The relevance of results in clinical research: statistical, practical, and clinical significance

Sebastian PINTEA

Babes-Bolyai University, Cluj-Napoca, Romania

Abstract

The article addresses the problem of three types of significance of research results: statistical, practical, and clinical significance. These issues are treated as chronological sequences in the evolution of how results of clinical research have been reported. For a long time, statistical significance was the only way of reporting research results. This method was subject to severe criticism showing that estimating the probability of results to be obtained by chance is not satisfactory from a clinically point of view. Statistical significance was followed by practical significance reporting, translated into size effect. Even though this change is a step forward, effect size says nothing about whether the intervention makes a real difference in the everyday life of the clients, or others whom the client interacts with. Thus, in recent years, the concept of clinical significance has been increasingly emphasized and operationalized most frequently by the quality of life, improvement in symptom level (improvement criteria), transition of patients from the dysfunctional to the functional distribution (recovery criteria) or a combination of them. Although this concept has also been subject to criticism, it has survived the debate and satisfies the set of criteria by which clinical research results are judged.

Keywords: statistical significance, practical significance, effect size, clinical significance, quality of life, reliable change, normative comparison, social significance

In the beginning there was statistical significance

For many years, statistical significance testing was the golden standard in analyzing data for many research domains, including clinical psychology and psychotherapy. This procedure proved so useful that even nowadays researchers
are impressed by its potential. For example, Abelson (1997) asserts that “significance tests fill an important need in answering some key research questions, and if they did not exist they would have to be invented” (p. 118). Similarly, Harris (1997) states that „null hypothesis significance testing (NHST) as applied by most researchers and journal editors can provide a very useful form of social control over researcher's understandable tendency to “read too much" into their data” (p. 145).

But what does statistical significance mean after all? Statistical significance refers to result significance tests, particularly the “p” value, which is the probability that research results (e.g., the difference between a control and an experimental group) be obtained when the null hypothesis is true (i.e., the two groups belong to the same population). Simply put, “the p value indicates the probability that observed findings occurred by chance” (Paquin, 1983, p. 38). In the context of clinical research, if we compare two types of therapies, A and B, statistical significance can prove if intervention A is better than intervention B.

In comparing an experimental group with a control group, the “p” value depends on three factors: 1. the magnitude of the effect (the performance difference between groups); 2. the number of observations or participants; and 3. the spread in the data (commonly measured as variance and standard deviation).

An important question for clinical research is if statistical significance testing is enough for the evaluation of intervention effectiveness and efficacy. If clinicians would only be interested in the superiority of one therapy over another, then yes, statistical significance testing would be sufficient. The problem is, however, that clinicians want to know more, such as how large is the outcome of a particular therapy?

Statistical significance does not provide information about the magnitude of change (i.e., effect size) or whether the relationship is meaningful (i.e., clinical significance), although sometimes researchers misinterpret statistically significant results as showing clinical significance. Moreover, it is quite possible, if the study involves a large sample, to have a statistically significant result but a small magnitude of change (i.e., a small effect size). On the other hand, outcomes with small “p” values (e.g., p < .01) are sometimes misinterpreted as having stronger effects than those with higher “p” values (p<.05). Actually, this misconception may have a correct basis: given a constant sample size, smaller “p” values are correlated with larger effect sizes (e.g., explained by the distance between the means or by the high homogeneity of the groups). Sometimes researchers also misinterpret statistically non-significant results as proof that two treatments have the same effect. In such a situation, a non-significant result may involve a large effect size but a small sample and implicitly not enough power to reveal that effect.

Based on such arguments, a growing volume of literature in the last decades addresses the need of calculating and reporting the effect size, as an
The relevance of results in clinical research

indicator of practical significance (Kirk, 1996; Kraemer, Morgan, Leech, Gliner, Vaske, & Harmon, 2003; Thompson, 2002).

Practical significance: Is effect size a step further?

