Statistics Series

Statistical Considerations for Performing Multiple Tests in a Single Experiment. 1. Introduction

PETER C. O'BRIEN, Ph.D., *Section of Biostatistics*; MARC A. SHAMPO, Ph.D., *Section of Publications*

In a previous series of articles,[1] we described elementary statistical concepts and procedures. Those discussions focused on methods for describing a specific data set and for making inferences to larger populations. The part on inferential statistics described some of the most commonly used statistical tests; on occasion, somewhat unrealistic, simplified assumptions were made in order to maintain clarity. One important assumption was that studies were conducted to test only one hypothesis. In practice, of course, investigators often want to test several different (although possibly related) hypotheses. For example, when comparing the efficacy of an experimental therapy with that of a placebo, an investigator may want to include several treatment options in the same experiment rather than conduct multiple experiments, each with its own group of controls (patients given a placebo). Another example is an experiment designed to compare multiple endpoints (for example, multiple measures of efficacy and safety) between two therapeutic modalities.

When studies such as these are reported in the medical literature, authors often refer to multiple comparison procedures (many are available), adjusted *P* values, Bonferroni adjustment, and per-experiment error rates. In this new series of articles (intended to supplement but not replace the earlier series), we explain the meaning of these terms and describe the statistical methods that are often used when multiple comparisons are made within the context of a single experiment.

Some of the topics that we discuss are controversial, even among statisticians. Fortunately, however, the arguments involved are, without exception, nontechnical. Because the stakes are often high (whether the results of a study are to be regarded as statistically significant is often at issue), it should be well worth the reader's time to become familiar with the issues and to decide for oneself the correct approach to the answer.

Before pursuing the topic of performing multiple statistical tests, we review how an investigator uses statistics to test hypotheses. The basic approach is as follows: one formulates a null hypothesis (for example, no difference in efficacy exists between two drugs), computes an appropriate test statistic and corresponding *P* value, and rejects the null hypothesis only if the *P* value is sufficiently small (for example, $P<0.05$). The *P* value is actually shorthand for the following statement: "If the null hypothesis is true, then the probability of observing a value of the test statistic as large as the value observed is equal to *P*."

Specifically, an investigator who wants to compare an experimental drug therapy with a standard (or placebo) therapy might administer each therapy to each patient in a random sequence and in a double-blind fashion so that neither the patient nor the physician is aware of which preparation is being administered. Using a paired *t* test, one could then test the null hypothesis of no difference between the two modalities. (In this example, as well as elsewhere throughout this series of articles, we will not address the important considerations of study design, which would have to be carefully considered in any actual study. Our purpose will be simplicity in order to

Individual reprints of this article are not available. The entire six-part series will be available for purchase as a bound booklet from the *Proceedings* Circulation Office in December.

enable us to focus on elementary concepts involved in the data analysis.)

Suppose, however, that the investigator also wants to evaluate nine other experimental drugs (agents B, C, D, E, F, G, H, I, and J). Is it still permissible for drug A to be compared with a standard? If so, is the *t* test, and the corresponding *P* value, still appropriate? Many statisticians argue that a direct comparison of drug A with standard therapy that ignores the other nine drugs is no longer permissible in this situation. Others argue that if the usual *t* test is performed, the resulting *P* value should be multiplied by 10 because 10 tests will ultimately be performed. Still others maintain that the analysis should not be altered because of inclusion of nine additional drugs in the study. Obviously, the position one takes in this controversy has a profound influence on the conclusions drawn about the efficacy of drug A. Fortunately, the principles that will ultimately determine the appropriate answer to this controversy are not mathematical. They center on an understanding of two types of error rates: the per-comparison error rate and the per-experiment error rate.

PER-COMPARISON AND PER-EXPERIMENT ERROR RATES

In a comparison of drug A with standard therapy in the preceding example, the experimenter may decide that drug A is superior only if the *P* value derived from the paired *t* test is less than 0.05 ($P<0.05$). Because the probability of making this decision when the null hypothesis is true (drug A is not superior) is 0.05, the probability of erroneously concluding that drug A is superior is 0.05. In statistics, this is called a type I error. Thus, the probability of a type I error is 0.05. (Another type of error—a type II error—can occur if one fails to identify superiority when drug A actually is better.) By concluding that a drug is superior only when the *P* value is less than 0.05, the investigator is confident that drugs will erroneously be declared superior only 5% of the time. Still smaller error rates can be achieved by requiring a smaller *P* value as the criterion (for example, $P<0.01$).

Similarly, the investigator may compare each drug separately (A through J) with a standard therapy by using the paired *t* test, in each case declaring superiority only if the corresponding *P* value is less than 0.05. The error rate for each such comparison, the *per-comparison error rate*, is 0.05.

Obviously, for an evaluation of the results of a study in which one or more drugs are compared with a standard therapy, knowing the per-comparison error rate is important. A second error rate is also of interest, however. Specifically, one may ask, If the efficacies of all the drugs in the entire experiment are the same, what is the probability that one or more of the drugs will erroneously be declared superior to the standard therapy? Statisticians refer to this error rate as the *per-experiment error rate*. Suppose, for example, that 100 different drugs, none of which is efficacious, are compared with a placebo. With a per-comparison error rate of 0.05, one would expect that five of the drugs will be erroneously declared efficacious. Clearly, in some studies it may also be important to know the per-experiment error rate.

CONCLUSION

The question that inevitably arises when the results of multiple tests from a single experiment are available is, Which error rate should I report, the per-comparison error rate or the per-experiment error rate? This question has been the source of much controversy, even among statisticians. We believe that much of the controversy is in large part due to the way in which the question has been posed. As stated, it suggests that one error rate is correct or relevant and the other is incorrect or irrelevant. This fallacy is further reinforced by a tendency to ask, What is *the P* value?—a further implication that only *one* error rate can and should be quoted.

On the contrary, we hope that the preceding discussion clarifies that both the per-comparison error rate and the per-experiment error rate provide qualitatively different types of information. In any given study, both types of information may be relevant and worth presenting. The relative importance of each error rate in a specific study will depend on the particular circumstances, especially the goals, of the study.

In the remaining articles of this series, we will consider some of the most common types of studies in which multiple tests are performed, using examples from studies done by researchers at the Mayo Clinic. In each case, we will discuss the appropriateness of the two types of error rates and the corresponding statistical techniques.

The discussion in part 2, "Comparisons Among Several Therapies," considers the problem of performing all possible pairwise comparisons. Two types of multiple comparison procedures are described: those that are based on a pooled estimate of error and those that are not. The statistical concepts are illustrated with a study in which several analgesics are compared with placebo and aspirin.

For the type of study described in part 3, "Repeated Measures Over Time," measurements are made on the same subjects at several fixed points in time (for example, at 30-minute intervals) after the administration of therapy. Statistical techniques are described for addressing questions such as, Is a treatment effect present? Are the changes that occur after treatment constant over time, or are they accelerating (or decelerating) over time? At what time does the treatment effect begin, reach a peak, and end? The example used evaluates the effects of head-down neck flexion on blood flow in the calf and forearm.

