The Statistical Comparison of Clinical Trials

Nina R. Schooler, Ph.D.

Being able to compare clinical trials using statistically proven analyses is essential, but certain protocols must be followed if such a comparison is to be made. The difficulty of making meaningful statistical comparisons is illustrated in 5 clinical trials comparing atypical antipsychotics. Certain judgments can be made on the basis of internal and external validity, for example, but there are many other areas—effect size, for example—in which it is impossible to make any statistical comparisons across these trials because they were not conducted in a uniform fashion. In the final analysis, it is doubtful that the differences between atypical antipsychotics covered in these trials are greater than those due to chance.
(J Clin Psychiatry 2001;62[suppl 9]:35–37)

Being able to compare clinical trials using statistically proven analyses is essential, but certain protocols must be followed if such a comparison is to be made. The difficulty of making meaningful statistical comparisons is illustrated in 5 clinical trials comparing atypical antipsychotics.[1–5] Certain judgments can be made, but there are many other areas—effect size, for example—in which it is impossible to make any statistical comparisons across these trials because these trials were not conducted in a uniform fashion.

TERMS AND CONCEPTS

A study may be said to be valid in 2 ways. A study has *internal validity* when "the observed differences between the control and comparison groups may, apart from sampling error, be attributed to the effect under study. *External validity* or *generalizability* means that a study can produce unbiased inferences regarding the target population, beyond the subjects in the study."[6(p575)] Randomization, which controls for selection bias, is an indicator of high internal validity in a trial. How assessments are made also affects internal validity. Blinded assessments contribute to high internal validity, the assumption being that blinded assessments are not biased by knowledge of the treatment being assessed. However, the protection against this bias in blinding is frail, particularly when the investigator wishes to see behind the blind. Ralph Horwitz, M.D., reminds us that a study can be unblinded "whenever treatment effects include a readily measured physiologic variable."[7(p504)] Under such circumstances, and others, there are many trials in which blinding remains a polite fiction.

In a study with high external validity, treatment medications are administered under the same conditions that they would normally be administered by clinicians. Studies that have high external validity, however, may well have low internal validity. A range of studies is useful (Table 1). In the trials under consideration, Tran et al.[1] and Conley, Mahmoud, et al.[2] have high internal validity. QUEST[4] lies midway along a scale of internal validity; it is randomized but lacks blind assessment. Of the 2 remaining studies, Conley et al.[5] has higher external validity than Ho et al.[3] The Ho et al. study has 2 serious problems. First, the population size (N = 42) is quite small, and it seems unlikely that treatment assignment was random. Second, the analysis of the follow-up data is compromised by the fact that only the subjects who continued receiving the study medication were included. Apparently, most of the subjects who were not included in the follow-up analysis had been switched to another treatment. In general, small study populations have other problems. Baigent[8] has pointed out the general inability of small studies to differentiate among a moderate benefit, a moderate hazard, and a trifling difference in major outcomes.

EFFECT SIZE

One way these studies might profitably be compared is through effect sizes. Effect size is the "observed or expected change in outcome as a result of an intervention. Expected effect size is used when the sample size necessary to achieve a given power is estimated, since, given a similar amount of variability, a large effect size will require a smaller sample size to detect a difference than will a smaller effect size."[6(p546)] If the data are distributed normally

From the Department of Psychiatry, Hillside Hospital, Division of Long Island Jewish Medical Center, Glen Oaks, N.Y.
Presented at the symposium "Evaluating Clinical Trial Data From Schizophrenia Research," which was held March 17, 2000, in Washington, D.C., and supported by an unrestricted educational grant from Janssen Pharmaceutica, L.P.
Reprint requests to: Nina R. Schooler, Ph.D., Department of Psychiatry, Hillside Hospital, 75-59 263rd St., Glen Oaks, NY 11004.

Table 1. Statistical Comparison in 5 Clinical Trials of Atypical Antipsychotics

Characteristic	Tran et al[1]	Conley, Mahmoud, et al[2]	Ho et al[3]	QUEST[4]	Conley et al[5]
Standard deviations	Reported	Not reported	Reported	Not reported	Reported
Effect size	Not reported	Not reported	Not reported	Not reported	Not reported
Randomization	Yes	Yes	No	Yes	No
Blinding	Yes	Yes	No	No	No

(i.e., approximately 80% of the values occur inside 1 standard deviation on either side of the mean[9]), then a standard deviation can be used as an estimate of variability in both groups.

The problem with using effect size as a means of comparing the 5 studies of atypical antipsychotics under consideration here[1–5] is that we do not have information necessary to calculate effect sizes and standard deviations in the unpublished studies. But even if that information were available, no 2 of these studies are directly comparable. Streiner and Joffe[10] surveyed 69 articles in 26 different journals reporting comparisons of 2 antidepressants and a placebo. They developed 3 sets of scores based on their eligibility for inclusion in meta-analyses. For example, articles earned a minimum criteria score of 1 if they included the initial sample size, the final sample size, and some index of variability. Of the 69 articles surveyed, only 9 earned a minimum criteria score of 1. Clarke and Stewart[11] recommend that biases be minimized through gathering the greatest possible amount of randomized evidence, including collection of details regarding every participant in every trial. Lohr and Carey[12] report that the U.S. Agency for Health Care Policy and Research undertook a project to evaluate how well its Evidence-Based Practice Centers were carrying out systematic reviews of the literature, known as evidence reports. They noted that there are many checklists and other instruments for assessing the methods or the clinical applicability of individual reports while admitting that how reliable, valid, feasible, and useful these instruments are changes or is undetermined.

