Psychotherapy
2011, Vol. 48, No. 3, 225–228

0033-3204/11/$12.00 DOI: 10.1037/a0022701

COMMENTARY

Statistical Significance Testing and Clinical Effectiveness Studies

Edward A. Wise
Mental Health Resources, Memphis, Tennessee

Effectiveness studies, by definition, must reflect patient and treatment variables that exist in the real world of practice. This includes identifying the covariates that contribute to both positive and negative outcomes. The use of clinically significant change and reliable change indices is reviewed to demonstrate that patient changes can be made on the basis of normative comparisons, using outcome relevant variables that reflect the diversity of problems that are of concern to our clients. Therefore, neither agreement on a single outcome index nor plotting the entire outcome frequency distributions of the comparison group, as called for by Krause, is required. Various methods currently exist and are used to address the limitations Krause cites with respect to randomized trials, outcome relevant covariates, file-drawer effects, and so forth. The identification and prediction of an individual's treatment response can be done in treatment as usual studies, where outcome relevant covariates have not been controlled, by studying change at the individual level. By studying such individuals and their combined continuums of treatment response, we can then identify clinically relevant outcome variables and alter the course of treatment accordingly.

Keywords: clinical significance, reliable change, effectiveness, outcome methods

Although Dr. Krause (this issue, pp. 217–222) primarily focuses on randomized clinical trials (RCTs), he includes "effectiveness in ordinary practice" (this issue, p. 217), which is highly relevant to practitioners. Dr. Krause is correct in his assertions that comparing differences in mean outcomes between groups does not take into account individual change. He is also correct in stating that it would be most helpful if we could predict who are "the types of clients who obtain . . . unsatisfactory outcomes" (this issue, p. 217). In fact, these statistical observations are true irrespective of research designs (e.g., RCTs, quasi-experimental prepost group with contrasts, regression discontinuity). However, it remains an empirical question as to whether "the most informative comparison of two different treatments' . . . effectiveness . . . is the comparison of their whole outcome distributions" (this issue, p. 217). Although the sample distributions provided by Dr. Krause are interesting in theory, the extent to which they reflect distributions in practice is an empirical matter, also awaiting investigation. Hence, Krause's paper represents a theoretical treatise that does not necessarily reflect the realities of practice. As another example, Krause stated that "there is no mathematically necessary relation between a distribution's (greater than zero) range and its variance" (this issue, p. 219). Similarly, he demonstrated in the Tables and Figures 3 and 4 that "it is mathematically possible for nonoverlapping outcome distributions to have statistically *insignificant* different mean outcomes and for overlapping outcome distributions to have statistically *significant* different mean outcomes" (this issue, p. 219, emphasis added). Although these scenarios are theoretically possible, no real-world clinical data were provided that reflect such outcomes, and these scenarios would not occur under the assumptions of normality.

In contrast, Shimokawa, Lambert, and Smart (2010) remind us that 5% - 10% of adults treated in RCTs and effectiveness studies will likely deteriorate with treatment, irrespective of what treatment is delivered, under what conditions (RCTs vs. routine practice), or patient characteristics (Hansen, Lambert & Forman, 2002). However, it is not essential to know the entire outcome distribution to address the problem of patients who "obtain . . . unsatisfactory outcomes" (Krause, this issue, p. 217). Howard, Moras, Brill, Martinovich, and Lutz (1996) were the first to call this "patient-focused research" (p. 1059). However, a normative continuum is needed to make comparisons.

