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.

(Eing 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)

B

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.” 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.” 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. and Conley, Mahmoud, et al. have high internal validity. QUEST lies midway along a scale of internal validity; it is randomized but lacks blind assessment. Of the 2 remaining studies, Conley et al. has higher external validity than Ho et al. 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 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.” If the data are distributed normally
(i.e., approximately 80% of the values occur inside 1 standard deviation on either side of the mean), 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 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 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 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 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. Three of the studies were not blind. 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 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 asso-

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Tran et al</th>
<th>Conley, Mahmoud, et al</th>
<th>Ho et al</th>
<th>QUEST</th>
<th>Conley et al</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviations</td>
<td>Reported</td>
<td>Not reported</td>
<td>Reported</td>
<td>Not reported</td>
<td>Reported</td>
</tr>
<tr>
<td>Effect size</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Randomization</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Blinding</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

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

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. 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 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

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