Part 4, "Performing Multiple Statistical Tests on the Same Data," considers the situation in which one statistical test indicates that a difference between groups is not statistically significant but another test indicates that the difference is significant. The example used is based on a randomized trial in which plasmapheresis was compared with sham-pheresis for the treatment of chronic inflammatory demyelinating polyradiculoneuropathy.

In part 5, "Comparing Two Therapies With Respect to Several Endpoints," a study that identifies individual endpoints for which one type of therapy may be more efficacious than the other is distinguished from a study in which a single overall test for efficacy is desired on the basis of the cumulative results of several measures of efficacy. A randomized trial comparing two approaches for treating diabetes is used as an example.

The problem in part 6, "Testing Accumulating Data Repeatedly Over Time," is concerned with comparing two groups serially over time during the course of a clinical trial, with a view toward terminating the trial early if one treatment is observed to be much superior to the other. The first example used is a randomized trial in which prednisone therapy was compared with prednisone plus vincristine for the treatment of leukemia. A second example compares two regimens of chemotherapy for extensive small cell lung cancer.

REFERENCE

1. O'Brien PC, Shampo MA: Statistics for clinicians. Mayo Clin Proc 56:45-46; 47-49; 126-128; 196-197; 274-276; 324-326; 393-394; 452-454; 513-515; 573-575; 639-640; 709-711; 753-754; 755-756, 1981

Statistical Considerations for Performing Multiple Tests in a Single Experiment. 2. Comparisons Among Several Therapies

PETER C. O'BRIEN, Ph.D., *Section of Biostatistics*; MARC A. SHAMPO, Ph.D., *Section of Publications*

In this article, we discuss the statistical issues that arise when one makes pairwise comparisons among several groups within the context of a single experiment. Several techniques for controlling the per-experiment error rate are described. These techniques are illustrated with examples that demonstrate how the usefulness of per-experiment and per-comparison error rates depends on the circumstances surrounding a particular study.

STATISTICAL TECHNIQUES FOR CONTROLLING THE PER-EXPERIMENT ERROR RATE

Overall Preliminary Test.—With use of an overall preliminary test, a null hypothesis is established specifying that no difference exists among any of the groups. If 10 drugs are being studied, for example, a test statistic (referred to as *F*, analogous to *t* in the *t* test for comparing two therapies) is derived by dividing a measure of the variability among the 10 group means by a measure of the variability expected by chance. One obtains the corresponding *P* value from suitable tables (the larger the *F* statistic, the smaller the *P* value) and rejects the null hypothesis of no difference among treatments if *P* is sufficiently small (for example, less than 0.05).

With this approach to controlling the per-experiment error rate, one adopts the convention that pairwise comparisons of individual treatments will be pursued only if the *P* value from the preliminary *F* test is less than 0.05. If pairwise comparisons are warranted, they may be done in two ways: (1) by performing separate *t* tests in the usual way or (2) by using modified *t* tests. The latter approach is called the least significant difference (LSD) method. It is similar to the *t* test, but it relies on the assumption that the variability is the same for all therapeutic modalities. (For further details, see the Appendix.)

Regardless of which method is used, the per-experiment error rate will be less than 0.05 because pairwise comparisons will be done only 5% of the time when no real differences exist among any of the therapies (because the initial *F* test will be significant only 5% of the time if no differences exist). Conversely, the per-comparison error rate will be somewhat less than the *P* value obtained from the *t* tests.

Bonferroni Adjustment.—Another method for controlling the per-experiment error rate is to perform *t* tests in the usual way but multiply each *P* value by the number of comparisons undertaken to obtain adjusted *P* values, called the Bonferroni adjustment. Treatments are judged to be significant only if the adjusted *P* value is less than 0.05; thus, the corresponding per-experiment error rate will be less than 0.05. Notice that the adjusted *P* values provide per-experiment, not per-comparison, error rates. The per-comparison error rates are indicated by the original (unadjusted) *P* values.

Informal Adjustment in Interpretation of* P *Values.—One source of dissatisfaction with the Bonferroni method is that it provides only an approximation of the per-experiment error rate,

Individual reprints of this article are not available. The entire six-part series will be available for purchase as a bound booklet from the *Proceedings* Circulation Office in December.

and sometimes the accuracy of the approximation is unsatisfactory. The true per-experiment error rate may be considerably less than the Bonferroni approximation. A second concern is that in some applications one may want to focus more on the per-comparison error rate while concurrently acknowledging the per-experiment error rate. This reasoning suggests quoting the usual P values (obtained from separate paired t tests) but being appropriately cautious about drawing conclusions. For example, one may consider only P values less than 0.01 as providing convincing evidence of a difference.

Other Multiple Comparison Procedures.— Other available procedures focus solely on the per-experiment error rate. The commonly used procedures are the Student-Newman-Keuls (SNK) procedure, Tukey's honestly significant difference (HSD) method, and Duncan's multiple-range (D) test. All these procedures are based on the same test statistic as the LSD procedure; however, instead of comparing the test statistic to tables of the t distribution, special tables are used that provide accurate control over the per-experiment error rate.

These procedures enable the investigator to make all possible pairwise comparisons in such a way that the probability of any treatment being declared significant when all therapies are equivalent (the per-experiment error rate) is less than or equal to a specified level. Conversely, none of these methods provides information on the per-comparison error rate. These types of procedures are discussed by Bancroft.[1]

Dunnett's Procedure.—Dunnett's procedure[2,3] is designed specifically for comparing several experimental therapies to a single standard therapy. It is similar to the SNK, HSD, and D procedures, except that comparisons are made only between the experimental therapies and the standard treatment.

EXAMPLE

We will consider a double-blind crossover study in which nine marketed analgesics and a placebo were evaluated in 57 patients with definite pain problems as a result of unresectable cancer.[4] Although all possible pairwise comparisons were of interest, only comparisons with aspirin and placebo were reported. Thus, the total number of comparisons was 17. If no difference in efficacy was detected among any of the preparations, separate paired t tests performed at the 0.05 level would probably show significant differences due to chance alone. Because the investigators wanted to be cautious and were reluctant to claim erroneously that a preparation was superior to placebo or inferior to aspirin, efforts to control the per-experiment error rate were deemed necessary.

Of the various analyses performed, the results of analysis of the mean percentage of relief of pain achieved among the 57 patients are listed in Table 2-1. Note that the P value reported in Table 2-1 is the per-experiment error rate obtained by using the SNK procedure. Thus, the probability of reporting any significant differences in this analysis would be less than 0.05 if no differences existed among all the preparations.

In addition to evaluating percentage of relief of pain, patients also were asked to rank the analgesics on the basis of relative efficacy. The analysis of these data (Table 2-2) was based on separate paired t tests; thus, the P values reported in Table 2-2 are per-comparison error rates. Because of the many comparisons undertaken, however, evidence of a difference was judged to be convincing only if the P value was impressively small—that is, less than 0.01. Although different data and different statistical tests were used for the analyses in Tables 2-1 and 2-2, the results were remarkably similar.

