Hypothesis and Hypothesis Testing in the Clinical Trial

Jeffrey A. Lieberman, M.D.

The hypothesis provides the justification for the clinical trial. It is antecedent to the trial and establishes the trial's direction. Hypothesis testing is the most widely employed method of determining whether the outcome of clinical trials is positive or negative. Too often, however, neither the hypothesis nor the statistical information necessary to evaluate outcomes, such as p values and α levels, is stated explicitly in reports of clinical trials. This article examines 5 recent studies comparing atypical antipsychotics with special attention to how they approach the hypothesis and hypothesis testing. Alternative approaches are also discussed. *(J Clin Psychiatry 2001;62[suppl 9]:5–8)*

When clinicians evaluate scientific reports about drug treatments, it is important to know if the reports are accurate and based on valid and properly conducted studies. To know this, certain elements must be considered. Formulation of a hypothesis and experimental design to test it are fundamental components of the scientific method. In biomedical and psychiatric research the hypothesis provides the justification for the clinical trial. It is antecedent to the trial, establishes the trial's direction, and determines whether the outcome is positive or negative. Even when a hypothesis is not explicitly stated for a clinical trial, in almost all cases there is, in fact, a hypothesis guiding the direction of clinical testing. As part of a review of comparative trials of atypical antipsychotic drugs, we examined the hypotheses of 5 recent clinical studies (Table 1).[1–5]

TERMS AND CONCEPTS

Lang and Secic, in their book *How to Report Statistics in Medicine*, define *hypothesis* as

> a testable statement about a proposed relationship between two or more variables—a statement that will be accepted or rejected on the basis of the results of the study. . . . If one hypothesis can be rejected, another, alternative hypothesis tends to be thought of as more plausible. That is, if a given hypothesis is supported, certain interpretations are more likely, and if the hypothesis is not supported, other interpretations are more likely.[6(p67)]

The most common hypothesis, and one frequently implicit in clinical trials, is the *null hypothesis*, the hypothesis that there is no difference between 2 (or more) groups being observed. The null hypothesis stands against the alternative hypothesis, the hypothesis that there is a difference between the 2 (or more) groups being observed. After the null hypothesis has been tested by the data obtained from a clinical trial, for example, the probability of finding a difference between the 2 groups as large or larger than that actually observed, assuming that the null hypothesis is correct, is calculated. This probability is the *p value*; the smaller the p value, the stronger the evidence against the null hypothesis. The p value is compared with a threshold of significance called the *α level*, arbitrarily set, usually at .05, at the outset of research. If $p < .05$, a statistically significant difference between groups on the mean response of interest can be concluded. The *α level* is the probability of committing a *type I error*, or wrongly concluding that the null hypothesis is incorrect,[6] meaning the possibility that the findings are due to chance. With an α of .05, the possibility of false or chance findings is less than 5 out of 100.

The ability of a test to detect a difference of a given size if it actually exists is called *statistical power*. The fact that no statistically significant difference is detected in a trial may mean that no such difference exists, but it may also mean that a sample size is too small to allow for the collection of sufficient data to show whether such a difference existed. Statistical power is defined as $1 - \beta$, where β is the probability of committing a *type II* error, or wrongly concluding that the null hypothesis is correct. β is a probability value established between 0 and 1, usually 0.1 for 90%

From the Department of Psychiatry and Mental Health and Neuroscience Clinical Research Center, University of North Carolina School of Medicine, Chapel Hill.

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: Jeffrey A. Lieberman, M.D., Department of Psychiatry and Mental Health and Neuroscience Clinical Research Center, University of North Carolina School of Medicine CB7160, 7025 Neurosciences Hospital, Chapel Hill, NC 27599-7160.

Table 1. Hypothesis and Hypothesis Testing in 5 Clinical Trials of Atypical Antipsychotics[a]

Characteristic	Tran et al[1]	Conley, Mahmoud, et al[2]	Ho et al[3]	QUEST[4]	Conley et al[5]
Hypothesis	Different efficacy and safety profiles	Different EPS rates	Not reported	Not reported	Not reported
Drugs compared	Olanzapine, risperidone	Risperidone, olanzapine	Risperidone, olanzapine	Quetiapine, risperidone	Clozapine, risperidone, olanzapine, haloperidol decanoate, fluphenazine decanoate
α Level	.05	.05	Not reported	Not reported	Not reported
β Level	Not reported	Not reported	Not reported	Not reported	Not reported
p Value	$p < .05$	$p < .05$	Not reported	Not reported	Not reported
Statistical power	Not reported	Not reported	Not reported	Not reported	Not reported
Statistical tests	1-tailed[b]	2-tailed	2-tailed	Not reported	Not reported

[a]Abbreviation: EPS = extrapyramidal symptoms.
[b]Test of primary efficacy; all other tests 2-tailed.

power or 0.2 for 80% power. Generally speaking, the smaller the difference a test is powered to detect, the larger the sample size necessary.[6]

One-tailed (or 1-sided) and *2-tailed* (or 2-sided) tests analyze deviations from the null hypothesis.[7] Only 2-tailed tests can analyze such deviations in either direction; they are more conservative than 1-tailed tests, requiring a larger change to achieve statistical significance. A 1-tailed test is appropriate only when the direction of the difference between 2 groups "is known in advance or when differences observed in the opposite direction are not of interest or are not possible,"[8(p271)] according to Lang and Secic. Rigby[9] is less sanguine, citing the 2-tailed test as the norm and preferring it over the 1-tailed test, which increases the chance of a type I error.