An increasing number of authors underline the gap between researchers who only report the statistical significance of their results and practitioners who need relevant information for their decisions in clinical, counseling, educational, and organizational practice (e.g., Jacobson, Follette, & Revenstorf, 1984; Paquin, 1983; Thompson, 1993). One solution to this problem was the adoption of a conventional way of reporting research results, consisting in both statistical significance and effect size measures (Rosnow & Rosenthal, 1996).

Despite the obvious limitations of significance testing and the growing need to report effect sizes, many studies published in scientific journals continue to report only statistical significance. Talking about this tendency, Cohen (1994) states that the statistical significance test “does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!” (p. 997). Rozeboom (1997), another critical voice regarding the exclusive use of statistical significance, asserts: “null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students . . . It is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism . . .” (p. 335). Similarly, Schmidt and Hunter (1997) state that “statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution” (p. 37). Other researchers discuss the addiction to statistical significance testing and its negative consequences, arguing that such procedures “have been so ingrained in our psychological research history, training, and curriculum that the casual observer may not be fully aware of the controversies surrounding statistical significance tests. These tests and their corresponding p values are, after all, what many learned in their doctoral training, often yearning earnestly for at least one p < .05 in their dissertations. The proclivity toward NHSTs has since been strongly reinforced by our research literature, which is riddled with p values and biased toward publication of statistically significant results.” (Henson, 2006, p. 603).

The need of changing or upgrading the statistics reported in research was also emphasized through formal channels. Thereby, in the fourth edition of the APA Publication Manual (1994), authors are encouraged to provide effect-size information. As empirical studies have documented the small impact of this recommendation on published research (e.g., Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000), the fifth edition of the APA Publication Manual stresses more firmly the importance of calculating and reporting effect sizes: “For the reader to fully understand the importance of your findings, it is almost always
necessary to include some index of effect size or strength of relationship in your Results section.” (APA Publication Manual, 2001, pp. 25-26).

This impulse changed the way psychologists report and interpret research, effect sizes becoming increasingly important for effective research interpretation (Henson, 2006).

But is it the effect size a real step further? Or, better put, how far is this step from statistical significance? As with all good things in life, effect size has its own critics.

Ogles, Lunnen and Bonesteel (2001) state that, despite the advantages of effect size measures, they do not provide information regarding within-group variation. Similarly, Jacobson, Roberts, Berns, and McGlinchey (1999) conclude that effect sizes do not directly indicate the proportion of individuals that have improved or recovered as a result of the intervention. Moreover, Campbell (2005) asserts that two studies with equal effect sizes (e.g., measured by Cohen’s d) might differ in their clinical significance (i.e., the proportion of individuals who have improved or no longer need intervention).

Jacobson & Truax (1991) indicate effect size statistics is only apparently an improvement over standard inferential statistics. Even though it does reflect the magnitude of the effect, it is still relatively independent of what is called clinical significance. The authors present a potential situation where a large effect has no clinical significance: “if a treatment for obesity results in a mean weight loss of 2 lb and if subjects in a control group average zero weight loss, the effect size could be quite large if variability within the groups were low. Yet the large effect size would not render the results any less trivial from a clinical standpoint.” (p. 184). While generally large effect sizes are more likely than small sizes to have clinical significance, there are also situations of large effects that are clinically insignificant (Jacobson et al., 1999).

Clinical significance: A real step further

What is clinical significance? Literature devoted to this topic presents several definitions of clinical significance. Some are general definitions; others are operational, sending to specific ways of calculation.

Among the general definitions, one of the most popular is given by Jacobson and Truax (1991), who define clinical significance of the treatment as "its ability to meet standards of efficacy set by consumers, clinicians and researchers" (p. 12). Another definition in this category is given by Kazdin (1999) who states that clinical significance refers to the practical or applied value or importance of the effect of the intervention - that is, whether the intervention makes a real (e.g., genuine, palpable, practical, noticeable) difference in the everyday life of the clients, or of others whom the client interacts with. Another general definition is given by Ogles et al. (2001), namely that clinical significance
The relevance of results in clinical research

shows whether clients have returned to normal functioning or the extent their lives have been positively altered.