Table 2-1.—**Relative Therapeutic Effects of Orally Administered Analgesics Based on Mean Percentage of Relief of Pain Achieved in 57 Patients**

Analgesic	Dose (mg)	Relief of pain (%)	
Aspirin	650	62	Significantly superior to placebo (P<0.05)†
Pentazocine*	50	54	
Acetaminophen	650	50	
Phenacetin	650	48	
Mefenamic acid	250	47	
Codeine	65	46	
Propoxyphene	65	43	Significantly inferior to aspirin (P<0.05)†
Ethoheptazine	75	38	
Promazine	25	37	
Placebo	...	32	

*Results are reported for 30 patients (statistical significance was calculated on the basis of patients receiving pentazocine with use of Dunnett's procedure for multiple comparisons with control).

†Student-Newman-Keuls method.

From Moertel and associates.[4] By permission of the Massachusetts Medical Society.

Table 2-2.—Relative Therapeutic Effects of Orally Administered Analgesics Based on Sum of Ranks Accorded by Each Patient

Analgesic*	Dose (mg)	Rank sum	
Aspirin	650	223.0	Significantly superior to placebo ($P<0.01$)†
Mefenamic acid	250	271.5	
Phenacetin	650	275.0	
Acetaminophen	650	280.5	
Codeine	65	284.5	
Propoxyphene	65	315.0	Significantly inferior to aspirin ($P<0.01$)†
Ethoheptazine	75	335.0	
Promazine	25	352.5	
Placebo	...	374.0	

*In 30 patients, pentazocine (50 mg) was in fifth position and significantly superior to placebo ($P<0.01$).

†Analysis by t test.

From Moertel and associates.[4] By permission of the Massachusetts Medical Society.

ALTERNATIVE EXAMPLE

Notice that the purpose of the study in the example determined the precise formulation of the study questions and the corresponding data analysis used to answer them. Suppose, however, that the same study had been performed for a different purpose and had addressed somewhat different questions. Specifically, suppose that the manufacturer of propoxyphene (Darvon) had performed this study and that the specific aims of the study had been to evaluate propoxyphene (1) relative to aspirin, (2) relative to placebo, and (3) relative to the seven other analgesics studied. We assume that these are three distinct objectives listed in order of importance. Under these circumstances, the company would have justifiably pursued the comparison of propoxyphene with aspirin by using the usual paired t test and quoting the per-comparison error rate. The per-experiment error rate would be irrelevant in addressing this specific aim.

If the company had obtained the same data as did Moertel and associates,[4] how would the conclusions by Moertel's group be altered? Recall that Moertel and colleagues required that the P value be less than 0.01 for statistical significance (Table 2-2), rather than the more conventional 0.05 level, because of concerns about the per-experiment error rate. With the per-experiment error rate no longer of interest, the observation of the P value being less than 0.01 would retain its more usual interpretation of being highly significant. Thus, the conclusion of the inferiority of propoxyphene relative to aspirin in this alternative example would be even more convincing. In this circumstance, it would be absurd to insist that the pharmaceutical firm could not compare its drug with aspirin unless the overall F test (comparing all 10 preparations) was significant. In fact, one could imagine the situation in which the company's drug was the only one that differed from aspirin, in which case an overall F test comparing all preparations simultaneously would probably not be statistically significant.

COMMENTS

At this point, readers may ask which of the many methods of analysis that have been described is the best. Perhaps the most important point to be made is that no *one* method is always the "best" method.

The questions that a study is intended to answer must be clearly stated beforehand. Failure to consider this basic principle often may lead to an overreliance on per-experiment error rates. As an illustration, suppose one investigator conducts an experiment to answer the following question: "Are the effects of treatments A and B different?" Now suppose that a second investigator conducts the same experiment and obtains the same data but also collects additional data on a separate group of patients receiving a third treatment (C) because the investigator also wants to compare treatments A and C. It seems apparent that these investigators should arrive at the same conclusions when treatments A and B are compared, and this result will prevail when per-comparison error rates are used. In these circumstances, it would be unreasonable to insist that the second investigator quote a higher (per-experiment) error rate in answering the original question of whether the effects of treatments A and B are different.

In the examples, each patient in the study was exposed to each treatment. If such an experiment includes only two treatments, the paired t test is the appropriate statistical test to use. The additional statistical procedures described herein may be considered extensions of the paired t test to incorporate more than two treatments. In many studies, however, the patients who receive the various treatments differ, and completely separate and distinct treatment groups are formed. When two such groups exist, the appropriate statistical test is the two-sample t test.[5] Extensions of the two-sample t test for performing multiple pairwise comparisons among the groups are available and are entirely analogous to the procedures described herein. Of course, the computations will differ, depending on whether or not distinct groups are used. No new concepts are involved, however.

In both examples considered in this article, the statistical tests pertained to hypotheses that had been formulated before the study was initiated. Often, however, inspection of the data may suggest additional hypotheses. One might suppose, for example, that no difference was observed among the therapeutic modalities but that progression of disease differed between male and female patients or between old and young patients. When the hypotheses to be tested are suggested by the data in this manner, the observed differences must be reported accordingly—that is, the investigators should clearly state that the data have identified hypotheses for further study rather than confirmed previously formulated hypotheses. Although reporting any P values in this context may be inadvisable, this situation is particularly well suited to the conservative approach of reporting per-experiment error rates.

The methods described in this article also provide an opportunity to remind readers of a basic principle that should always be kept in mind when statistical methods are used: always know what assumptions are being made and be sure that they conform to your own beliefs. Specifically, many of the methods described herein (LSD procedure, Dunnett's method, SNK procedure, HSD method, and the D test) assume that the variability associated with all therapies is the same, even though the efficacy may differ. One should be aware of this assumption when choosing an appropriate method of analysis.

Finally, it is unfortunate that many persons who review manuscripts for medical journals are under the mistaken impression that one type of analysis should always be used or that one type of error rate should always be quoted and the others excluded, and this attitude is reflected in much of the medical literature. Because the issues involved are essentially nontechnical in nature, we hope that readers will be encouraged to respond with their own informed judgment as these situations are encountered.

For additional, nontechnical discussions of the use and misuse of multiple comparison procedures, readers should refer to the articles by O'Brien[6] and Little.[7]

ACKNOWLEDGMENT

We thank Charles G. Moertel, M.D., and his colleagues for the use of their data on pain relievers.

REFERENCES

1. Bancroft TA: Topics in Intermediate Statistical Methods. Vol 1. Ames, Iowa, Iowa State University Press, 1968
2. Dunnett CW: A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc 50:1096-1121, 1955
3. Dunnett CW: New tables for multiple comparisons with a control. Biometrics 20:482-491, 1964
4. Moertel CG, Ahmann DL, Taylor WF, Schwartau N: A comparative evaluation of marketed analgesic drugs. N Engl J Med 286:813-815, 1972
5. O'Brien PC, Shampo MA: Statistics for clinicians. 6. Comparing two samples (the two-sample t test). Mayo Clin Proc 56:393-394, 1981
6. O'Brien PC: The appropriateness of analysis of variance and multiple-comparison procedures. Biometrics 39:787-788, 1983
7. Little TM: If Galileo published in HortScience (editorial). HortScience 13:504-506, 1978

APPENDIX

As mentioned in the text, the LSD procedure may be considered a modified version of the t test, and the necessary assumption is that the variability is the same for all therapies. For two therapeutic modalities, the paired t test statistic is determined as follows:

$$\begin{aligned} t &= \bar{\Delta}/\mathrm{SE}_{\bar{\Delta}} \\ &= \sqrt{n}\left(\frac{\bar{\Delta}}{s_{\Delta}}\right) \end{aligned}$$

in which $\bar{\Delta}$ is the mean of the differences between the two treatments in the series of patients, $\mathrm{SE}_{\bar{\Delta}}$ is the standard error of $\bar{\Delta}$ ($s_{\Delta}/\sqrt{n}$), s_{Δ} is the standard deviation of the differences, and n is the sample size (the number of pairs of values).