Jacobson, Follette, and Revenstorf (1984) were interested in clinically significant (CS) change at the individual level, which they defined as a client moving "from the dysfunctional to the functional range during the course of therapy" (p. 340). They went on to operationalize the dysfunctional to functional continuum and Jacobson and Truax (1991) further quantified these efforts by providing statistical methods that could be applied to any variable of clinical interest in any setting. Although there are several variations on the initial formulas, the important concept is that in order for a client to have made CS change, they must have moved from the dysfunctional range toward the functional range. A traditional example is based on symptom improvement, where a patient's score can be obtained at the beginning and end of treatment and then compared with a normative database. However, depression scores, global distress indices, client satisfaction, step down to lower levels of care, functional improvements, and number of days of abstinence from substance abuse have been used as CS variables (Wise, 2003, 2004, 2010). For example, if a depres-

Correspondence concerning this article should be addressed to Edward A. Wise, Mental Health Resources, PLLC, 1037 Cresthaven Road, Memphis, TN 38119. E-mail: edwisephd@gmail.com

sion rating scale were used for which normative data were available, pre- and post-treatment scores could be used to determine the client's location on the normative continuum (CS) compared with a relevant group (e.g., community sample, outpatients, inpatients).

To determine whether these CS changes are statistically reliable, and not simply the result of measurement error, Jacobson and Truax (1991) developed the Reliable Change Index (RCI). RCIs are calculated by subtracting the post-treatment score from the pretreatment score and dividing by the standard error of estimate. (The Jacobson and Truax (1991) RCI formula to determine statistically reliable change is: $RCI = X_1 - X_2/SE$, where $X1$ = pretest score; $X2$ = posttest score; $S1 = {}_{s1}\sqrt{(1\text{-}r_{xx})}$; $s1$ = the standard deviation [*SD*] of control group, normal population, or pretreatment group; and r_{xx} = the test–retest reliability. The reader is referred to Wise, 2004, for more detailed formula descriptions and examples of CS and RCI.) If a client's score is CS and the RCI 1.96, reliable change has occurred and the patient would be considered to have recovered or improved. In its simplest form, an RCI for a depression scale score that was ≥1.96 and moved two *SD*s toward the mean, or whose score moved into the functional or normative sample, would be considered both statistically reliable and CS. Using these criteria, one could then classify each individual "in a treatment outcome study as *Recovered* (passed both CS normative and RCI criteria), *Improved* (passed RCI criteria alone), *Unchanged/Indeterminate* (passed neither), or *Deteriorated* (passed RCI in the negative direction)" (Wise, p. 52, 2004). Hence, RCIs allow us to not only identify those who have recovered, but also to identify and examine the "improved," "deteriorated," "unchanged," or "indeterminate" patient(s), on the basis of the normative continuum, and these patients could be further studied. Note, however, that the RCI cut-off points could also be changed, for example from 1.96 (confidence interval [CI] = 95%) to 1.28 or .84 (90% and 80% CI, respectively), to capture patients who are likely to have changed, but not at the 95% CI. These patients might be identified as "on their way" to recovery.

These individual RCI scores can be compared with a known normative group (using means and *SD*s) and supplemented with any quantifiable variable of clinical interest to further assess CS change to determine individual recovery, improvement, no change, and deterioration rates. In this way, knowing individual patients' change score(s) allows one to determine their relative location as compared with the normative continuum.

In fact, these recovery categories can be refined in numerous ways. For example, in the aforementioned dual diagnosis effectiveness study (Wise, 2010), the sample presented with comorbid psychiatric symptomatology comparable with psychiatric inpatients and was using substances an average of 5 days per week. Because the combination of these outcome variables was of great interest, the Global Severity Index (GSI) from the Symptom Checklist-90-R (Derogatis, 1994) was used as the primary RCI variable to determine statistically significant change in psychiatric distress and this was combined with Percentage of Days Abstinent (PDA) from alcohol or drugs as the CS variable because of its outcome relevance. Thus, just as there is flexibility in the RCI criteria, the CS variables may also be selected to better reflect clinical realities. It should be noted that the treatment cited used a harm reduction model and that abstinence per se was not an outcome, but that quantity and frequency of reduced substance use was a treatment goal. The data analysis showed that by using an 80% CI and one-day less of substance use, 56% of the sample achieved statistically reliable significant improvements with respect to the combined variables of symptom severity and PDA (Recovered). In contrast, 23% did not change PDA, but 73% of these showed reliable symptom improvements (Improved). Another 21% showed increased PDA, but none of these significantly reduced the severity of symptoms (Unchanged). Only 2% increased use and RCI criteria (Deteriorated). Although these measures portray somewhat liberal estimates of CS, it should also be noted that at the completion of treatment, 53% had maintained 100% abstinence and 16% reported having used only one day in the previous week. Of these two latter groups, 79% also achieved reliable GSI improvements. The RCI methodology provided for a clinically meaningful assessment of the treatment effects on the combined outcome variables of psychiatric symptom severity index (GSI) and PDA at the individual level. The interaction of psychiatric symptoms and PDA is very complex and there are multiple strategies to measure the outcomes of these comorbid conditions.

