Increasing the Power of Clinical Trials Through Judgment Analysis

THOMAS R. STEWART, PhD, C. R. B. JOYCE, PhD

A method for increasing the statistical power of clinical trials to detect clinically important differences is described. Inconsistency in physicians' overall judgments of treatment effectiveness adds "noise" to a trial that may mask either the superiority or the inferiority of particular treatments. The method described here uses "judgment analysis" to reduce errors in an individual's overall judgments of treatment effectiveness. The method can also be used to reduce error variance due to differences in judgment between physicians and may thus be particularly useful in multicenter trials. The method is illustrated with results from a recent trial. Key words: statistical power; experimental design; multi-investigator trials; judgment analysis; clinical judgment. (Med Decis Making 8:33–38, 1988)

In clinical trials, several different measures of treatment effects are often necessary to record the physiologic or psychological changes relevant to the course of the disease, the action of the treatment, or the management of the patient. As a result, there is no single objective indicator of the overall effect of the treatment, and there are several sets of results to be combined, and perhaps transformed, to enable the physician to decide whether the treatment received by each patient modified the outcome.

Even simple observations such as the reading of an oscillating sphygmomanometer value frequently contain a judgmental element, and the more complex ones that come from the autopsy room or microbiology laboratory may depend heavily upon judgment. Nevertheless, such information must be combined by the physician with other material available in memory or on other forms of data base to arrive at a diagnosis, to choose a treatment, and to evaluate the outcome. All these activities involve judgment, an activity in which the behavior of individuals varies from occasion to occasion as well as among themselves.

Although separate analysis of single measures may be valuable, the participating physician's (or other observer's) judgment is often the only basis for estimating a treatment's overall effect on the patient. As a result, it is common to find that the key dependent variable in the analysis of the results of a clinical trial is based on one or more observers' judgments of the treatment effects.

Reliance on observers' judgments of treatment effects can be a source of error variance because the reliability of judgments is limited and because, in trials involving more than one observer, judgments may vary across observers. Such intra- and inter-individual differences increase error in the measurement of treatment effects, and the result is a decreased power to detect differences that are potentially clinically important. See Cohen for a discussion of statistical power and Stewart et al. for examples of the application of power analysis to clinical trials.

Despite their importance, however, inter- and intra-individual differences in observers' judgments are seldom looked at. Typically, the error due to differences between observers, if recognized at all, has at best been controlled, to an uncertain extent, by holding training sessions and/or discussions between the participants before and perhaps during the trial. This may be largely due to the frequent assumption (not only by laypeople) that the training of the physician results in such reliable and objective performance that no account need be taken of variations resulting from differences in education and experience.

The importance of the judgment problem increases when, as is more and more frequently the case today, it becomes necessary, for a variety of reasons (pressure of time, demands of national drug regulatory authorities, comparative rarity of the disease in the practice of any one physician), to resort to multicenter or international clinical trials, which may involve the participation of doctors having very different backgrounds.
and medical training. Even within a single center, it is very common for more than one investigator to be involved—whether explicitly or not.

Judgment error, i.e., unreliability within an observer, on the other hand, is almost invariably lumped in with residual error (which, of course, it inflates), so the statistical power is reduced. Although statistical significance is less relevant to medical practice than clinical importance, the judgment problem is not solved by ignoring it.

The neglect of power in trial design is frequently complained of by statisticians and others. For example, Ward recently found that 12 of the 15 randomized double-blind trials in acute severe asthma identified from Index Medicus between 1974 and 1984 had a less than 60% chance of detecting a true 25% difference between treatments. Indeed, the trial that demonstrated the largest difference was unable to show that it was statistically significant, having a less than 30% chance of showing a 50% difference. Evidence also suggests that many trials have insufficient power to detect clinically important differences.