In the overall F test, we formulate the null hypothesis that all therapies have the same effect, so that in this case the variability of paired differences will be the same regardless of which two treatments are being compared. If we use s_P to denote the pooled estimate of variability obtained in the overall F test, the test statistic for the modified procedure is determined by the following:

$$t' = \sqrt{n}\left(\frac{\bar{\Delta}}{s_P}\right)$$

Another modification with use of the LSD method is that the tables used to obtain P values will be entered in a way that reflects the use of all the data to estimate variability. Thus, even if s_P were exactly equal to s_Δ, the P value associated with the modified (LSD) t test would be smaller than the P value obtained by using the usual paired t test.

The other procedures that necessitate the assumption that the variability is the same for all treatments are the Dunnett, SNK, HSD, and D procedures.

Statistics Series

Statistical Considerations for Performing Multiple Tests in a Single Experiment. 3. Repeated Measures Over Time

PETER C. O'BRIEN, Ph.D., *Section of Biostatistics*; MARC A. SHAMPO, Ph.D., *Section of Publications*

In part 2, we considered the situation in which patients are successively exposed to different treatment regimens, with the intention of comparing the various treatments. A similar situation arises when patients are observed at baseline, exposed to an experimental therapy, and then observed at multiple predetermined times after therapy. With use of the analogy that equates observation times with experimental therapies, one might suppose that the same types of analyses discussed in our previous article could be applied to this type of "repeated measures study"; indeed, this mistake often occurs in practice. In this situation, however, the methods discussed previously are inappropriate because the questions that these two types of studies are intended to answer are different. The following questions are usually of interest in the type of study that we discuss in this article: Is a treatment effect present? Are the changes that occur after treatment constant over time, or are they either accelerating or decelerating over time? When does the treatment effect begin, reach a peak, and end? In this article, we describe some statistical techniques used to answer these questions.

IS A TREATMENT EFFECT PRESENT?

It will help to answer this question if the investigator can specify in advance the nature of the anticipated treatment effect. If the investigator is confident that any effect should become apparent immediately but may be of short duration, a paired *t* test would be an appropriate statistical method to compare the baseline measurement with the first posttreatment measurement. Alternatively, if the investigator can specify that treatment should result in fairly constant posttreatment measurements over time, using a paired *t* test to compare the mean of the posttreatment values with the baseline value would be appropriate.

Another possibility is that measurements may be expected to increase or decrease steadily over time. This possibility can be evaluated by computing the change from baseline at each time point for each patient. For each patient, the observed changes can then be related to time by regression analysis.[1] Specifically, for each patient, an equation is obtained relating the changes from baseline (Y) to time (T): $\Delta = a + bT$ (in which a and b are fitted values for the intercept and slope, respectively—values that will vary from patient to patient). Using a *t* test in which *t* is the mean of the slopes (b) divided by the standard error of the mean,[2] one can then test the hypothesis that the treatment effect is steadily increasing. The hypothesis of no association with time will be rejected if the corresponding *P* value is small (for example, $P<0.05$).

Unfortunately, because the properties of the experimental therapy often are not sufficiently well understood, predicting whether or when a response will occur is often not possible. In this case, performing various analyses to investigate the possibilities we have described may be appropriate. When this strategy is adopted, however, one should be aware that the possibility of erroneously observing a statistically significant

Individual reprints of this article are not available. The entire six-part series will be available for purchase as a bound booklet from the *Proceedings* Circulation Office in December.

treatment effect increases when multiple statistical procedures are used to test the same null hypothesis (no treatment effect). This type of problem will be the subject of a future article. For now, we point out that if the conclusions suggested by different analyses are divergent, the results should be interpreted with caution.

IS THE RATE OF CHANGE CONSTANT OVER TIME?

We will suppose that one has performed the aforementioned regression analysis and concluded that the effect of treatment is dependent on time. A logical question to ask under these circumstances is, "Is the rate of change constant over time, or is the rate of change either accelerating or decelerating?" This question can be formally tested by fitting a more complicated regression model: $Y = a + bT + cT^2$. This model is fitted for each patient (the computations are easily performed by using standard computer software). One then computes $t = \bar{c}/SE_c$, in which $\bar{c}$ is the mean of the fitted c values and SE_c is the standard error of the mean. If $\bar{c}$ differs significantly from 0, one would conclude that the rate of change is not constant over time. For a more thorough evaluation of the precise nature of the association with time, the descriptive methods to be discussed next are useful.

WHEN DOES THE TREATMENT EFFECT BEGIN, REACH A PEAK, AND END?

When these types of questions are addressed, analysis necessarily focuses on description rather than on formal testing of hypotheses. This descriptive method is simple to perform and is extremely valuable. The simplicity of this type of analysis should not be construed to imply that it is less informative than the more sophisticated types of analyses described previously. On the contrary, it sometimes may be more informative.

One of the most useful methods of evaluating changes over time is to graph the mean and the standard error of the mean at each time point in a study, the measurement of interest being graphed on the vertical axis and time being graphed on the horizontal axis (Fig. 3-1). The reason for graphing the standard error of the mean rather than the standard deviation is that the former indicates the precision with which the mean is estimated (usually of greater interest in this context), whereas the latter indicates the amount of variability among individual data points. In such an evaluation, however, either could be used. (Alternatively, an indication of the overlap at various time points also could be provided by graphing the minimal and maximal values rather than the standard error.)

Performing multiple paired *t* tests in which each time point is compared with the baseline (or compared with the immediately preceding time point) is often helpful for interpreting such a graph. The resulting *P* values, however, must be interpreted with caution. For example, if statistical significance is observed at only one time point, one must consider the possibility of a chance occurrence. One should insist that such a *P* value be convincing ($P<0.01$, for example) before concluding that an actual treatment effect has been observed. Conversely, if a systematic trend is noted (for example, *P* values become consistently smaller over time and then increase), conclusions can be drawn with greater confidence.