Fox and Mathers[10] analyzed 1422 statistical tests in 85 articles appearing in the *British Journal of General Practice* for statistical power. The tests, with a median power of 0.71, had a slightly greater than two thirds likelihood of rejecting false null hypotheses. Twenty-one articles had a likelihood of attaining significant results poorer than those reached by a coin toss when the null hypothesis is false.

Hypothesis testing involves the identification of effects that are statistically significant. Statistical significance should not be confused with clinical significance. The former is a reflection of the influence of chance on the outcome, while the latter reflects the biological or clinical value of the outcome. Statistical results are derived from groups of individuals, while clinicians regard distinct individuals. Statistical conclusions demand sufficient data to be valid, while medical decisions must necessarily always be made in the face of insufficient data. Statistical answers indicate probabilities, while medical treatment demands commitment to decisions. Statistical analysis always needs measurement, while medicine can make intuition necessary. The statistical and clinical applications of the term *normal* are often confused and vague.[11]

We examined the small number of studies that compared atypical antipsychotic drugs for methodological elements and results.

TRAN ET AL.

In 1997, Tran et al.[1] carried out a randomized double-blind comparison of olanzapine and risperidone in the treatment of 339 patients with schizophrenia and other psychotic disorders over 28 weeks. The researchers noted the pharmacologic differences between the 2 agents, including receptor binding affinities, preclinical pharmacology, pharmacokinetics, and in vivo neuroimaging profiles. They hypothesized that olanzapine and risperidone would have different safety and/or efficacy characteristics as a result. The Tran et al.[1] study was also designed to test a number of secondary hypotheses assessing the comparative atypicality of the 2 agents.

The primary efficacy analysis in the Tran et al.[1] trial was 1-tailed, analyzing deviation from the norm for only one patient group. In all other analyses, main effects were tested at a 2-tailed α level of .05, indicating that p levels smaller than .05 would yield statistical significance. Most results favored olanzapine and the differences were often statistically significant.

This trial did reveal efficacy and safety differences between olanzapine and risperidone. However, questions about design and drug doses and the use of a 1-tailed test affect the interpretation of the results. This study was sponsored by Eli Lilly and Company.

CONLEY, MAHMOUD, ET AL.

Conley, Mahmoud, et al.[2] compared risperidone with olanzapine in a randomized double-blind treatment trial of schizophrenia and schizoaffective disorder over 8 weeks. On the basis of published rates of extrapyramidal symptoms (EPS) associated with the agents, the researchers hypothesized that they would not find a difference in EPS rates between risperidone- and olanzapine-treated patients, and the study was powered to detect a difference if one existed. Doses of both drugs were limited. The study found no statistically significant difference in the rate of adverse events due to EPS; the severity of EPS was reduced in both treatment groups.

This study was sponsored by Janssen Pharmaceutica, L.P. All statistical tests were 2-tailed at an α level of .05. Analyses controlled for investigator, baseline score for continuous variables, and baseline age for the Extrapyramidal Symptom Rating Scale and wherever age was a statistically significant correlate.

HO ET AL.

Ho et al.[3] conducted a study comparing the effectiveness of risperidone and olanzapine in treating schizophrenia. The researchers examined a cohort of subjects who had been treated with risperidone or olanzapine under actual clinical practice conditions over a 6-month period. No a priori hypothesis was proposed at the outset of this open-label study. This study was funded by the National Institute of Mental Health and the Nellie Ball Trust Foundation, so it would seem that the null hypothesis (that there is no difference between the 2 variables) was being tested in this small (N = 42), open-label study. The researchers found risperidone more effective for treatment of psychotic symptoms at 6 months, but otherwise the 2 medications were equally effective in the routine clinical care of patients with schizophrenia.

All tests of significance were 2-tailed. No α level was stated. The limitations of the study—its small sample size, nonrandomized open-label design, and relatively brief treatment period—were readily acknowledged by the authors. The β level of significance was not stated. Given the small sample size, this information would be necessary to know whether the trial had the statistical power to detect differences between the medications if they occurred.

QUEST

QUEST[4] compared the tolerability and efficacy of quetiapine and risperidone in a 4-month, multicenter, open-label trial in 751 adult outpatients with psychotic disorders. No hypothesis was stated in the published abstracts, but given the trial's sponsorship by Zeneca Pharmaceuticals, it seems likely that an alternative to the null hypothesis—namely, that quetiapine would outperform risperidone—was being tested.

As of this writing, QUEST has not been published, although results have been presented at a number of conferences. No statistical information has yet been published. Statistical power, α and β levels of significance, and the nature of statistical analyses (1- or 2-tailed) are unknown.

