Running head: CLINICAL SIGNIFICANCE DECISION

The Clinical Significance Decision

David J. Weiss, Ward Edwards, Jie W. Weiss

Abstract

An important element in using evidence to select therapy is the determination of whether a treatment is clinically superior to its competitors. Statistical significance tells us that an observed difference is reliable, but whether the difference is large enough to be important cannot be resolved by applying a statistical formula. The determination of clinical significance is a decision. As a decision, it depends upon utilities, which are inherently subjective. Those who summarize the research literature are urged to provide sufficient information that the various stakeholders – patients, practitioners, and payers - can make that assessment from their own perspectives.

The Clinical Significance Decision

In recent years, there have been proposals to make medicine (Evidence-Based Medicine Working Group, 1992), dentistry (Chiappelli & Prolo, 2002), and psychotherapy (Kazdin & Weisz, 2003) rely more upon recent evidence than upon tradition to select among possible treatments. Practitioners are urged to consult the research literature in order to determine whether a new regimen has demonstrated superiority over the one upon which they have relied. However, interpreting the literature is not as simple as one might hope. Results are typically presented in terms of whether one treatment is statistically significantly superior to another.

What the practitioner wants to know, on the other hand, is whether the new contender will generate patient outcomes that justify its implementation. Adopting a new therapy has costs beyond actual expenses needed to carry out the program. Training in the new procedure may be needed; and even after training, lack of experience with the new technique may inspire increased uncertainty about the prognosis. If that uncertainty is transmitted honestly to the patient, the patient may lose confidence and possibly seek traditional treatment with a different professional.

The determination, made before treatment, of whether one treatment is more worthwhile than another is a decision about the clinical significance of the research results in the context of this patient's disease and other circumstances. Asking whether there is a clinically significant difference is asking whether there is a difference in the applied values of the treatments; that is, whether the data cause us to believe that the treatments lead to recognizably different results, and that the new one clearly leads to

results that are better than those produced by the old one. Most of the discussion on this question of management of beliefs has been located within the psychological literature.

In the present paper, we emphasize that the determination is truly a decision, requiring both kinds of information that are necessary in decision analysis: the probabilities and values associated with the possible outcomes. It is debatable whether significance tests answer questions about probabilities in a form suitable for decision making. But significance tests cannot answer questions about the comparative values of different treatments. The preferable option, we believe, is the one with the highest expected utility, where expected utility is the product of probability times utility (Edwards, 1954).

The frequencies observed for the various possible outcomes of treatments (including side effects), which serve to estimate the probabilities, could be provided in a research report, but sometimes are not. In an abstract, the raw material for the reviews that support adoption of one treatment over another, these details are glossed over in favor of a significance statement and a presentation of averages. The significance statement tells us that the observed difference is unlikely to be a chance result, but does not speak to the magnitude of the effect. The reason is that by using sufficiently large samples, a researcher can effectively guarantee a statistically significant difference. Therefore, achieving a statistically significant result means little in terms of importance.

Utilities are more arguable than probabilities, because they are inevitably subjective. Someone has to judge whether an observed difference is large enough to matter. Renjilian et al. (2001) reported that participants in a group program lost (statistically) significantly more weight than those getting one-on-one intervention. The

researchers then provided a theoretical rationale for the efficacy of the group program, an approach that has the additional advantage that it is cheaper to implement. However, one of that study's authors, Michael Perri, (quoted in Huff, 2004) recently characterized the mean difference, 1.9 kg, as not clinically significant. The official stance of the U. S. government (National Institutes of Health, 1998) is that only a reduction in body weight of 10% or more is clinically significant. This subjective evaluation suggests that the statistical significance test does not capture what those who work with this patient population consider to be important. That is, in the opinion of the experts, a 1.9 kg reduction in weight yields too small a difference in utility to play more than a bit part in the drama of treatment selection. In fact, so small a difference might be used as an argument against continuing to conduct research on the new treatment. "Pursuing that line of thought just wasn't worth the trouble."

Clearly, the magnitude of the effect matters, and not just in clinical decisions. One of the present authors was involved in a study in which gender differences in acceptance of rape myth were anticipated. The results showed a significant difference in the expected direction, but the mean difference was "small" (~.5 on a 7-point scale) – much smaller than expected, and smaller than other differences observed within the study. The researchers essentially dismissed the gender difference, creating a new story that accounted for the similarity across gender.

