreover, one cannot interpret finding any statistically significant result as proof of the adequacy of the design. Such a result may well arise by chance or because of design bias. The logic underlying the concept of *assay sensitivity* is flawed. But should one nevertheless include *P* in any comparison of *I* versus *A* for another reason?

The ethical principle of *clinical equipoise*[10] precludes proposing an RCT involving patients (a) in the absence of a theoretical rationale and empirical justification for the hypothesis to be tested or (b) after the answer is already scientifically known. The first criterion seems reasonably well-understood, but what does *scientifically known* mean? Surprisingly, researchers tend not to believe the results of their own studies; every study ends with an appeal for yet more studies, with no end in sight. With the exceptions noted below, such additional studies do not clarify comparisons of clinical effectiveness.

A proposal: if a meta-analysis of all existing valid RCTs comparing *I* versus *C* in a population results in a 95% confidence interval for an outcome in which the ES lies completely above d*, *I is known to be clinically superior to C*; if that confidence interval lies completely below –d*,[10] *C is known to be clinically superior to I*; if that confidence interval lies completely between –d* and +d*, *I and C are known to be clinically equivalent*. Any other result would indicate that more studies are needed.

The answer to the question of how *I* and *C* compare in terms of clinical effectiveness, then, is virtually never known on the basis of a single RCT, and it is seldom known on the basis of fewer than 2 or 3 RCTs.[11] However, it would seldom require more that 3–5 adequately powered, valid RCTs to know the answer. Such a meta-analysis is the basic principle underlying the Cochrane Collaboration approach to evidence-based medicine. As Leon points out, a single RCT with a nonstatistically significant result seldom proves equivalence, nor do 1 or 2 RCTs with statistically significant results ($P < .05$) necessarily establish clinical superiority.

So, to gather these ideas together to envision the process of clarifying the role of an *I* for clinical decision-making:

- The earliest RCTs should be (as they are now) efficacy RCTs, with a *P*, in the population most likely to respond to *I*, to establish that *I* is clinically superior to *P* in *some* population. The rationale and justification for such RCTs stem primarily from Phase 1 and 2 studies, and from translational research. This approach is both ethical and logical, for if *I* is not clinically superior to an intervention that essentially does *nothing* (*P*) in the population in which *I* is likely to have its *greatest* effect, it makes no sense to invest the time and resources to further develop it or to impose an unnecessary burden of multiple RCTs on patients.
- Once it is known that *I* is clinically superior to *P* in the most favorable circumstances, effectiveness RCTs should follow, with a *P*, sampling the population with the targeted indication. The primary goal is to establish the ES in this target population, but of major concern also are moderators of treatment,[12–14] ie, identification of subpopulations in which *I* is clinically superior to *P*. It is quite possible that *I* is clinically superior to *P* in some subpopulation, and either equivalent to or less effective (harmful) than *P* in others, which is the concern of personalized medicine.[15–18] Thus, if *I* is effective and safe only for those with a certain genotype, in a certain age range, in the absence of certain comorbidities, at this stage those limits should be established and clarified for medical consumers.

- Once it is known that *I* is clinically superior to *P* in a specific subpopulation of those with the targeted indication, the question is whether there are other *A*'s also known to be clinically superior to *P* in that subpopulation. If so, one would seek to find in which subpopulations *I* is clinically superior to *A*, in which subpopulations *A* is clinically superior to *I*, and in which subpopulations *A* and *I* are clinically equivalent. (All 3 different subpopulations might exist.)

A and *I* should be dealt with symmetrically. One company's *I* is another company's *A*, and neither should get preferential treatment. Until it is known that both *I* and *A* (active interventions) are clinically preferable to *P* in the same subpopulation, it makes no sense to compare *I* versus *A* in an RCT in any population. As soon as it is known that both are clinically preferable to *P* in a subpopulation, including a placebo control is unethical. Thus, *P* should *not* be included in a study comparing 2 active treatments, *A* and *I*.

Currently, the US Food and Drug Administration (and therefore drug companies) emphasizes *P* values rather than ES, and it puts little emphasis on how representative the sample is of the population to which the results may be applied. There has been little attention to moderators of treatment response (the concern of personalized medicine). The emphasis is often on what can be shown in *selected* individual studies, not on the *cumulative* results of all valid RCTs done to date on a particular question. All of these approaches have a negative impact on the quality of medical decision-making, and they should be reconsidered.

Author affiliations: The Department of Psychiatry and Behavioral Sciences (Emerita), Stanford University School of Medicine; Palo Alto, California, and the Department of Psychiatry, University of Pittsburgh School of Medicine, Pennsylvania.
Potential conflicts of interest: None reported.
Funding/support: None reported.

REFERENCES

1. Leon AC. Comparative effectiveness clinical trials in psychiatry: superiority, noninferiority and the role of active comparators. *J Clin Psychiatry*. 2011;72(10):1344–1349
2. US Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research. Center for Biologics Evaluation and Research. *Guidance for Industry Non-Inferiority Clinical Trials (draft guidance)*. Rockville, MD: Food and Drug Administration; 2010. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf. Accessed November 9, 2010.
3. Jones LV, Tukey JW. A sensible formulation of the significance test. *Psychol Methods*. 2000;5(4):411–414.
4. Meehl PE. Theory testing in psychology and physics: a methodological paradox. *Philos Sci*. 1967;34(2):103–115.
5. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
6. Kraemer HC, Thiemann S. *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA: Sage Publications; 1987.
7. Altman DG, Schulz KF, Moher D, et al; CONSORT GROUP (Consolidated Standards of Reporting Trials). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med*. 2001;134(8):663–694.
8. Maxwell SE. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychol Methods*. 2004; 9(2):147–163.
9. Kazdin AE, Bass D. Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *J Consult Clin Psychol*. 1989;57(1):138–147.
10. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med*. 1987;317(3):141–145.
11. Lau J, Antman EM, Jimenez-Silva J, et al. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med*. 1992; 327(4):248–254.
12. Kraemer HC, Frank E, Kupfer DJ. Moderators of treatment outcomes: clinical, research, and policy importance. *JAMA*. 2006;296(10):1286–1289.
13. Arnold LE, Farmer C, Kraemer HC, et al. Moderators, mediators, and other predictors of risperidone response in children with autistic disorder and irritability. *J Child Adolesc Psychopharmacol*. 2009;20(2):83–93.
14. Kraemer HC, Wilson GT, Fairburn CG, et al. Mediators and moderators of treatment effects in randomized clinical trials. *Arch Gen Psychiatry*. 2002;59(10):877–883.
15. Garber AM, Tunis SR. Does comparative-effectiveness research threaten personalized medicine? *N Engl J Med*. 2009;360(19):1925–1927.
16. Richmond TD. The current status and future potential of personalized diagnostics: Streamlining a customized process. *Biotechnol Annu Rev*. 2008;14:411–422.
17. Lesko LJ. Personalized medicine: elusive dream or imminent reality? *Clin Pharmacol Ther*. 2007;81(6):807–816.
18. Abrahams E, Ginsburg GS, Silver M. The personalized medicine coalition: goals and strategies. *Am J Pharmacogenomics*. 2005;5(6):345–355.

© Copyright 2011 Physicians Postgraduate Press, Inc.