Regarding operational definitions, Jacobson et al. (1984) employ an illustrative one, stating that "a change in therapy is clinically significant when the customer moves from the dysfunctional to the functional range during the course of therapy" (p. 340).

In trying to define the concept of clinical significance, a question that helps us catch the complexity of what it means is „whom is the treatment effect clinically significant to?” (Kazdin, 1999). There are at least two different perspectives on the relevance of a therapeutic result: (1) the investigator’s perspective, emphasizing symptoms change. Researchers may come up with reasonable definitions and conventions for an outcome to be considered clinically significant, but the change in symptoms may not reflect what is actually important to the client. (2) The client’s perspective was often neglected in the evaluation of clinical significance. Kazdin (1999) states that the client has been largely excluded from the process of defining a clinically significant change. From this point of view, a change should be considered clinically significant if, at the end of treatment, the client acknowledges the change as very important or if the change has a palpable impact on his or her life. However, there are cases when researches are not as much concerned with the opinion of the client (e.g., a young child with autism, an adolescent with conduct disorder, an adult with borderline personality disorder). We must also consider that this is not an unambiguous criterion; there are all sorts of possible biases that can influence the client’s opinion.

Arguments for using clinical significance

Several relevant authors in the research methods field have underlined the necessity of calculating and reporting the clinical significance of results. One of them, Kendall asserts that „as the mental health field moves forward toward a better understanding of the psychosocial treatments that can be said to be efficacious, we must address the degree to which treatment outcomes possess clinical significance” (Kendall, 1999, p. 283). In this respect, when research results are published, the indicators of statistical significance and practical relevance (effect size) should be accompanied by clinical significance analysis: “Evaluations of the outcomes of psychological treatments are favorably enhanced when the published report includes not only statistical significance and the required effect size but also a consideration of clinical significance” (Kendall, 1997, p. 3). Other authors also observe that there is increasing tendency toward reporting clinical significance when therapeutic interventions are tested. For example, Ogles, et al. (2001) state that as investigators return to the roots of psychotherapy research, methods for investigating the clinical meaning of changes are becoming a standard addition to the typical therapy outcome study.

Treatments that produce reliable effects (i.e., statistical significance) may be quite different in their impact on client functioning, and clinical significance
brings this issue to light (Kazdin, 1999). Ogles et al. (2001) illustrate this situation using the following example. A group of subjects who received a therapy for loosing weight had an average of 16 pounds weight loss. Some of them lost 30 pounds while others lost no weight or even gained weight. The question is how relevant is a 16-pound weight loss for a morbidly obese individual? Does this change reduce the risk of mortality or improve the quality of life of the individual? Statistical significance (e.g., the value of „p”) and effect size (e.g., Cohen’s „d”) do not answer these questions.

Methods for establishing clinical significance

In what methods for establishing clinical significance are concerned, four major categories can be identified: (1) methods that use "quality of life" as an outcome variable for determining the relevance of interventions (e.g., Gladis, Gosch, Dishuk, & Crits-Christoph, 1999); (2) methods that assess if the amount of change exhibited by an individual is large enough to be considered meaningful – improvement criterion (e.g., Reliable Change Index - RCI; Jacobson, Roberts, Berns, & McGlinchey, 1999; Jacobson & Truax, 1991); (3) methods concerned if, following treatment, treated individuals are indistinguishable from normals with respect to the primary complaints – recovery criterion (e.g., normative comparisons; Kendall & Grove, 1988; Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999). In addition to these „pure” categories, we can talk about a „mixed” category, which uses both improvement and recovery criteria.

“Quality of life” methods

The interest for including quality of life in clinical significance evaluation comes not only from the researchers, but mainly from health care providers and funding agencies. A problem related to this aspect of the clinical significance is quality of life measurement. Measuring this concept may include several dimensions such as life satisfaction, social and emotional functioning, social support and so on. The existing instruments for measuring quality of life can be general measures, including well-being, life satisfaction, social support and/or health-related quality of life measures (e.g., SF-36), assessing symptoms, impairment, function, disability (Czaja & Shoultz, 2003).