EXAMPLE

A study by Essandoh and associates[3] evaluating the effects of head-down neck flexion on limb blood flow can be used as an example. Blood flow in the calf and forearm was measured in eight

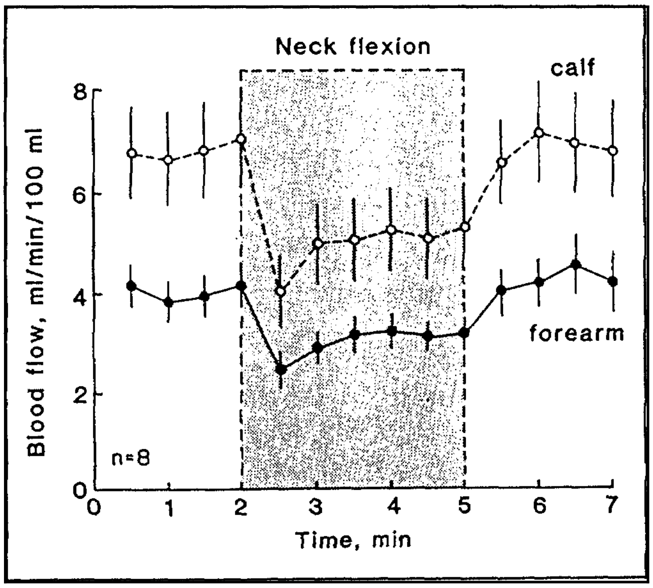

Fig. 3-1. Effects of head-down neck flexion on blood flow in the calf and forearm of eight healthy male subjects. (From Essandoh and associates.[3] By permission of the American Physiological Society.)

healthy male subjects while the subjects were in the prone position for 7 minutes. After the first 2 minutes, the head of the subject was maximally flexed and lowered. Three minutes later, the head was returned to the initial position. Paired *t* tests that compared each time point with the mean of the four baseline values showed statistically significant reductions in blood flow throughout the time that the neck was lowered (Fig. 3-1). With resumption of the initial position, significant differences from baseline were not observed during the final 2 minutes. In that study, the primary interest was in the occurrence of a change immediately after lowering of the neck; thus, the *t* test comparing the value at 2½ minutes with the baseline value was appropriate without consideration of a per-experiment error rate. The use of *t* tests at subsequent time points was helpful for interpreting the consistent pattern observed in Figure 3-1. We believe that, in this instance, the use of a graphic display aided by *t* tests appropriately conveyed the pertinent information obtained by the investigators and that the use of more sophisticated statistical analysis would not have been helpful.

ACKNOWLEDGMENT

We thank Louis K. Essandoh, M.D., and his colleagues for the use of their data regarding the effects of neck flexion on limb blood flow.

REFERENCES

1. O'Brien PC, Shampo MA: Statistics for clinicians. 7. Regression. Mayo Clin Proc 56:452-454, 1981
2. O'Brien PC, Shampo MA: Statistics for clinicians. 4. Estimation from samples. Mayo Clin Proc 56:274-276, 1981
3. Essandoh LK, Duprez DA, Shepherd JT: *Reflex constriction of human limb resistance vessels to head-down neck flexion.* J Appl Physiol 64:767-770, 1988

Statistics Series

Statistical Considerations for Performing Multiple Tests in a Single Experiment. 4. Performing Multiple Statistical Tests on the Same Data

PETER C. O'BRIEN, Ph.D., *Section of Biostatistics*; MARC A. SHAMPO, Ph.D., *Section of Publications*

In this article, we compare two groups of patients with respect to a single measurement. For example, in a clinical trial to evaluate the ability of a drug to decrease blood pressure, one might compare the reductions noted in a group of patients receiving the drug with the reductions recorded in a comparable group of patients receiving a placebo. A standard statistical test for this type of comparison is the two-sample *t* test.[1]

The *t* test, however, is only one of many statistical tests that could be used to test the hypothesis of no difference between therapies. How should one decide which test to use? In this article, we will provide guidelines for choosing an appropriate test and for interpreting the results when more than one test has been used. In general, one must consider two goals: (1) ensuring that the final probability (*P* value) statement is valid and (2) maximizing the possibility of detecting a treatment effect when experimental therapy actually is efficacious.

VALIDITY OF THE PROBABILITY STATEMENT

The *t* test assumes that the data are not highly skewed and that no outliers are present. If these assumptions are satisfied, the *t* test will be valid—that is, the *P* value statement associated with the *t* test will be accurate. If departures from these assumptions are large, however, the *P* value statement may be inaccurate. (Usually, the skewness or outliers that one needs to be concerned about are obvious on inspection of a graphic display of the data. A good rule of thumb is that the *t* test should not be used if the mean value differs appreciably from the median.)

When the *t* test cannot be used, a common alternative is the Wilcoxon rank sum test.[2] It consists of pooling the data from both samples and ranking the values from smallest to largest. One computes the sum of the ranks in each sample and obtains the corresponding *P* value from special tables.

Which test should one use? If the assumptions required by the *t* test are satisfied (at least approximately), the *t* test should be used because it is generally more likely to detect a true difference in therapeutic modalities under these circumstances. When large departures from these assumptions occur, the test will not yield a valid probability statement; then the rank sum test should be used instead. An important aspect of this decision is that it should be made before performing either test. For example, if one were to perform both tests and choose the one that yielded the smaller *P* value, the selected *P* value would not provide a valid probability statement.

DETECTING A TRUE DIFFERENCE IN THERAPIES

If multiple statistical tests (*t* or rank sum, for example) are available and each provides valid probability statements, which one should be used? The general guide is to use the test that is most likely to detect an actual treatment effect. Thus, in comparing the *t* and rank sum tests, we advocate use of the *t* test when the necessary assumptions are satisfied because under these conditions it is generally a more sensitive test.

Individual reprints of this article are not available. The entire six-part series will be available for purchase as a bound booklet from the *Proceedings* Circulation Office in December.

Typically, however, the ability of any test to detect a treatment effect will depend on the nature of the effect. For example, both the *t* test and the rank sum test assume that the effect of the experimental therapy is the same in each patient. In our example, this assumption would mean that the biologic effect of the drug is to reduce blood pressure by the same amount in each patient (homogeneous effect). Both the *t* and rank sum tests will perform well in detecting this type of a treatment effect if the assumption is at least approximately true. Conversely, these procedures may perform poorly if the effect of therapy varies considerably among patients (heterogeneous effect). Because it is typically difficult, if not impossible, to specify the precise nature of a treatment effect in advance, the need may arise to perform multiple statistical tests to investigate various possible types of treatment effects. A generalization of the standard *t* test to explore the possibility of a heterogeneous treatment effect is described next.[3]

Step 1.—Let $W_1 < W_2 < \ldots < W_{n_E + n_P}$ represent the values in the two samples arranged from smallest to largest, in which n_E and n_P represent the number of patients in the experimental and placebo groups, respectively. For each value of W_i ($i = 1, \ldots, n_E + n_P$), the variable Z_i is defined as follows:

$Z_i = 1$, if the patient corresponding to W_i is in the experimentally treated group and
$Z_i = 0$, if the patient is in the placebo group

Step 2.—Using standard statistical computing software, fit the following model:

$$Z_i = a + bW_i + cW_i^2$$

in which a, b, and c are coefficients to be estimated.

Step 3.—If c differs significantly from 0, the suggestion is that the effect of treatment is heterogeneous among patients. In this case, an overall analysis for group differences is to test the hypothesis that the true values for both b and c are 0. If a significant treatment effect is observed, which seems to be heterogeneous, one should be careful *in measuring the magnitude of* the effect. For example, overall group summary statistics such as group means may be misleading. A graphic display of all the data will be more informative. If possible, subgroup analyses identifying patients who are most responsive to therapy are indicated.

Step 4.—If c does not differ significantly from 0, the implication is that the treatment effect (if any) is approximately homogeneous among patients. Under these circumstances, the *t* test will perform well. In terms of the methods described in the foregoing steps 1 through 3, one could fit the model

$$Z_i = a' + b'W_i$$

and test that b′ differs significantly from 0. In fact, this test is algebraically identical to the standard *t* test.