As shown in Table 1, the 5 studies under consideration are not comparable in terms of critical design elements. Two of the studies did not involve randomization.[3,5] Three of the studies were not blind.[3–5] Finally, the duration of treatment exposure varied markedly across the trials. Thus, even if the statistical information were available to allow calculation of effect sizes in each trial, a comparison of the effect size would be inappropriate.

MATCHING PATIENT TO DRUG

In 1965, Klett and Moseley[13] attempted to develop predictor profiles establishing which patient would respond best to which (then new) antipsychotic medication. These, and similar attempts by others, ultimately failed; they were irreproducible. Clinical experience confirmed that certain patients fared better taking one antipsychotic than another, but a particular type of patient could not be associated with response or nonresponse to a particular antipsychotic. With the advent of the atypical antipsychotics, this question has become important to clinicians again. The trials mandated by the U.S. Food and Drug Administration as part of any New Drug Application require new drugs to demonstrate a significant difference against placebo or active control. Initially, drug companies chose to compare atypical antipsychotics with a placebo, because it was felt that they would not be able to establish a significant difference with the other antipsychotics. Drug companies are now interested in assessing the differences among competing antipsychotics, for example. Further, determining differences among the new antipsychotics and specifically linking patient characteristics to differential response are of great interest and importance to patients and clinicians. John M. Kane, M.D.,[14] points out that although it is widely assumed a patient who fails to respond to one drug might respond to another, this idea has rarely been tested. In addition, he cites the lack of studies testing whether switching a patient from one antipsychotic drug to another has any value.

SECONDARY ANALYSES

Secondary analyses that are conducted post hoc should be regarded as exploratory and as hypothesis-generating for new studies. Performing a new analysis to test a hypothesis suggested by the data but not explicitly tested by a study increases the possibility of a type I error, or finding a difference between the groups studied when, in fact, no difference exists.[15] Secondary, post hoc analyses can be very interesting, but their important limitations must be recognized. Treating these analyses as if they were primary violates sound statistical practice.

CONCLUSION

The studies under review[1–5] offer a cautionary lesson. We have published results for only 2 of the 5 studies. Even so, they do not seem to avail themselves of meaningful comparisons. If medical studies are to be compared in meaningful, useful ways, they must be designed in a manner that makes these comparisons easy to accomplish.

Disclosure of off-label usage: The author has determined that, to the best of her knowledge, no investigational information about pharmaceutical agents has been presented in this article that is outside U.S. Food and Drug Administration–approved labeling.

REFERENCES

1. Tran PV, Hamilton SH, Kuntz AJ, et al. Double-blind comparison of olanzapine versus risperidone in the treatment of schizophrenia and other psychotic disorders. J Clin Psychopharmacol 1997;17:407–418
2. Conley RR, Mahmoud R, and the Risperidone Study Group. Risperidone versus olanzapine in patients with schizophrenia and schizoaffective disorder. Presented at the 38th annual meeting of the American College of Neuropsychopharmacology; Dec 12–16, 1999; Acapulco, Mexico
3. Ho B-C, Miller D, Nopoulos P, et al. A comparative effectiveness study of risperidone and olanzapine in the treatment of schizophrenia. J Clin Psychiatry 1999;60:658–663
4. Mullen J, Reinstein M, Bari M, et al. Quetiapine and risperidone in outpatients with psychotic disorders: results of the QUEST Trial. Presented at the biennial meeting of the International Congress on Schizophrenia Research; April 17–21, 1999; Santa Fe, NM
5. Conley RR, Love RC, Kelly DL, et al. A comparison of rehospitalization rates between patients treated with atypical antipsychotics and those treated with depot antipsychotics. Presented at the 54th annual convention and scientific program of the Society of Biological Psychiatry; May 13–15, 1999; Washington, DC
6. Iverson C, Flanagin A, Fontanarosa PB, et al. American Medical Association Manual of Style: A Guide for Authors and Editors. 9th ed. Baltimore, Md: Williams & Wilkins; 1998
7. Horwitz RL. Complexity and contradiction in trial research. Am J Med 1987;82:498–510
8. Baigent C. The need for large-scale randomized evidence. Br J Clin Pharmacol 1997;43:349–353
9. Lang TA, Secic M. Guide to statistical terms and tests, part 2. In: How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers. Philadelphia, Pa: American College of Physicians; 1997: 241–290
10. Streiner DL, Joffe R. The adequacy of reporting randomized, controlled trials in the evaluation of antidepressants. Can J Psychiatry 1998;43: 1026–1030
11. Clarke MJ, Stewart LA. Systematic reviews of randomized controlled trials: the need for complete data. J Eval Clin Pract 1995;1:119–126
12. Lohr KN, Carey TS. Assessing "best evidence": issues in grading the quality of studies for systematic reviews. Jt Comm J Qual Improv 1999;25: 470–479
13. Klett CJ, Moseley EC. The right drug for the right patient. J Consult Psychol 1965;29:546–551
14. Kane JM. Pharmacologic treatment of schizophrenia. Biol Psychiatry 1999;46:1396–1408
15. Lang TA, Secic M. Comparing groups, II: the multiple testing problem. In: How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers. Philadelphia, Pa: American College of Physicians; 1997:81–92