“Reliable change” methods

The most popular strategy in this category, also known as „improvement methods”, is the one introduced by Jacobson et al. (1984) and developed by Jacobson and Truax (1991). The authors use the concept of „reliable change” and calculate a Reliable Change Index (RCI) for each individual, using the following formula: \( \frac{(X_{post} - X_{pre})}{S_{diff}} \), where the numerator is the posttest-pretest scores difference for each subject, and the denominator \( S_{diff} \) represents the standard error of the difference between scores. \( S_{diff} \) actually describes the spread of the distribution of change scores that would be expected if no actual change had
The relevance of results in clinical research

occurred). The value of Sdiff is calculated by the formula \( \sqrt{2S^2(1-r_{xx})} \), where S² is the variance of pre-test scores and rxx is the test-retest correlation. The Sdiff in this formula is a correction introduced by Christensen and Mendoza (1986) to a previous formula described by Jacobson et al. (1984) and accepted by them. The authors consider that a change for a specific patient is reliable and the individual can be considered improved if RCI is higher than 1.96 and if this change is in the desired direction.

Another concept related to this category of methods is „noticeable change”. This approach argues that a change may be statistically reliable but not noticeable by the clients or others. Studies like those of Ankuta and Abeles (1993) shows that satisfaction with therapy is higher for improvers, while Lunnen and Ogles (1998) prove that perceived change and therapeutic alliance are significantly higher for individuals who reliably improve than for nonchangers and deteriorators, from both client and therapist perspectives. It can be concluded therefore that the statistical argument of clinical significance correlates with the qualitative argument of clinical significance.

“Normative comparison” methods

These methods are based on the following principle: “if we can demonstrate that clients are easily distinguished from a group of peers before treatment while after treatment their behavior is indistinguishable from peers, we have demonstrated a clinically meaningful change” (Ogles et al., 2001). In this case, the comparison is empirical and based on normative distributions of outcome measures (Kendall & Grove, 1988; Kendall et al., 1999; Ogles et al., 2001).

The empirical testing of a return to normal functioning can be done in several ways. One of them is to conduct a diagnostic interview and to show that, before intervention, the patient met the criteria for a specific disturbance, while after treatment he/she no longer meets the criteria.

Another method is a statistical one and consists of the use of existing normative data for a given measure of pathology. The clients’ scores are then compared to the normative distribution, based on a cutoff score (e.g., 9 or less on the Beck Depression Inventory – Beck, Steer & Garbin, 1988) or percentile levels.

A similar method is extensively described in Kendall et al. (1999) and shortly in Kendall and Sheldrick (2000). Briefly put, “the normative comparison comprises a five-step procedure. First, a range of closeness is specified (δ₁, δ₂) within which the post-treatment group mean and the normative group mean will be considered to be clinically equivalent. Second, t tests are conducted to test for equivalency between the two means. Third, a traditional t test of means is conducted to test for a statistical difference between the two means. Fourth, the results of Steps 2 and 3 are combined and classified as equivalent (if the result of
Step 2 is positive, and Step 3 negative), different (Step 2 negative, Step 3 positive), clinically equivalent but statistically different (Steps 2 and 3 positive), or equivocal, indicating that more power is required (Steps 2 and 3 negative)" (p. 767).

Other solutions are offered by Jacobson et al. (1984). The authors conceptualize return to normal functioning in three ways: (1) the level of functioning subsequent to therapy should fall outside the range of the dysfunctional population, where range is defined as extending to two standard deviations beyond (in the direction of functionality) the mean of the population; (2) the level of functioning subsequent to therapy should fall within the range of the functional or normal population, where range is defined as within two standard deviations of the mean of the population; (3) the level of functioning subsequent to therapy places that client closer to the mean of the functional population than it does to the mean of the dysfunctional population (when the score satisfies this criterion, it is statistically more likely to be drawn from the functional than from the dysfunctional population). In this case, the cut-off point is 

\[ C = \frac{(S0M1+S1M0)}{(S0+S1)} \]

where S0 and S1 are the standard deviations for the functional and dysfunctional population, and M0 and M1 are the means of the functional and dysfunctional population.