Corresponding methods generalizing the rank sum test are available. Specifically, one uses the rank of W_i in place of W_i in the preceding computations. With either method, one needs to be aware that multiple statistical tests are being used to evaluate the possibility of a treatment effect, and this approach increases the possibility of erroneously concluding that an effect exists when in truth it does not. Consequently, these types of explorations that go beyond conventional *t* and rank sum tests should be reported with caution. The results of such tests will be more convincing if they are supported by biologic considerations.

EXAMPLE

We illustrate the considerations discussed thus far with an example from a study performed at the Mayo Clinic by Dyck and colleagues.[4] This randomized, double-blind study of patients with chronic inflammatory demyelinating polyradiculoneuropathy compared a group of 15 patients receiving plasmapheresis with a group of 14 patients having sham-pheresis. A primary endpoint was the change from baseline in the neurologic disability score observed at 3 weeks (Table 4-1).

Because one value (-69.0) was an obvious outlier and because the data were also highly skewed (notice that, ignoring the outlier, the larger values were generally more spread out than the smaller values), a *t* test was inappropriate. Therefore, a rank sum test was used. The sums of the ranks for the plasmapheresis and sham-pheresis groups were 241 and 194, respectively (data from Table 4-1). A negative result was obtained by using a one-sided rank sum test ($P = 0.249$). One would expect, however, that plasmapheresis may benefit some patients more than others. A test of this hypothesis (see the aforementioned steps able 4-1.—**Changes in Neurologic Disability Score (NDS) in Patients With Chronic Inflammatory Demyelinating Polyradiculoneuropathy Who eceived Either Plasmapheresis or Sham-Pheresis**

Change in NDS*	Ranked value	Group†
-69.0	1	0
-7.0	2	1
-6.8	3	1
-6.1	4	1
-6.0	5	0
-5.2	6	0
-4.0	7	1
-3.0	8	1
-2.5	9	0
-2.0	10	1
-0.5	11	1
+1.5	12	0
+4.0	13	0
+5.0	14	0
+5.5	15	0
+8.0	16	0
+13.0	17.5	0
+13.0	17.5	1
+15.0	19	0
+16.0	20.5	0
+16.0	20.5	1
+21.5	22	0
+23.5	23	1
+26.0	24	0
+32.4	25	1
+33.5	26	1
+47.3	27	1
+83.0	28	1
+96.0	29	1

*Positive (+) values indicate improvement.
†1 = patient received plasmapheresis;
0 = patient received sham-pheresis.

hrough 3) indicated that the coefficient of the
adratic term for the generalized rank sum
del differed significantly from 0 (P = 0.020).
e overall test for a difference between groups
s also significant (P = 0.027). These results con-
m to prior expectations that plasmapheresis
y benefit only a subgroup of patients, and this
tcome is reflected in the data, wherein the five
tients who improved the most all received plas-
pheresis, although some patients receiving
smapheresis failed to show any improvement.
ecause of the small P values associated with
generalized rank sum test and the strong a
ori justification for expecting a heterogeneous
atment effect, the authors appropriately con-
ded that plasmapheresis may be beneficial in
e treatment of chronic inflammatory demyelin-
ng polyradiculoneuropathy, at least for some
tients. Unfortunately, the sample size was too
all to pursue meaningful subgroup analyses.

CONCLUSION

Even in the relatively simple situation of comparing two therapies with respect to a single endpoint, one often is led to consider multiple statistical tests. When the goal is to achieve a valid probability statement, if the necessary assumptions are satisfied, the interpretation of resulting P values is unaffected. For example, one may intend initially to base the comparison on a t test but, on completion of the study, find that the occurrence of an outlier makes this impractical. Under these circumstances, use of a rank sum test will not alter the interpretation of the resulting P value.

The problem is considerably more complex when an investigator performs multiple tests to study different types of treatment effects that may occur. In general, we recommend the use of a single primary test, supplemented by additional secondary tests as may be needed. This approach was illustrated in the example of chronic inflammatory demyelinating polyradiculoneuropathy, in which the primary test was the rank sum test, and the corresponding P value was duly reported. In this case, the conclusions were appropriately altered by the results of secondary tests supported by biologic considerations. Under these circumstances, reporting of the results of secondary tests should also be accompanied by suitable cautions regarding the increased possibility for error when multiple statistical tests are used to evaluate efficacy.

ACKNOWLEDGMENT

We thank Peter J. Dyck, M.D., and his colleagues for the use of their data on the treatment of chronic inflammatory demyelinating polyradiculoneuropathy.

REFERENCES

1. O'Brien PC, Shampo MA: Statistics for clinicians. 6. Comparing two samples (the two-sample t test). Mayo Clin Proc 56:393-394, 1981
2. Snedecor GW, Cochran WG: Statistical Methods. Sixth edition. Ames, Iowa, Iowa State University Press, 1967, pp 130-134
3. O'Brien PC: Comparing two samples: extensions of the t, rank-sum, and log-rank tests. J Am Stat Assoc 83:52-61, 1988
4. Dyck PJ, Daube J, O'Brien P, Pineda A, Low PA, Windebank AJ, Swanson C: Plasma exchange in chronic inflammatory demyelinating polyradiculoneuropathy. N Engl J Med 314:461-465, 1986

Statistical Considerations for Performing Multiple Tests in a Single Experiment.´ 5. Comparing Two Therapies With Respect to Several Endpoints

PETER C. O'BRIEN, Ph.D., *Section of Biostatistics*; MARC A. SHAMPO, Ph.D., *Section of Publications*

In this article, we suppose that an investigator has conducted a comparative study between two groups of patients who have a certain disease, for the purpose of determining whether one therapy is more efficacious than another. Although our discussion of this topic applies to general settings, we will suppose that the study was a randomized clinical trial in which an experimental therapy was compared with a conventional therapy and that efficacy was determined by one or more quantitative measurements (for example, change in blood pressure observed during the course of the study).

In practice, the data analysis for such a study would address several different, but related, questions. A primary question, of course, would be, Is a treatment effect present? Subsequent questions would be, What specific patient characteristics are affected? and, Is the effect of sufficient magnitude that the benefits offset the risks? Other questions would be related to characterizing the effect: When does it begin? How long does it last? In this article, we address only the first question, Is a treatment effect present? As in the other articles in this series, many statistical issues that would need to be considered in any actual study will be omitted from the discussion. Instead, we focus on the use of statistical tests of hypotheses and the problems that arise when multiple tests are used.

If efficacy is measured by only one patient characteristic, testing for a difference between the two therapies is easily done by using standard statistical procedures, such as the two-sample *t* test.[1] On the basis of the calculated value of *t*, a *P* value is obtained from suitable tables or computing equipment. The *P* value is the probability of observing a value for *t* as large as in our study if no actual difference existed between therapeutic modalities. If the *P* value is sufficiently small (for example, $P<0.05$), one can conclude that the observed *difference in efficacy* is statistically significant and that an actual difference between therapies probably exists.