For a desired level of protection against type I error, the statistical power of a clinical trial depends upon the population difference between treatment means, the population error variance, and the sample size. If power could be improved only by increasing sample size, a longer and presumably more expensive trial would be needed. Error variance can be controlled, to some extent, by careful experimental design and measurement procedures. We propose a method for further reducing error variance by using judgment analysis to eliminate the unreliability of judgment. This method can provide an additional increase in power beyond that afforded by careful design and conduct of the trial, and it can do so at less cost than would be required to increase the sample size.

**Reducing Error Variance by Applying Judgment Theory**

The physician attends to a number of different factors ("cues") in judging the overall effectiveness of a drug. These cues are drawn from all the effects associated with the treatment that are of interest in the given trial. Physicians will be less than perfectly reliable in combining these cues and making their judgments. This unreliability is a result both of the complexity of the task itself and of the consistency with which the judge approaches it. The effect of unreliability is to add error variance or "noise" to the systematic variance of the physician's judgments.

The power of a trial (defined as 1 minus the probability of a type II error) can be improved by reducing the estimate of the population variance \( \sigma^2 \) instead of increasing the sample size. When the dependent variable is a judgment of effectiveness or tolerability, the removal of judgmental "noise" will increase the power of the test.

Theory and research on judgment offer a method for doing this based on the finding that simple, additive models can adequately describe, in a variety of contexts, the component of judgment that is systematically related to the cues. Multiple regression analysis, regressing the physician's judgments on the cues, can therefore be used to separate reliable variance (i.e., that which is systematically related to the cues) from error variance. The use of multiple regression analysis for this purpose ("judgment analysis") has been discussed.

In judgment analysis (JA), the magnitude of the error component is estimated by \( \sqrt{1 - R^2} \), where \( R^2 \) is the squared multiple correlation coefficient. Removal of the error component leaves only the reliable portion of the judgmental variance (s\(_j^2\))—rather than the total judgmental variance (s\(^2\))—as an estimate of the population variance \( \sigma^2 \). Since s\(_j^2 = R^2 \sigma^2 \) and 0.0 \( \leq R^2 \leq 1.0 \), use of only the reliable portion of the judgmental variance must result in a lower estimate of \( \sigma^2 \). Consequently, the sample size required to achieve error probabilities \( \alpha, \beta \) for a specific alternative hypothesis is reduced. Alternatively, for a given sample size, type I error probability \( \alpha \) and a specific alternative hypothesis, the power of the clinical trial will increase.

This method is valid if the regression model for the judge is correct, so that only random error variance is discarded. It has been shown for estimation of the severity of a medical condition and evaluation of its treatments that a linear model is usually adequate, but care must be taken in any application to avoid pitfalls, such as overfitting data from a given sample of judgments, that can lead to incorrect models and erroneous results. Methods for evaluating judgment models are discussed below.

**Differences within Observers**

The effect of using JA to reduce errors of judgment is illustrated using data derived from a clinical trial designed to compare the efficacy of two dose levels of a new analgesic with that of an older remedy, and that of a placebo, in patients suffering from painful osteoarthrosis of the hip and/or knee. The multicenter, double-blind trial involved five physicians who administered an active treatment or placebo, in approximately equal numbers, to between 23 and 47 patients each. Patient records were completed at the time of the initial treatment, a week later, and two weeks later. On the final report, the physicians recorded a judgment of overall treatment effectiveness on a scale ranging from 0—none, no good at all, ineffective drug; to 4—excellent, ideal, best possible effect.

A traditional one-way ANOVA based on the complete
A sample of 167 patients revealed no statistically significant differences among judgments of overall treatment effectiveness of the three active treatments (although these differed significantly from that of the placebo). JA was applied to each physician's judgments of therapeutic effectiveness to determine whether inconsistency in these judgments might have masked a significant difference among treatments.