Mixed methods (both improvement and recovery criteria)

Several studies use both improvement (reliable change) and recovery (normative comparison) criteria in calculating clinical significance. The most common method is known as the Jacobson-Truax Method (Jacobson & Truax, 1991). The authors suggest that clinical significance should be determined by using both criteria (i.e., reliability of change and normative comparison). In other words, for the changes following intervention to be clinically significant, the client should shift from the theoretical dysfunctional population to a functional population, and this shift should be also reliable. In the normative comparison, the two authors also clarify how we decide for one of the three criteria proposed by Jacobson et al. (1984). Thus, when there are norms for the population, criteria 2 or 3 are preferable. In this situation, if there is an overlap between the functional and dysfunctional population, criterion 3 is preferable, while if there is no overlap, criterion 2 is recommended. If population norms are not available, criterion 1 is recommended. As the reliability of change is concerned, the RCI formula is the one described above.

Based on the simultaneous use of the two criteria (RCI and a cut-off point on one of the above criteria) individuals who received an intervention can be classified into three categories: "recovered" (reliable shift to the functional population), "improved" (reliable shift in the functional direction) and "unimproved or deteriorated" (unreliable shift or a shift in the dysfunctional direction). The proportion of each category is a measure of clinical significance.
The use of clinical significance methods

In a review published in the Journal of Consulting and Clinical Psychology, covering a 9-year period (1990–1998), Ogles et al. (2001) identified 74 studies that conducted analyses to examine clinically significant change. Approximately 35% (26 of 74) of the studies sampled used the Jacobson-Truax (1991) method for determining clinical significance (both improvement and recovery criteria), 28% (21 of 74) chose to calculate only a clinical cutoff score, not determining the RCI, 2.7% (2 of 74) used only the RCI, while the rest used other methods.

The Jacobson-Truax method has its own critics, not so much against the principles of the method (i.e., using RCI and the normative comparison) but regarding calculation procedures. These critics generated several alternatives. The main alternatives are the Gulliksen-Lord-Novick (GLN) Method (Hsu, 1989; 1999), the Edwards-Nunnally (EN) Method (Speer, 1992), the Hageman-Arrindell (HA) Method (Hageman & Arrindell, 1999) and the Hierarchical Linear Modeling (HLM) Method (Speer & Greenbaum, 1995). Bauer, Lambert, and Nielsen, (2004), using a sample of 386 outpatients who had undergone treatment in a routine clinical practice, compared the way these methods work. Their results show that the statistical method used to calculate clinical significance has an effect on estimates of meaningful change and also reveal the superiority of the Jacobson-Truax method.

McGlinchey, Atkins and Jacobson (2002) using data from 128 participants treated for major depressive disorder, compared classification rates based on the above five clinical significance methods and the accuracy of these methods in predicting depression relapse of depression two years after treatment. The authors found no significant differences between methods. In a simulation study, Atkins, McGlinchey and Beauchaine (2005) systematically explored data parameters such as reliability of measurement, pre–post effect size, and pre–post correlation that might yield differing results among the most widely considered clinical significance methods. Their results indicated that classification across methods was more similar than different, especially at greater levels of reliability, concluding that the existing methods of clinical significance appear highly comparable.

Reflections on methods for determining clinical significance

Methods for determining clinical significance are not invulnerable to criticism. Ogles et al. (2001) have classified criticisms of clinical significance methods into the following categories: the validity of the instruments (e.g., a decrease in reported symptoms may not correspond to behavioral changes); multiple measures (e.g., the number of measures of a specific problem on which the client should prove clinically significant change in order to be considered meaningfully improved); potential rater bias (e.g., some self-report instruments are too reactive to be used for judging clinical meaningfulness); regression to the
mean (e.g., individuals with high pretreatment scores may be the most likely to make large improvements); base rates of change (e.g., methods for classifying clients who shift from the dysfunctional to the functional distribution do not consider the base rate of movement between the two distributions); and the limits of a functional distribution (e.g., using the Jacobson method, a client who was severely ill and then improved to the level of a mild disturbance would not be considered clinically improved).