Suppose, however, that the investigator has measured many different patient characteristics and that each characteristic provides an indication about whether the therapy has been efficacious. For example, the efficacy of therapy for heart disease may be determined by data obtained from echocardiograms, electrocardiograms, angiograms, and various clinical measurements. How can an investigator perform a statistical test in these circumstances to compare the two therapies? We describe some approaches in the next section.

STATISTICAL PROCEDURES FOR COMPARING TWO THERAPIES WITH RESPECT TO MULTIPLE ENDPOINTS

Identify a Single Endpoint for Analysis.— A commonly used approach to the problem, and perhaps the simplest, is to identify one primary endpoint (measure of efficacy), with the understanding that the decision to accept or re-

Individual reprints of this article are not available. The entire six-part series will be available for purchase as a bound booklet from the *Proceedings* Circulation Office in December.

ect the new therapy will depend on the results of a statistical test applied to this single endpoint. This approach has the advantage of providing a single, unambiguous, valid probability statement for comparing the therapies. Furthermore, if only one of the endpoints is expected to reflect the benefit associated with the experimental therapy reliably, this method may be a sensitive technique for identifying a difference. The obvious disadvantage, of course, is that this method fails to use the information from the other measurements.

Perform Multiple t *Tests.*—A second approach is to compare the therapies with respect to each endpoint by using a t test (or any other two-sample test procedure); as described previously. The P value associated with each test (that is, the per-comparison error rate, as described in the introduction to this series) provides an indication about whether therapy is efficacious, and when these P values are considered collectively, an overall judgment about efficacy can be made. This approach has the advantage of using all the available information; however, it fails to provide a single overall P value as an objective criterion for accepting or rejecting the experimental therapy.

Use the Bonferroni Adjustment.—Suppose that on inspecting the results of separate t tests, as previously described, the investigator noted that the smallest observed P value was some number, which we will denote by P^*. For example, suppose $P^* = 0.017$. In these circumstances, a logical question would be, "If no difference existed between therapies, what is the probability that the smallest observed P value would be less than or equal to P^*?" A common approximation is obtained by using a Bonferroni adjustment, as described in part 2 of this series, whereby P^* is multiplied by the number of statistical tests that were performed. In our example, if $P^* = 0.017$, and this was the smallest P value observed among 10 tests, the Bonferroni-adjusted P value (P_A) would be 0.17 (that is, $0.017 \times 10 = 0.17$). This approach provides another possible criterion for judging whether to accept the experimental therapy. That is, one may decide to accept the new therapy as superior only if the *Bonferroni-adjusted P value* is less than 0.05. This strategy has the desirable property that, if no actual difference exists between therapies, the probability that the experimental therapy would be erroneously accepted is less than 0.05. (Notice that this Bonferroni-adjusted error rate is analogous to the concept of a per-experiment error rate introduced in part 1.)

A disadvantage of this procedure is that once the endpoint associated with the smallest P value is identified, the information available from the other endpoints is discarded. For example, if 10 measures of efficacy had been evaluated and all 10 separate t tests had produced $P = 0.02$, the Bonferroni-adjusted P value would be $P_A = 0.20$ (that is, $10 \times 0.02 = 0.20$), an indication of no difference among therapies. This conclusion is counter to intuition, because statistical significance was achieved for each measure of efficacy when each was considered individually.

A second disadvantage of the Bonferroni approach is that it provides only an approximate overall error rate, and this approximation may be poor, especially if the endpoints are highly correlated. Consider an extreme example. Suppose that in the previous example all 10 endpoints were perfectly correlated—that is, any 1 measurement on a patient could predict the 9 other measurements exactly, without error. In this case, the results of all 10 tests would be identical, and any 1 test could be selected as a criterion for accepting or rejecting the experimental therapy. For example, if a P value of 0.04 was observed, this would appropriately be judged statistically significant. The Bonferroni-adjusted P value (P_A), however, would be 0.40 (that is, $0.04 \times 10 = 0.40$), an indication of no difference. The method we describe next overcomes both of these difficulties.

Perform a Global Test.—Although all the aforementioned methods are commonly used in practice, each has one or more deficiencies. Intuitively, one wants an overall test procedure that will accumulate the separate pieces of information available from the several endpoints. In the following paragraphs, we describe a method for performing such a test.[2]

The first step is to consider each endpoint separately. Rank all the measured values (pooling the data from both groups) from worst to best in terms of efficacy. Next, replace the measured value with the assigned rank. For example, the lowest ejection fraction in a cardiology study would be replaced by a value of 1, and the highest ejection fraction would be replaced by a value of $n_E + n_C$ (n_E is the number of patients in the experimental therapy group and n_C is the number of patients in the conventional therapy group).

The second step is to consider each patient separately and add the rank values received. For example, in our hypothetical cardiology study, a patient's rank value for ejection fraction would be added to the rank values for all other measurements. The summated values provide a single number for each patient in the study that indicates his or her response to therapy relative to the other patients. Thus, the final step is to compare the summated values between the two groups by using any standard statistical procedure for comparing two samples (for example, a two-sample *t* test).

Table 5-1.—**Distribution of 34 *P* Values Comparing Two Groups of Patients With Diabetes, Six Treated With Conventional Insulin Therapy and Five Treated With Continuous Subcutaneous Insulin Infusion**

P value interval	Percentage of 34 variables*
0-0.1	38
0.1-0.2	12
0.2-0.3	9
0.3-0.4	15
0.4-0.5	6
0.5-0.6	2
0.6-0.7	9
0.7-0.8	3
0.8-0.9	0
0.9-1.0	3

*Percentages do not total 100% because they were rounded off.

EXAMPLE

We illustrate the foregoing procedures with an example from a study conducted at the Mayo Clinic by Service and colleagues,[3] in which two therapies for diabetes mellitus were compared. Twelve patients with insulin-dependent diabetes mellitus who were deficient in C peptide were randomly assigned to either conventional insulin therapy or continuous subcutaneous insulin infusion. Eleven patients (six receiving conventional therapy and five receiving infusion therapy) completed the study, which focused on the effects of therapy on peripheral nerve function. Nerve conduction was studied by measuring 34 electromyographic variables.

Because all 34 variables were of interest, identifying a single primary endpoint was not feasible. Separate comparisons between the two groups were made for each of the 34 measurements, and the results are summarized in Table 5-1. Because outliers and skewness were observed for some of the variables, *t* tests were not used for these comparisons; rank sum tests were used. With this procedure, the smallest possible *P* value was 0.002, which was observed for one of the tests. The second smallest *P* value was 0.015. A pattern of small *P* values was observed, suggesting that the experimental therapy may be more efficacious than conventional therapy for improving nerve conduction. Consideration of the 34 *P* values separately, however, failed to provide a single objective criterion for judging efficacy.

In this situation, using a Bonferroni-adjusted *P* value was not feasible. Because the smallest possible *P* value with the rank sum test (with six patients in each group) was 0.002, the smallest possible adjusted *P* value with 34 variables would be 0.002 × 34 = 0.068. With use of the global test based on summated ranks, the *P* value was 0.033, a finding that supports the original impression of a treatment benefit. For further understanding of the nature of the effect, subgroupings of the data were considered. The results indicated that the effects were most apparent proximally (Table 5-2).