The intuitive value of effect size has been obscured by its specialized meaning in statistical analysis. Researchers have been urged to report differences in standardized units, a practice that has the advantage of fostering the integration of results across studies (Wilkinson & APA Task Force on Statistical Inference, 1999). Unfortunately, the

use of standardized units robs the effect size of its everyday meaning, which is the one that both practitioners and patients understand. To convey clinical significance, empirical results must be presented to stakeholders in comprehensible units, whether those units be expressed as life expectancy, or quality of life (Gladis, Gosch, Dishuk, & Crits-Christoph, 1999), or functionality. Only with appreciation of the magnitude of difference between treatments can the stakeholders make a reasoned choice about which option is best for them. If results are presented in units that are unfamiliar to the practitioner (whose duty it would be to explain the units to the other stakeholders), then it is unlikely that any opinions will be swayed by the study's results.

Variables are sometimes selected for their ease of measurement; typically, those that are more judgmental are harder to measure. In order to achieve statistically significant results that foster professional advancement, researchers may prefer to study variables that show rapid, dramatic effects, although slow-acting accumulative processes may well be more important. The emphasis on easily observed variables militates against the kind of long-term, multifaceted investigations that have contributed so much to our understanding of the connections between, for example, lifestyle and health (Lloyd-Jones, Larson, Beiser, & Levy, 1999; Stamler, Wentworth, & Neaton, 1986).

Rare is the treatment that has only one effect. Usually, the decision to use a new treatment requires assessment of the relative importance of the therapeutic effects and various so-called side effects, some of which can be quite undesirable. It is typical of researchers focused on statistical significance to analyze one variable that purports to capture the most important aspect of the treatment. Multivariate analysis, a superficially attractive alternative, is generally ineffective because value-related dimensions are

weighted according to their variance rather than their importance. Assessment of the tradeoffs between therapeutic effects and deleterious side effects is the heart of practical clinical decision making. The tool for doing this, called multi-attribute utility, is discussed next.

Utilities

There may well be differences of opinion among the stakeholders with regard to utilities, the worth of the anticipated outcomes of the treatment. Caregivers, patients, and payers may view differently the tradeoffs among the core components of utility - anticipated improvement, suffering, and costs. These differing views need to be faced squarely (Bauer, Spackman, Prolo, & Chiappelli, 2003).

The usual goal for a patient is complete symptom relief and restoration of functionality. Practitioners are more likely to see value in intermediate steps toward a goal, and would consider a treatment that goes farther along a promising path to be clinically significantly superior to one that merely begins to address the problem. On the other hand, a patient may consider any failure to achieve her goal as a treatment failure. For example, a dieter who wants to fit into a costume she wore in high school might attach little value to even a large weight loss. If the prevailing evidence suggests that the goal is unlikely to be achieved using any of the contending therapies, the patient may view the difference among regimens as clinically insignificant. The practitioner can try to persuade the patient that the goal is unrealistic. If that persuasion is unsuccessful, the patient might best be served by finding a different consultant.

In order for the interested parties to have an informed discussion about treatment options, those who summarize the research evidence need to provide meaningful utility

information. One can perhaps rely upon domain experts to assess utility, or it might prove worthwhile to employ focus groups composed of people for whom the particular treatments under discussion are relevant.

Formulaic Approaches

The problem for the researcher is that clinical significance is subjective, and science worships objectivity. Classical statistical significance testing has survived a host of challenges (Schmidt & Hunter, 1997), primarily because applying the techniques is very much like following a recipe, with little judgment involved once the dish has been chosen. Accordingly, researchers have sought to quantify clinical significance in a similar manner. The late Neil Jacobson and his colleagues (Jacobson, Roberts, Berns, & McGlinchey, 1999; McGlinchey, Atkins, & Jacobson, 2002) have been leaders in the movement to establish similarly routine procedures for assessing clinical significance.

Jacobson was concerned specifically with patients in psychotherapy, though his logic can easily be generalized. He considered the situation in which the patients were measured on a continuous scale of functionality, so that statistical significance could be determined in a study comparing groups of patients receiving different therapeutic approaches. Jacobson's departure from standard practice was to impose a criterion of "normal functioning" on the continuous scale. If a patient moved from the "disturbed" region below the criterion to the "normal" region above the criterion, then the therapy has achieved a clinically significant result for that patient. Any other improvement was not considered to be worthy of note. The therapies were compared with respect to the number of patients who attained this clinically significant improvement.

Jacobson's index has had wide influence, but we consider it misguided. Technically, the two-point scale, which throws away metric information about the patient's status, has the potential for peculiar, nonmonotonic, results. A therapy might change one patient from terribly dysfunctional to mildly dysfunctional (which would count as "no change"), while another patient might be changed from mildly dysfunctional to normal ("clinically significant change"). The smaller change counts as a success for the therapy and the larger change does not. Nor does the index eliminate subjectivity; the subjectivity is hidden within the imposition of the criterion for normal functioning. Demarcating a zone of normalcy along the continuum of functionality is a process that calls for expert judgment.