The Jacobson-Truax method is not a perfect method either; some of its problems have been described by the authors themselves. One of the problems has to do with the fact that the formulas in this method assume normal distribution, and it is therefore unclear whether the method will be robust to violations of the assumption that the dysfunctional and functional distributions are normal. The second issue concerns the fact that translating clinical significance in terms of recovery or return to normal functioning may not be appropriate for all disorders treated by psychotherapy (e.g., schizophrenia, autism). Other problems are related to the fact that the validity of the method is dependent on the psychometric qualities of the instruments that measure the outcome of interventions and to the fact that for numerous measures we lack normative data for functional and dysfunctional populations in order to develop standardized cut-off points (Jacobson & Truax, 1991).

Even though existing methods for determining clinical significance still present certain limitations, the inclusion of clinical significance data in research reports provides very useful information regarding the individuals who are involved in treatment and the meaningfulness of treatment outcome. Wise (2004) state that “there is a need for real world measures designed to assess both symptoms and functional capacities in intervals or increments that can be expected to respond to psychotherapy dosage, that are quantifiable, normed, and assess multiple functional domains” (p. 57).

The same author highlights the fact that the analysis of clinical significance marks a paradigm shift from studying treatment groups to studying individual change within those groups. This paradigm shift is captured and positively evaluated by other authors, showing that “the examination of the clinical significance of therapy-produced changes is a welcomed step in the evolution of the study of treatment and builds on a base of methodological advances” (Kendall et al., 1999).

In the last years, methods for examining improvement and/or recovery are becoming more and more accepted and used. Their criticism has not killed them but made them stronger. As Wise (2004) asserts, clinical significance methodology “has withstood rigorous debate and survived stronger than originally conceived” (p. 57).
Conclusions

This paper is a plea for reporting all three types of results significance in clinical research: statistical, practical and clinical. As noted above, for a long time, statistical significance was the only kind of result reported in clinical research. Even though statistical significance is very useful, reflecting the likelihood that research findings are due to chance or highlighting the superiority of an intervention over another, this information is not sufficient or satisfactory from a clinically point of view. Based on this argument, pointed out by a number of relevant authors in the research methods domain but also on the emphasis of the need to report the practical relevance of results, conveyed through The APA Publication Manual, clinical researchers have began to report both statistical significance and effect size.

Even if practical significance/ effect size is a step forward in relation to statistical significance, it has also proved insufficient to analyze the effectiveness of interventions. Thus, effect size is a statistical indicator that is not independent of statistical significance (in particular, the p value) and it says nothing about a number of issues highly relevant from a clinical point of view, such as intra-group variation or the proportion of patients who improve or recover after an intervention.

In this context, the concept of clinical significance is recently increasingly being used in reporting results of clinical research. As discussed in the article, the most common methods of reporting clinical significance are concerned with the extent to which the intervention improves quality of life, the extent to which there is an improvement in symptoms level (i.e., improvement criterion), the transition of patients from the dysfunctional to the functional distribution (i.e., recovery criterion) or combinations of these criteria. As noted, the best-known method of calculating clinical significance is the Jacobson-Truax method. Although it has its own limits, this method has proved its superiority over alternative strategies.

Finally, it is worth mentioning that the three types of significance are not mutually exclusive but complementary in reporting results of clinical research. In this regard, Kirk (2001) recommends the following research questions: (1) Is the observed result an actual one, or should it be attributed to chance? (2) If the result is real, how large is it? and (3) Is the result large enough to be meaningful and useful? An important step even further would be the social relevance of clinical results, meaning not just economic indicators such as cost-effectiveness but also the impact of the results on society at large. Of course, this would require precise evaluation strategies, which are still a challenge for researchers.
REFERENCES

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.). What if there were no significance tests? (pp. 117-141). Mahwah, NJ: Erlbaum.


Articles Section

Harlow, S. A. Mulaik & J. H. Steiger (Eds.). *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.