Seven cues were derived from the patient records. The rating of each cue was determined by the difference between an assessment recorded on the final visit and the same assessment recorded on the initial visit. Assessments used were severity of joint pain on full passive movement, severity of pain while walking, average duration of starting problems on arising, severity of starting problems on arising, status of arthritic condition, discomfort or inconvenience caused by the treatment, and patient's reported sense of well-being. These seven cues were used in multiple linear regression analyses to predict physicians' judgments of therapeutic effectiveness. Resulting adjusted R²'s for the five physicians ranged from 0.774 to 0.906. These values approach the reliabilities typically found in judgment studies, suggesting that the seven cues were sufficient to model the reliable variance. For this example, the (testable) assumption is made that valid judgment models were obtained for each judge.

The regression model (consisting of weights for each of the seven cues and a constant term) for each physician was used to compute judgments for each patient seen by that physician. These computed judgments are systematically related to the cues by a function extracted from the physician's raw judgments. Their reliability is limited only by the reliability of the cues, and they exclude the inconsistencies that introduce random error into the physician's judgments. The computed judgments were substituted for actual judgments of therapeutic effectiveness as the dependent variable in a one-way ANOVA of the three active treatments. Two important results emerged. First, the within-groups mean square error, which is the basis for the estimate of population error used in significance tests, was reduced by 11%, from 1.027 to 0.912. This reduction means increased power to detect reliable differences among treatments without increasing sample size or increasing the probability of type 1 error.

Table 2 illustrates one consequence of the reduction in error variance that resulted from JA. The table compares 95% confidence intervals for the differences between treatment means based on raw judgments with the same confidence interval based on computed judgments. In each case, the use of computed judgments narrows the confidence interval, thus improving the precision of the estimate of the treatment means.

In this example, the application of JA did not result in statistically significant differences between treatments. The lack of statistical significance was therefore not solely a result of unreliability in the physicians' judgments, which were in fact fairly reliable. The use of JA increased confidence in the failure to reject the null hypothesis. In other trials, however, JA, by eliminating unreliability, may unmask a significant difference between treatments.

To summarize, the following procedure can reduce error variance and improve the power of clinical trials in which the outcome is assessed by physicians' judgments of individual patient benefit.

1. Identify the important effects that are used as cues to overall treatment effectiveness. This should of course be done during the design of the trial.

### Table 1

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Raw Judgment</th>
<th>Computed Judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 1</td>
<td>1.476</td>
<td>1.476</td>
</tr>
<tr>
<td>Treatment 2</td>
<td>1.674</td>
<td>1.680</td>
</tr>
<tr>
<td>Treatment 3</td>
<td>1.625</td>
<td>1.625</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Dependent Variable</th>
<th>95% Confidence Interval for Difference Between Treatment Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 1 vs. Treatment 2</td>
<td>Raw judgment of therapeutic effectiveness</td>
<td>−2.28 to 1.89</td>
</tr>
<tr>
<td>Treatment 1 vs. Treatment 3</td>
<td>Raw judgment of therapeutic effectiveness</td>
<td>−2.03 to 1.74</td>
</tr>
<tr>
<td>Treatment 2 vs. Treatment 3</td>
<td>Computed judgment</td>
<td>−1.94 to 1.64</td>
</tr>
</tbody>
</table>

*Cues are of course based upon observations of one kind or another and judgments inevitably enter into those observations. However, variation in cues due to judgment error is ignored in the present discussion, although it can be studied in a similar way.
2. Use JA to describe the policy of each judge. If necessary (e.g., if the number of patients seen by each physician is insufficient) "paper patient" data derived from a similar population may be used for JA. Carefully examine the validity of the JA results.

3. Replace the judgments of each physician with those based upon his or her computed policy.

4. Use standard statistical tests with the computed judgments as the dependent variable.

This method will indicate whether the inconsistencies of individual judges are hiding a significant difference between treatments.

Differences Between Observers

In addition to helping to reduce the problems caused by inconsistencies of judgment within physicians, as illustrated above, JA can be useful in reducing problems caused by differences between physicians in multicenter or other multi-investigator trials.