CONLEY ET AL.

Conley et al.[5] compared rehospitalization rates between patients treated with atypical antipsychotics and those treated with depot antipsychotics in 368 patients discharged from Maryland State Mental Health facilities between January 1, 1997, and December 31, 1997. The rehospitalization rates of patients discharged on risperidone, olanzapine, and clozapine were, respectively, 12% (N = 156), 13% (N = 109), and 14% (N = 49). The rehospitalization rate for those on the decanoate forms of haloperidol or fluphenazine therapy was 34% (N = 58).

This study included no a priori hypothesis. The study was funded in part by a nonprofit institution, the Theodore and Vada Stanley Foundation. As of this writing, the study is unpublished, and little statistical information has been reported.

ALTERNATIVES TO HYPOTHESIS TESTING

Hypothesis testing is the method of assessing the success of a clinical trial, and it is the one employed most often in the medical literature.[12] Nonetheless, only 2 of the 5 trials covered here included the explicit statement of a hypothesis. Among the guidelines offered by Lang and Secic is "If the study was designed to test one or more a priori hypotheses, state the hypotheses."[13(p5)]

Despite its popularity, hypothesis testing has many critics, and alternatives have been proposed. The International Committee of Medical Journal Editors, in its "Uniform Requirements for Manuscripts Submitted to Biomedical Journals," recommends that researchers "avoid relying solely on statistical hypothesis testing, such as the use of p values, which fails to convey important quantitative information.[14(p312)] The Committee also recommends, "When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals)."[14(p312)] Verdoux and Salamon[12] note the limitations of hypothesis testing and significance, including the fact that the arbitrary nature of the .05 α level is often forgotten, with statistical or biological relevance being confused with clinical significance as a result. The p value fails to convey the strength of the association and its direction or the variability in the sample precisely. As a result, some experts promote a method based on estimation and confidence interval, allowing the direction of the association and the effect size (i.e., the strength of the association) to be known. The variability of the estimation in the sample can be determined by the width of the confidence interval; the narrower the confidence interval, the more precise the estimation. The clinical relevance of the findings is easier to infer as a result.[12]

Others have taken up this call. Borenstein[15] agrees that p value is often mistakenly regarded as an indication of clinical significance rather than statistical significance, the latter having little to do with the former. Using confidence intervals to report results has had the concrete effect of shifting the focus from what the effect is not to what the effect is.[15] Glaser, addressing the controversy over null hypothesis significance testing, decries "the lack of accom-

panying statistics, such as effect sizes and confidence ntervals, that would provide a broader picture into the researcher's data analysis and interpretation."[16(p291)] Gardner and Altman, who have long advocated the use of confidence intervals in medical reporting, agree that medical investigators are generally more concerned with finding the extent to which a measured outcome differs between groups than with stating whether that difference is significant.[17]

CONCLUSION

Hypotheses too often are unreported and left for the reader to infer or guess. Hypotheses should be clearly stated in reports of clinical trials, and the statistical information necessary for the reader to evaluate the trial must also be included. Hypothesis testing and its emphasis on p values and significance may be replaced by or complemented with other methods of reporting clinical trials, but it currently remains the most widely used method of evaluating a clinical trial. As a consequence, reporting hypotheses and the statistical information necessary to assess clinical trials is paramount. In general, more rigorously designed trials to evaluate the efficacy, safety, and effectiveness of atypical drugs compared to conventional drugs and to each other are needed.

Drug names: clozapine (Clozaril and others), haloperidol (Haldol and others), olanzapine (Zyprexa), risperidone (Risperdal), quetiapine (Seroquel).

Disclosure of off-label usage: The author has determined that, to the best of his 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. Lang TA, Secic M. Comparing groups, I: reporting p values. In: How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers. Philadelphia, Pa: American College of Physicians; 1997:65–80
7. 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
8. 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
9. Rigby AS. Statistical methods in epidemiology, I: statistical errors in hypothesis testing. Disabil Rehabil 1998;20:121–126
10. Fox N, Mathers N. Empowering research: statistical power in general practice research. Fam Pract 1997;14:324–329
11. Lang TA, Secic M. Differences between clinical and statistical significance. In: How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers. Philadelphia, Pa: American College of Physicians; 1997:xxiii–xxv
12. Verdoux H, Salamon R. Resultats statistiques: quel mode de presentation choisir? [Statistical results: which method of presentation to choose?] Encephale 1997;23:19–21
13. Lang TA, Secic M. Asking questions and finding answers: reporting research designs and activities. In: How to Report Statistics in Medicine: An notated Guidelines for Authors, Editors, and Reviewers. Philadelphia, Pa: American College of Physicians; 1997:3–37
14. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. N Engl J Med 1997;336:309–315
15. Borenstein M. The case for confidence intervals in controlled clinical trials. Control Clin Trials 1994;15:411–428
16. Glaser DN. The controversy of significance testing: misconceptions and alternatives. Am J Crit Care 1999;8:291–296
17. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ 1986;292:746–750