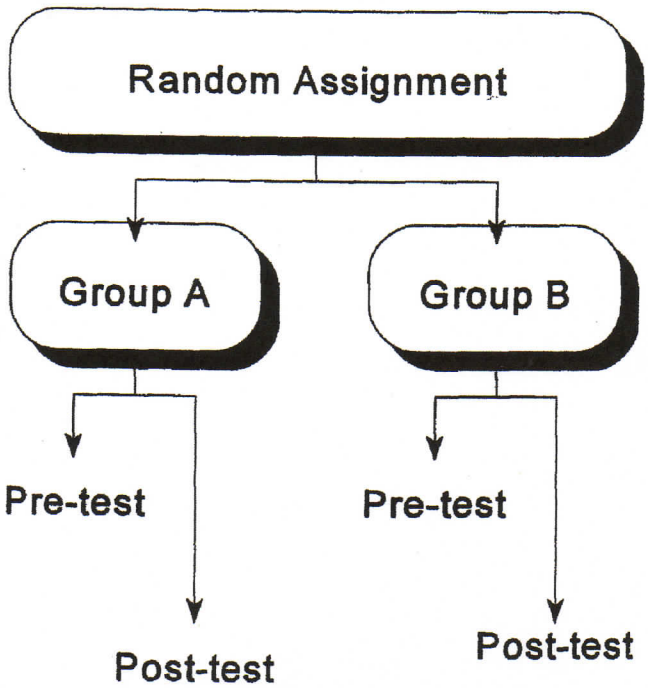

Schematic diagram of true-experimental design showing random assignment of subjects to groups followed by pretest and posttests.

is true or false, to the probability that a result would be replicated, or to treatment effects, nor is it a valid indicator of the magnitude or the importance of a result."

Several problems associated with statistical significance testing in clinical research, including small sample sizes, low statistical power, and failure to establish the magnitude of a treatment effect, are subsumed in what Cook and Campbell[3] call statistical conclusion invalidity, which they define as the inappropriate use or interpretation of statistical tests. Various authorities have argued that statistical conclusion invalidity can reduce the sensitivity of experimental procedures and produce quantitative results that do not accurately reflect the impact of treatment.[10-12]

Ottenbacher and colleagues[13-19] have argued that the presence of statistical conclusion invalidity leads to misinterpretations in clinical research and that these misinterpretations reflect negatively on the effectiveness of rehabilitation practice and contribute to dissatisfaction with traditional research approaches. This article illustrates how problems of statistical conclusion invalidity negatively affect the interpretation of rehabilitation research and reduce the usefulness of clinical trials. Three specific problems are discussed: clinical versus statistical significance, low statistical power, and replication.

CLINICAL VERSUS STATISTICAL SIGNIFICANCE

In the simplest quantitative clinical trial, a treatment effect is reflected as a difference between mean scores for the treatment and control groups (fig). The magnitude of the raw treatment effect can be determined by subtracting the mean posttest score for the control group from the mean posttest score for the treatment group. For example, if—after a program of rehabilitation—the mean elbow range of motion (ROM)[20] score for a group of patients with arthritis is 90° of flexion and the mean ROM score for the control group, not receiving intervention, is 60° of flexion, then the treatment effect would be 90° − 60°, or 30°. This assumes an adequate research design where subjects were randomly assigned to treatment and control groups resulting in equivalent groups and the outcome is blindly recorded (see figure).

Raw estimates of effect size (eg, the 30° of ROM in the previous example) have relatively little usefulness in establishing clinical or practical significance outside the context of the trial in which they occur. The two problems with raw estimates of treatment effect are: (1) they often represent values that cannot be compared across trials, or even across different measures of a similar outcome within the same trial; and (2) they do not provide information on what should be considered a small, medium, or large treatment effect for a given outcome.

Cohen[21] has pioneered the development of standardized measures of effect size that allow the results of different tests (raw scores) measuring the same outcome to be compared across trials. Cohen suggests that an effect size called a d-index be used to measure the difference between the means of two groups in terms of a common standard deviation. For instance, if the mean score for the treatment group following intervention is 85 (with a standard deviation of 10) and the mean score for the control group is 80 (with a standard deviation of 10), then the d-index is 85-80/10 or 0.50. This d-index represents a standardized metric and indicates that 5/10ths of a standard deviation separates the average subjects in the two groups. Another way of interpreting the standard deviation units is that the average person in the treatment group (receiving intervention) scored better than 69.1% of the persons in the control group (not receiving the intervention). See Cohen[21] for additional information on interpreting standardized effect size measures.

Rosenthal[22] presented an excellent illustration of the importance of effect size in interpreting statistical significance. Suppose a clinical researcher, White, reports statistically significant results for a study using a two-group design similar to that presented in the figure. A second investigator, Black, decides to replicate the original study. Black uses an identical research design and the same independent and dependent variable, but Black reports a statistically nonsignificant result. A closer examination of the two studies provides the following quantitative information: White's study: $t = 2.20$ ($df = 78$, $p < .05$) and Black's study: $t = 1.05$, ($df = 18$, $p < .30$).

Cohen,[21] Rosenthal,[22] and others[19] have noted that effect size measures, such as the d-index, can be computed directly from the results of statistical tests. The formula to compute the d-index for a standard t test is $2t/\sqrt{df}$. Using this formula, the d-index for White's study is $2(2.20)/\sqrt{78}$ or 0.50 and the d-index for Black's study is $2(1.05)/\sqrt{18}$ or 0.50. The d-index (0.50) is identical for both studies, but White reported statistically significant results (ie, rejected the null hypothesis) and Black reported statistical nonsignificant results (ie, failed to reject the null hypothesis).

Cohen[21] defines an effect size as the degree to which a null