Suppose that several physicians make (independent) judgments of effectiveness and that these ratings are based on the same patients.

In general, it will turn out that some physicians will find the difference in the effects of the treatments to be greater than will others; a t-test for treatment 1 versus treatment 2 may even be significant for the former and nonsignificant for the latter. Since all physicians judged the same patients, it can be surmised that the differences occurred because the former attached more importance to those cues for which the treatments differed most, while the latter put higher weights on those cues for which the treatments differed least.

If differential judgment policies produce different test results when applied to the same set of patients, the problem is likely to be even more serious when different policies are applied to different sets of patients, as is almost always the case in real-life trials. In this situation, it is not known whether differences in the physicians' judgmental policies or lack of generality in the treatment effects leads to inconsistent or even paradoxical results. Large differences among physicians will deprive the statistical test of power to discriminate all but the most substantial differences in treatment effects. In other words, since the power of the test against a modest difference is too low, it will frequently be concluded that there is no difference between treatments when in fact there is. The same will be true for judgments about tolerability, although this will be harder to demonstrate quantitatively because adverse effects of dosages that have reached the stage of clinical investigation are much rarer than desirable effects. However, the consequence in either case may be the loss of potentially useful treatments or, conversely, the acceptance of useless ones, due to the commission of type II error. This is known to occur, although the frequency with which it does so is unknown. An important part of the known attenuation between synthesis and marketplace of drugs in development of some 4,000:1 or even more is likely to be brought about in this way.

The problem of multiple judgment policies may be resolved by generalizing the approach to the problem of unreliability of judgment described above. Depending on the nature of the problem in a particular trial, one or more of the following steps may prove useful:

1. Examine and compare the judgment policies of the separate judges. Are the patterns of regression weights similar or different?

2. Compute judgments for each physician by applying his or her policy to all patients (i.e., including those seen by the other physicians). Examine correlations among these computed judgments. The correlations may be high even if the policies are different because the cues are intercorrelated. If the correlations are low, then differences among judges may have substantially inflated the error variance.

3. Compute statistical significance tests using as dependent variables, in turn, a) each cue singly, b) the computed judgments (for all patients) based on each judge's policy, c) an average, over judges, of the computed judgments, and d) computed judgments based on a policy that applies equal weight to each of the cues. Examination of these tests will indicate which, if any, of a variety of models for combining the cues yields significant differences between treatments.

4. If the results of all the above tests are negative, calculate a multiple group discriminant analysis to determine whether any linear combination of the cues is capable of discriminating among treatment effects.

5. If the results of the above steps indicate that some policies produce significant differences while others do not, then a "best" policy might be chosen. The policy chosen will depend on the purpose of the trial and the nature of the differences among policies. Its choice involves substantive, as well as methodologic, considerations. In general, the policy should be chosen in consultation with the participating physicians and, as before, during the planning period rather than a posteriori (or, at the latest, after an initial pilot phase). In any event, the goal must be to increase the validity of the trial, and not just to obtain statistically significant differences.
Although JA can help to develop a single policy that can be applied to observations on all patients, only the judges themselves can reconcile their real differences. Once a joint policy has been established, by whatever means, it can be applied to the cue reports from all investigators or centers. Computed judgments derived from the application of the joint policy can then be used in the appropriate statistical test for analysis.

The solution to the problem of differences among judges is complex. It depends on statistical properties, such as cue and judgment means and intercorrelations, of the information collected from the trial. It also depends on the context and design of the trial and on the nature of the desired treatment effect.

**Limitations of Judgment Analysis**

The success of any application of JA depends on producing a correct model of judgment. If an incorrect model is used, residual variance from the regression may include not only error variance but also a portion of the reliable variance of judgment. Discarding such variance can decrease the power of the experiment and produce misleading results.

