Psychotherapy
2011, Vol. 48, No. 3, 223–224

0033-3204/11/$12.00 DOI: 10.1037/a0023198

COMMENTARY

Some More Fundamental Problems in Clinical Research: Comment on "Statistical Significance Testing and Clinical Trials"

Stefan G. Hofmann
Boston University

The article "Statistical significance testing and clinical trials" by Krause (this issue, pp. 217–222) provides a thought-provoking and critical discussion of the conventional statistical testing in clinical research. The author argues that, by focusing exclusively on mean differences between groups and their statistical significance, important information about the individual participant is being ignored. Krause calls for a different methodology that examines client covariates in relation to the outcome and then compares the treatment outcome distributions and their overlaps for each of the covariate-defined subgroups. The problem is well described, and the possible solutions well articulated. At the same time, however, the problem the author is tackling stays at the initial stage of a multilevel problem.

One of the central issues of Krause's (this issue) argument relates to client characteristics. Clinical researchers typically deal with populations that are defined by a medical classification system that categorizes people with a different history, course of illness, and etiology, as well as cultural and social features, into the same diagnostic group that is defined on the basis of more-or-less arbitrary symptom patterns. The resulting diagnostic groups show, not surprisingly, a considerable degree of heterogeneity. Despite these challenges imposed by the diagnostic system, psychologists have been extraordinarily successful in developing a number of effective interventions. Krause's call for examining client characteristics (i.e., the moderators) in treatment research is well taken. At the same time, it is further important to examine the treatment mechanism through which these interventions work (i.e., the mediators) and to further explore whether certain mediators are specific to certain subgroups and whether certain moderators are linked to one specific mechanism (i.e., moderated mediation and mediated moderation). These are complex issues that future generations of clinical researchers will have to tackle and that will likely take away from the conventional nosological system.

Krause (this issue) further notes that it is difficult to publish statistically nonsignificant findings, even if the trial was well conducted. This so-called file-drawer problem has long been recognized by methodologists (Rosenthal, 1979). Meta-analyses include the fail-safe *N* statistic and funnel-plot method to quantify the problem. Nevertheless, it is considerably more difficult to publish nonsignificant findings from adequately powered and well-designed studies than significant findings from flawed trials. The disregard of the Type II error rate runs deep among clinical researchers. Effective strategies will require changes in editorial policies, reviewers' judgments, and guidelines of funding agencies. The requirement for National Institutes of Health (NIH)–funded investigators to register clinical trials prior to ending a trial is an active attempt to work against the publication bias in clinical research.

The error rate problem is part of a larger, more fundamental problem, which was only briefly mentioned by Krause, (this issue) namely, the basic meaningfulness of the Null Hypothesis Significance Testing (NHST). At various places, the author touches upon the NHST but stops short of critically evaluating the basic logic of it. Specifically, it has been argued that the NHST is based on a misapplication of deductive syllogistic reasoning because probabilistic statements are incompatible with the rules of deductive reasoning (Cohen, 1994). Briefly, *deductive* and *inductive* statements are two different types of syllogistic arguments. In the case of an inductive argument, its conclusion is likely (but not necessarily) true if its premises are true. In contrast, the conclusion of a deductive argument has to be true if the premises are true. This characteristic of a deductive argument is known as *formal validity*. Philosophers distinguish a number of different rules through which a valid conclusion can be derived (deduced) from its premises. One of these rules is the modus tollens (denying the consequence), which is the argument that NHST is based on: As I had discussed earlier (Hofmann, 2002; see also Krueger, 2001), NHST is based on the following argument:

P1: If P (H_0 is true), then Q (there is no difference between the groups).
P2: Not Q (there is a difference between the groups).
C: Not P (H_0 is not true).

The problem arises when we are dealing with probabilistic premises because the probability statements can lead to false conclusions. Such an example is provided by Cohen (1994):

P1: If a person is an American, then he is probably not a member of Congress.
P2: The person is a member of congress.
C: Therefore, he is probably not an American.

This example illustrates that NHST, which is based on probability statements, can easily lead to wrong conclusions. Because of its probabilistic nature, it is possible that the conclusion of an argument is false although all of the premises of the argument are true. NHST is such a case.

Correspondence concerning this article should be addressed to Stefan G. Hofmann, Department of Psychology, Boston University, 648 Beacon Street, 6th Floor, Boston, MA 02215. E-mail: shofmann@bu.edu

To add another level of complexity to this problem, we do not test H_1 directly, but assume it to be true if H_0 is likely to be untrue. This "there-is-probably not-nothing-method" of NHST has been criticized by a number of authors before, encouraging authors to base their conclusions on effect size estimates rather than on the arbitrary $p < .05$ level of statistical significance (e.g., Meehl, 1978; Rosenthal, 19955). However, effect size estimations and attempts to quantify the Types I and II error rates do not solve the fundamental problem of the logic associated with NHST. Unfortunately, there is no easy solution to this problem because, without a good alternative, it is difficult and unwise to completely abandon NHST. Single-subject methodologies could provide a useful alternative for clinical researchers.

In summary, Krause (this issue) provides a thoughtful discussion of some problems of clinical research. However, a number of more fundamental issues that underlie these problems were insufficiently explored. Solutions to the issues raised in this commentary are substantial and may cause us to reconsider selecting participants on the basis of a symptom-based nosological system, using RCT methodology as opposed to single-subject designs as the primary method of doing treatment research, placing an overemphasis on outcome variables over treatment mechanism and moderator research, and the uncritical acceptance of the flawed logic of NHST.

References

Cohen, J. (1994). The earth is round, $p < .05$. *American Psychologist, 49,* 997–1003.

Hofmann, S. G. (2002). Fisher's Fallacy and NHST's flawed logic (letter). *American Psychologist, 57,* 69–70.

Krause, M. S. (2011). Statistical significance testing and clinical trials. *Psychotherapy, 48,* 217–222.

Krueger, J. (2001). Null hypothesis significance testing. On the survival of a flawed method. *American Psychologist, 56,* 16–26.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

Rosenthal, R. (1995). Progress in clinical psychology: Is there any? *Clinical Psychology: Science and Practice, 2,* 133–150.

Received November 23, 2010
Accepted November 24, 2010 ■