In effect, using RCIs and CS change scores provides a dimensional framework for each individual patient, ensuring that change scores are statistically reliable and CS. Furthermore, the aforementioned traditional recovery groups can be further subdivided in many clinically relevant ways, as indicated earlier. In fact, Jacobson and Revenstorf (1988) advocated using multiple measures and "accept[ing] the fact that *no single index* of CS change will capture all components of the disorder under study" (emphasis added, p. 140). Additionally, Wise (2004) elaborated that in some cases, such as in depressed inpatients:

> . . . comparatively few patients actually achieve a full remission of depressive symptoms, and [since] functional capacities are the last improvements, the return to normal criterion appears unrealistic . . .. In addition to varying the number of *SD* units required to demonstrate normative CS change, a symptom scale might also be used by applying varying confidence levels to the RCI formula (e.g., 1.96, 1.28, and .84, corresponding to 95%, 90%, and 80% confidence levels, respectively). Such varying RCI confidence levels could be used in conjunction with real-world measures . . . (p. 55).

In the aforementioned dual diagnosis study, a subgroup of substance abuse, depressed patients were identified who appeared not to improve, only to discover that the majority of this subgroup had entered treatment abstinent in an effort to prevent relapse and were successful in doing so (Wise, 2010). So, although it is true that "The only statistical *concept* that is specifically informative about every individual case in a group is the *distribution* of outcome values for *all* the members in the group" (Krause, this issue, p. 219; emphasis added), it is not necessary to know every individual's score in the normative comparison group. Additionally, to state that "*the only* statistic that RCTs can provide that is relevant for clinical (or, indeed any truly psychological) purposes" is the "outcome distributions" (Krause, this issue, p. 220; emphasis added) is rather extreme. Plotting frequency distributions would not provide the kind of clinically meaningful information that CS scores convey and would be extremely labor-intensive. In most practice settings, and virtually all independent practice settings, obtaining a control group or a second treatment option are huge

obstacles in themselves, not to mention the added burden of plotting all the outcome distributions of the comparison patients.

In Footnote 5, Krause states that given "the limited information [from two outcome distributions] the *t* test . . . is not appropriate for clinical purposes" (this issue, p. 220). If there are reasons to believe that data do not meet the assumptions of normality, this would be true, and in this case, a nonparametric test would be appropriate. However, Footnote 5 goes even farther than declaring the *t* test inappropriate for the comparisons of treatment groups:

> It is also important to appreciate that so long as a variety of different outcome variables whose functional interrelations remain unknown are used in the meta-analysis of RCTs that include the same pair of treatments, we shall never be able to adequately compare the results of these RCTs and integrate these in a single meaningful estimate of the treatments' relative efficacy. Therefore, it is crucially important that the psychotherapy research community finally settle on a single measure of mental health status, difficult as this may be to accomplish (p. 220).

"Different outcome variables", such as symptom severity and functional impairment will vary by patient, but this has not impeded the use of meta-analyses in the comparisons of RCTs. Note, however, that Krause is proclaiming all currently existing comparisons as less than adequate, despite the existence of numerous strategies to assess the quality of RCTs (Kocsis et al., 2009). Furthermore, the idea that there will be an ultimate, universally accepted, "single meaningful estimate" of psychotherapy, while theoretically ideal, is far from the world of psychotherapy, where agreement about how to define "outcome" varies with each individual patient, not to mention clinician (Jacobson & Revenstorf, 1988). Even if "a single meaningful estimate of the treatments' relative efficacy" and "a single measure of mental health" could be theoretically derived, agreement about such a "single measure" appears highly unlikely. Methods for the comparisons of multiple outcome measures, such as RCI, meta-analysis, and so forth, are currently widely accepted and frequently used.