Incorrect models may result from 1) using the wrong set of cues, e.g., overlooking one or more cues that are important to the observer, 2) using the wrong functional relations between the cues and the judgment, e.g., using linear relations when nonlinear relations are important, or 3) overfitting the data by having too few observations relative to the number of cues. A complete discussion of methods for avoiding these problems is beyond the scope of this paper. Following is a brief description of some of the critical tests that should be used to assess the adequacy of a JA model.

*The model should be crossvalidated.* The major test for overfitting is provided by crossvalidation, that is, applying the model to a set of judgments that were not used to fit the model.

*Alternative models should be compared.* The most appropriate functional relations can be identified by fitting, and crossvalidating, alternative models and comparing the results.

*Treatment means computed on the residuals should not differ.* If the residuals represent the unreliability, or error variance of judgment, then they should not be systematically related to any other variable. In particular, they should not be related to the treatments in a clinical trial. A significant difference across treatments would be evidence that some of the reliable variance of judgment is contained in the residuals. A corollary is that treatment means computed on the modeled judgments should not differ from those computed on the raw judgments.

Even if the judgment model is correct, the potential effect of reducing error variance by using JA depends on the obtained difference between treatment means and on the reliability of the observers. The application of JA will not change the difference between treatment means, and hence cannot cause a difference of a clinically important size to appear if it is not present. The potential benefit of the method will be realized to the extent that the observers' unreliability reduces the power of the trial to detect true differences.

It is obvious that JA can be used only when the cues for judgment are known and measured in the trial. But if this is the case, then an independent investigator could dispense with the physician's judgment altogether (though not his or her observations) and analyze the individual measures or some combination of them. What, then, is the advantage of the method described in this paper? If there is an agreed way to combine the cues to get a measure of overall treatment effects, then there is no need to apply JA. If, on the other hand, the investigator is unsure (or multiple investigators disagree) about which cues should be included in a composite and how they should be weighted and combined (as is often the case) then JA helps to develop an appropriate composite. It also provides a method for validating the composite against the physician's own judgment.

**Conclusion**

The application of JA to variation in the judgmental component of observation has not been attempted here. Instead, observations have been assumed to be both reliable and valid. Such assumptions are at least implicit in most clinical research. However, the assumption that judgments are also consistent and valid is usually even more deeply buried, and has been the object of the present study.

A first practical demonstration of the potential utility of the method proposed in this paper to clinical trials has been provided by Bech et al., who applied a variety of policies computed from the judgments of 26 psychiatrists participating in a comparative trial of two antidepressants. They were able to demonstrate a sharper and earlier differentiation between the treatments, as predicted.

In regard to multi-center or multi-investigator trials, judgmental disagreements can obscure differences in the application of inclusion and exclusion criteria as well as between evaluations of outcome and can also lead to inconsistent inferences from the results—in short, the likelihood of committing one or more type II errors will be increased if judgmental inconsistency is ignored. The application of a single, explicit, agreed (and especially, an appropriate) policy to all observa-
tions, on the other hand, will increase the power of the trial. This may be reflected in a reduction of type II error likelihood, or in economies in the size of patient and/or investigator samples needed to detect true treatment differences.

A cost is associated with the use of JA: the design of the trial and the analysis of the results will be more complicated. As is the case with any method, new or old, regardless of its fundamental value or soundness, there is also a risk that it will be naively or carelessly applied. JA requires study to determine its most appropriate uses and its value in practice. It may prove to be highly cost-effective in increasing the power of experiments.

The authors thank Tamra Pearson for assistance in analyzing the data presented in this paper, and Mary Luhring and Teresa Harnisch for typing several drafts of the manuscript. Per Bech, Mark Chaput De Saintonge, Ray Cooksey, Alvan Feinstein, Hans-Ueli Fisch, Austin Heady, and John Kirwan provided helpful comments on an earlier draft.

References

15. Ward MJ: Clinical trials in acute severe asthma: are type II errors important? Thorax 41:824–829, 1986