Another of Jacobson's suggestions for quantifying clinical significance is the reliable change index. Originally introduced in Jacobson, Follette, & Revenstorf (1984), the index has been improved so that its present form (Jacobson & Truax, 1991; Tingey, Lambert, Burlingame, & Hansen, 1996) purports to estimate true change. By itself, the index does not indicate clinical significance; rather, it specifies the amount of pre-to post-treatment change that would be statistically reliable. It is essentially a standard score. Thus, the reliable change index places a lower bound on the difference needed to declare that clinical significance has been achieved. The imprecision in the measuring instrument is taken into account, an idea whose importance seems undeniable to us. The literature on improved statistical methods for assessing clinical significance continues to expand (e.g., Bauer, Lambert, & Nielsen, 2004; Hageman & Arrindell, 1999; Speer & Greenbaum, 1995). Beutler and Moleiro (2001) have contributed to the discussion by

clarifying the meaning of equivalence testing, in which the crucial question is whether a treated group has become comparable to a nonpatient control group or normative sample.

Discussion

There can scarcely be an issue of greater importance to an applied field than determining whether an experimental result is worth incorporating into practice. Accordingly, practitioners in several of the therapeutic sub-domains within psychology have weighed in on the topic. Among those we have noted are Chorpita (2001) from developmental psychology, Donders (2000) from neuropsychology, Drotar (2000) from pediatric psychology, Ogles, Lunnen, & Bonesteel (2001) from clinical psychology, and Thompson (2002) from counseling. There has been a clarion call for standardization of methods from a distinguished methodologist (Kirk, 1996).

Our views accord closely with those expressed in a comprehensive discussion by Kazdin (1999). Kazdin notes that there are usually multiple dimensions of change brought about by successful treatment. The magnitude of observable symptoms and degree of impairment experienced by the patient are surely correlated, but the correlation need not be close to 1. Kazdin (1999) also recognized that the patient's perspective on treatment success may not accord with the practitioner's perspective, and that this disagreement does not connote methodological failings but rather reflects their different goals.

When the decision must be faced about whether to adopt a new treatment, someone has the responsibility of being the final arbiter. The practitioner is the domain expert, and that expertise can be enhanced by well-communicated information from the research literature (Prolo, Weiss, Edwards, & Chiappelli, 2003). Ethics discussions in

recent years have made practitioners aware that the final decision ought to rest with the patient (Corey, Corey, & Callanan, 1998), informed by as much expert consultation as possible.

The researcher's responsibility is to provide evidence in terms that the stakeholders can grasp. Clinical significance is an important component of that evidence. To the extent possible, the evidence should incorporate feedback from previous patients who have experienced the treatments under consideration. The determination of clinical significance is inescapably a subjective process, but that does not imply it is chaotic. Expert researchers are so designated because they have the knowledge that allows them to extract the vital information from their data. Those who undertake systematic reviews of the literature (Cook, Mulrow, & Haynes, 1997) must include information about clinical significance as they coordinate the existing evidence. Practitioners and patients must be able to rely upon their expertise as well.

The schism between researchers and clinicians is a barrier to both the determination of clinical significance and the implementation of evidence-based treatments. The knowledge that comes from direct interaction with patients is crucial in assessing clinical significance. Most of this knowledge comes from experience, experience that those whose primary emphasis is research may not have had occasion to absorb. Clinical significance is not addressed because researchers lack the knowledge to assess it. Implementation fails because practitioners question the relevance of research results, feeling as though studies that proclaim statistically significant outcomes fail to address important issues. If research is to be more than an academic exercise, and if

practice is to be evidence-based, the central issue of clinical significance – what do the results really mean to the stakeholders – will be the bridge between the two subcultures.

As psychologists, we are familiar with the scientist-practitioner model of professional instruction, which folds several years of internship training into the curriculum of those who plan to specialize in clinical research. However, in the medical and dental domains, vast amounts of technical material need to be absorbed, and consequently it is typical for students who plan to be practitioners to get only a brief introduction to research issues and for students who plan to be researchers to get only a smattering of clinical experience.

We do not foresee changes in the structure of medical or dental instruction, so we expect the reality that researchers in those domains lack extensive clinical experience to persist. Our reverence for clinical significance leads to the pragmatic recommendation that researchers routinely include practitioners on the research team. The role of the practitioner in the study should not be confined to technical aspects such as administering the treatment or measuring its impact. We view the input of the experienced clinician as invaluable in the conceptualization of the research, especially with respect to determining the dependent variables to be assessed. This sea change in how research is conducted will not occur unless the economics of the enterprise encourage this collaborative enterprise. It is doubtful that the revolution will occur unless funding agencies make it so.

References

Bauer, J., Spackman, S., Prolo, P., & Chiappelli, F. (2003, October). *Clinical decision tree of oral health.* Paper presented at the First International Brain Aging Meeting, Bucharest, Romania.

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment, 82*, 60-70.