Krause also stated that "Two comparison groups of clients are for clinical purposes critically heterogeneous insofar as they have some nontrivial range of outcomes" (this issue, p. 220). Although I agree with Kraus that there are a number of problems with the RCT methodology, I do so for different reasons. For example, by definition, RCTs were not intended to generalize to treatment as usual because of the lack of representativeness of patients as a result of the exclusion of those who present with substance abuse, suicidal ideation, comorbid Axis I and II disorders, and so forth, which inherently limits generalization to treatment as usual settings. However, the problem Krause is concerned with was that the group of patients selected to participate in RCTs remains "critically heterogeneous." This seems to ignore the fact that once patients clear the selection process, random assignment works because "it ensures that alternative causes are not confounded . . . reduces the plausibility of threats to validity . . . equates groups on the expected value of all variables at pretest, measured or not . . . (and) allows computation of a valid estimate of error variance . . ." (Shadish, Cook & Campbell, 2002, p. 248). That is, random assignment is specifically designed to reduce the probability that the groups are systematically different or similar (although that probability is never zero and of course heterogeneity exists). Additionally, the client-descriptive covariates that pose the greatest threats to randomization are the covariates that are unknown and therefore uncontrollable. Meta-analytic studies make significant contributions by summarizing accumulated research and further minimizing threats to validity. In circumstances where covariates are identified, as in pretreatment scores, stratified randomization or regression discontinuity methods could be used (e.g., where a patient is assigned to a treatment based on predetermined criteria, such as symptom scores). Similarly, if client matching on known covariates is desired, propensity scores might be used. Propensity scores "will, on average, remove all of the bias in the background covariates" (D'Agostino, 1998, p. 2267). Nonetheless, hidden influences as a result of variables that were not matched or accounted for appear to be inescapable.

While I agree that "client descriptive covariates" are important contributors to outcomes, Krause envisions "63 subgroups within which like comparisons can be made" (this issue, p. 220). He acknowledged the burden of "sample size and so of degrees of freedom" required to produce "covariate-defined subgroups with zero within-subgroup variable range and therefore zero outcome variance" (this issue, pp. 220–221), but such an exhaustively defined sample would also appear to increase the likelihood of Type I errors. The repeated analyses of the same data set, using such a large number of covariates, would also threaten the internal validity of the study as a result of prediction bias. This, in turn, would require further analyses to control for multiple comparisons (Norman & Streiner, 2000). It also begs the question whether these covariates that led to 63 subgroups exhaust all possible covariates, or whether there are still others that have not been recognized.

Much has been written about the issues of publishing statistically insignificant results from RCTs and the file-drawer effect (Rosenthal, 1979). In fact, Rosenthal himself provided three equations to calculate the number of studies to offset significant results (the so called "fail safe number") and others have introduced additional methods to address this problem. For example, a funnel plot (Light & Pillemer, 1984) represents a graphical depiction of the studies used in a meta-analysis to determine publication bias. Essentially, a funnel plot consists of plotting an estimate of the precision, such as sample size, along the vertical axis and the size of the treatment effect on the horizontal axis. The more symmetrical the funnel, the more likely that the results are not affected by publication bias. If the funnel plot is asymmetrical, publication bias is present because the smaller, nonsignificant, publications have not been published. In this case, the trim and fill technique (Duvall & Tweedie, 2000) could be used to estimate and adjust for the number of unpublished studies. This method estimates what data the missing studies would have contributed and the findings are then revised accordingly. More recently, Howard et al. (2009) provided additional strategies to reduce file-drawer effects through the combined use of null hypothesis statistical tests, effect size statistics, and Bayesian analyses, which also allow us to "adequately compare the results of these RCTs and integrate these in a . . . meaningful estimate of the treatments' relative efficacy" (Krause, this issue, p. 220). Hence, although there are limitations to all techniques, there are numerous ways to compensate for the file-drawer effect.