Beutler, L. E., & Moleiro, C. (2001). Clinical versus reliable and significant change. *Clinical Psychology: Science and Practice, 8*, 441-445.

Chiappelli, F., & Prolo, P. (2002). Evidence based dentistry for the 21st century. *General Dentistry, 50*, 270-273.

Chorpita, B. F. (2001). Reflections on clinical significance: What do our best treatments accomplish and how can we best find out? *Clinical Psychology: Science & Practice, 8*, 451-454.

Cook, D. J., Mulrow, C. D., & Haynes, R. B. (1997). Systematic reviews: Synthesis of best evidence for clinical decisions. *Annals of Internal Medicine. 1997, 126*, 376-380.

Corey, G., Corey, M. S., & Callanan, P. (1998). *Issues and ethics in the helping professions* (5th ed.). Pacific Grove, CA: Brooks/Cole.

Donders, J. (2000). From null hypothesis to clinical significance. *Journal of Clinical and Experimental Neuropsychology, 22*, 265-266.

Drotar, D. (2000). Enhancing reviews of psychological treatments with pediatric populations: Thoughts on next steps. *Journal of Pediatric Psychology, 27*, 167-176.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin, 51*, 380-417.

Evidence-Based Medicine Working Group. (1992). Evidence based medicine: A new approach to teaching the practice of medicine. *Journal of the American Medical Association, 268*, 2420-2425.

Gladis, M. M., Gosch, E. A., Dishuk, N. M., & Crits-Christoph, P. (1999). Quality of life: Expanding the scope of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 320-331.

Hageman, W. J., & Arrindell, W. A. (1999). Establishing clinically significant change: Increment of precision between individual and group level of analysis. *Behavior Research and Therapy, 37*, 1160-1193.

Huff, C. (2004). Teaming up to drop pounds. *Monitor on Psychology, 35*, 56-58.

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., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67*, 300-307.

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

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 332-339.

Kazdin, A. E., & Weisz, J. R. (Eds.) (2003). *Evidence-based psychotherapies for children and adolescents.* New York: Guilford.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Lloyd-Jones, D. M., Larson, M. G., Beiser, A., & Levy, D. (1999). Lifetime risk of developing coronary heart disease. *Lancet, 353*, 89-92.

McGlinchey, J. B., Atkins, D.C., & Jacobson, N. S. (2002). Clinical significance methods: Which one to use and how useful are they? *Behavior Therapy, 33*, 529-550.

National Institutes of Health, National Health, Lung, and Blood Institute (1998). *Clinical guidelines on the identification, evaluations, and treatment of overweight and obesity in adults: The evidence report* (Publication No. 98-4083). Washington, DC: U. S. Government Printing Office.

Ogles, B. M., Lunnen, K. M., & Bonesteel, K. (2001). Clinical significance: History, application, and current practice. *Clinical Psychology Review, 21*, 421-426.

Prolo, P., Weiss, D. J., Edwards, W., & Chiappelli, F. (2003). Appraising the evidence and applying it to make wiser decisions. *Brazilian Journal of Oral Science, 2*, 200-203.

Renjilian, D. A., Perri, L. G., Nezu, A. M., McKelvey, W. F., Shermer, R. L., & Anton, S. D. (2001). Individual versus group therapy for obesity: Effects of matching participants to their treatment preferences. *Journal of Consulting and Clinical Psychology, 69*, 717-721.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.

Speer, D. C. & Greenbaum, P. E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology, 63,* 1044-1048.

Stamler, J., Wentworth, D., & Neaton, J. D., (1986). Is relationship between serum cholesterol and risk of premature death from coronary heart disease continuous and graded? Findings in 356,222 primary screenees of the Multiple Risk Factor Intervention Trial (MRFIT). *Journal of the American Medical Association, 256,* 2823-2828.

Thompson, B. (2002). "Statistical", "practical". and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64-71.

Tingey, R. C., Lambert, M. J., Burlingame, G. M., & Hansen, N. B. (1996). Assessing clinical significance: Proposed extensions in method. *Psychotherapy Research, 6,* 109-123.

Wilkinson, L., & APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Authors' Note

David J. Weiss, Department of Psychology, California State University, Los Angeles. Ward Edwards, Professor Emeritus, Department of Psychology, University of Southern California, Los Angeles, California. Jie Wu Weiss, Division of Kinesiology and Health Sciences, California State University, Fullerton.

Preparation of this manuscript was partially supported by grant #FA9550-04-1-0230 from the Air Force Office of Scientific Research. We wish to thank Janet Bauer for valuable comments on an earlier draft.

Correspondence concerning this paper should be directed to David J. Weiss, Department of Psychology, California State University, Los Angeles, 5151 State University Drive, Los Angeles, CA 90032. email: dweiss@calstatela.edu.