Because effectiveness studies tend to produce lower effect size statistics compared with efficacy studies (Streiner, 2002), the former are more likely to end up in the file drawer as compared with the latter. However, practice settings, and therefore effective-

ness studies, are the most likely and important conditions where outcome relevant client covariates will be present because they have not been screened out. Hence, when effectiveness studies do prove to be significant, not only are they more likely to generalize to other typical practice settings and patients, but are also far more likely to contain outcome relevant covariates. Therefore, effectiveness studies are not only more relevant to practitioners and their patients, but also contain rich sources of data for predictive outcome variables.

Most importantly, however, "different outcome-influential types of clients" can be identified through various methods currently available, "to understand why we get the mean outcome differences" without plotting each individual score from the comparison group (Krause, this issue, p. 221). Despite the less than perfect methods available, outcome relevant covariates can be meaningfully identified in the treatment group and the use of multiple methods improves the accuracy of our treatment predictions. However, we can do better than assessing group mean outcome differences obtained from RCTs by conducting effectiveness studies and studying change at the individual level, by using RCI and CS techniques and identifying those who fail to significantly improve. By further analyzing those who do and do not improve in effectiveness studies, we can identify practice relevant covariates, as best currently exemplified by Lambert (2010), and alter the course of treatment outcomes.

References

D'Agostino, R. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistical Medicine, 17,* 2265–2281.

Derogatis, L. (1994). *SCL-90-R: Administration, scoring and procedures manual-II.* Minneapolis, MN: NCS, Pearson.

Duval, S., & Tweedie, R. (2000). Trim and Fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463.

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2002). Dose-response relationships do exist: More treatment is better. *Clinical Psychology Science and Practice, 9,* 329–343.

Howard, G., Hill, T., Maxwell, S., Baptista, T., Farias, M., Coelho, C., . . . Coulter-Kren, R. (2009). What's wrong with research literatures? And how to make them right? *Review of General Psychology, 13,* 146–166.

Howard, K., Moras, K., Brill, P., Martinovich, & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness and patient progress. *American Psychologist, 51,* 1059–1064.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15,* 336–352.

Jacobson, N. S., & Revenstorf, D. (1988). Statistics for assessing the clinical significance of psychotherapy techniques: Issues, problems and new developments. *Behavioral Assessment, 10,* 133–145.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19.

Kocsis, J., Gerber, A., Milrod, B., Roose, S., Barber, J., Thase, M., . . . Leon, A. (2009). A new scale for assessing the quality of randomized clinical trials of psychotherapy. *Comprehensive Psychiatry, 51,* 319–324.

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

Lambert, M. J. (2010). *Prevention of treatment failure.* Washington, DC: American Psychological Association.

Light, R., & Pillemer, D. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Norman, G., & Streiner, D. (2000). *Biostatistics: The bare essentials.* Hamilton, Canada: B. C. Decker.

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Shimokawa, K., Lambert, M., & Smart, D. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology, 78,* 298–311.

Streiner, D. S. (2002). The 2 "Es" of research: Efficacy and effectiveness trials. *Canadian Journal of Psychiatry, 47,* 552–556.

Wise, E. A. (2003). Psychotherapy outcome and satisfaction methods applied to intensive outpatient programming in a private practice setting. *Psychotherapy: Theory, Research and Practice, 40,* 203–214.

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change and recommendations for future directions. *Journal of Personality Assessment, 82,* 50–59.

Wise, E. A. (2010). Evidence based effectiveness of a private practice intensive outpatient program with dual diagnosis patients. *Journal of Dual Diagnosis, 6,* 25–45.

Received December 6, 2010
Accepted December 17, 2010 ■