COMMENT

In this article, we focused attention on evaluating the efficacy of an experimental therapy, when efficacy was determined by multiple patient characteristics. Both per-comparison error rates (from considering each characteristic individually) and the per-experiment error rate (the single over-

Table 5-2.—**Overall *P* Values Comparing Two Groups of Patients With Diabetes, Six Treated With Conventional Insulin Therapy and Five Treated With Continuous Subcutaneous Insulin Infusion, According to Various Groupings of Electromyographic Measurements**

Variables included	No. of variables	*P* value
All	34	0.033
Speed of conduction	21	0.044
Sensory conduction	13	0.045
Lower extremity function	13	0.156
Distal function	8	0.164
Proximal function	6	0.001
Elevated from normal initially	14	0.156
Most reliably measured	8	0.029

Modified from Service and associates.[3] By permission of Springer-Verlag.

all P value) were informative for answering the question of whether experimental therapy was superior to conventional therapy.

We have not considered a related important but qualitatively different question: Which individual electromyographic characteristics demonstrate a benefit from the experimental therapy? This question must necessarily be addressed by consideration of each characteristic individually. In the example, a per-comparison error rate of $P = 0.002$ was observed for median nerve somatosensory-evoked potential latency at the neck. When so many statistical tests are considered, however, Bonferroni adjustment suggests that such a small P value may occur by chance. Because of the very small per-comparison error rate, this endpoint may be justifiably identified as a potential candidate for further study. One might want to temper any definitive conclusions, however, because the per-experiment error rate based on *the Bonferroni adjustment is more than 0.05. In* general, conclusions may depend on the presence or absence of similar trends in related measurements and on other circumstances surrounding the study.

ACKNOWLEDGMENT

We thank F. John Service, M.D., and his colleagues for the use of their data on the treatment of diabetes mellitus.

REFERENCES

1. O'Brien PC, Shampo MA: Statistics for clinicians. 6. Comparing two samples (the two-sample *t* test). Mayo Clin Proc 56:393-394, 1981
2. O'Brien PC: Procedures for comparing samples with multiple endpoints. Biometrics 40:1079-1087, 1984
3. Service FJ, Rizza RA, Daube JR, O'Brien PC, Dyck PJ: Near normoglycaemia improved nerve conduction and vibration sensation in diabetic neuropathy. Diabetologia 28:722-727, 1985

Statistics Series

Statistical Considerations for Performing Multiple Tests in a Single Experiment. 6. Testing Accumulating Data Repeatedly Over Time

PETER C. O'BRIEN, Ph.D., *Section of Biostatistics*; MARC A. SHAMPO, Ph.D., *Section of Publications*

Suppose that an investigator wants to perform a clinical trial to compare a group of patients who receive an experimental drug with a control group that receives only a placebo. For simplicity, assume that the result of therapy is dichotomous (for example, success or failure) and that the result is known soon after therapy has been administered. How can the investigator perform a statistical test to determine whether the experimental therapy is superior to the placebo? The simplest approach is to wait until all the patients have been entered into the trial and then perform a χ^2 test on the results.[1]

In practice, however, this approach of waiting to analyze the data until all patients have completed the trial is often impractical. For example, ethical considerations may necessitate periodic monitoring of the accumulating data, with the understanding that if one therapy is found to be much superior to the other, the trial will be discontinued. Under these circumstances, how can one perform statistical tests at each of the monitoring time points and still obtain a single, overall probability statement (*P* value) for judging the efficacy? Statistical methods for answering this question will be addressed in this article.

We note at the outset that it would be inappropriate to perform the usual χ^2 tests and obtain the corresponding *P* values at each analysis. The problem with this approach is that the probability of incorrectly stopping the trial would be too large. For example, suppose an investigator had decided to perform one interim test with the understanding that the experimental therapy would be judged superior if (1) the interim test yielded $P<0.05$ (in which case, the trial would be terminated) or (2) the interim test yielded $P>0.05$ but the final test at the conclusion of the study yielded $P<0.05$. With this strategy, the probability of incorrectly deciding that the experimental therapy is efficacious equals (1) the probability of obtaining $P<0.05$ in the interim analysis (this probability is 0.050) *plus* (2) the probability of obtaining $P>0.05$ in the interim analysis and $P<0.05$ in the final analysis (this probability has been evaluated[2] and equals 0.030). Thus, with this strategy, the overall probability of incorrectly concluding that the experimental therapy is efficacious is $0.050 + 0.030 = 0.080$. Similarly, if one were evaluating a treatment that was actually ineffective and performed five tests at the 0.05 level, one would incorrectly conclude that the treatment was efficacious with a probability of 0.14.

METHODS FOR REPEATED SIGNIFICANCE TESTING WITH ACCUMULATING DATA

If multiple testing is to be taken into account, each test must be performed at a more stringent level of significance (at a lower value of *P*). We will describe three of the most commonly used methods for doing this. With each method, one computes the usual test statistic by using all available data in each analysis. The critical value for judging significance, however, is altered to account for multiple testing.

Individual reprints of this article are not available. The entire six-part series will be available for purchase as a bound booklet from the *Proceedings* Circulation Office in December.

Pocock Boundary.—Instead of using $P = 0.05$ as the criterion of significance for each test, the Pocock method[3] uses a smaller value—P'. The actual value for P' will depend on the number of analyses to be conducted. For example, with one interim test, both the interim test and the final test would be conducted at the $P' = 0.029$ level to provide an overall significance level of 0.05. With five tests, each would need to be conducted at the $P' = 0.016$ level to achieve an overall significance level of 0.05.

A disadvantage of this method arises in trials in which early termination does not occur. For example, if four interim tests are planned and early termination does not occur, the final test at the end of the study must achieve a value of $P' = 0.016$ in order to conclude significance at $P = 0.05$. If the observed value was 0.02, for example, the investigator would be unable to claim statistical significance. This situation would be awkward: if the investigator had not contemplated early termination and had performed only one test (at the end of the study), no adjustment for multiple testing would have been necessary, and the results of the study would have been significant at $P = 0.02$. The next two procedures address this problem. Because these procedures require more stringent criteria for the initial interim tests, the adjustment needed in the final test (assuming the study is not terminated early) becomes negligible.

O'Brien-Fleming Boundary.—With use of the O'Brien-Fleming boundary,[4] the adjustment for interim testing is large early in the study but constantly decreases and is negligible by the end of the study. For example, with two interim tests and an overall adjusted P value of 0.05, the necessary levels of significance are 0.0006, 0.015, and 0.047 at the first, second, and third analyses, respectively. If more than two interim tests are planned, the possibility of stopping at the first analysis becomes extremely remote.-To some extent, this situation may be desirable. (As we will discuss, there are important reasons for not terminating a study early.) If a less stringent criterion is desired for the initial test, however, $P' = 0.001$ could be substituted, with negligible effect on the overall adjusted P value.

Haybittle/Peto Boundary.—The Haybittle/Peto method[5,6] uses a constant boundary (P') for the interim tests but makes no adjustment in the analysis at the end of the study if early termination does not occur. Although the actual error rate will exceed 0.05, the difference will be negligible if the interim level for testing is sufficiently stringent ($P' = 0.001$, for example).

Comment.—The stopping boundaries of the aforementioned three methods are shown in Table 6-1 and Figure 6-1. In an actual study, the choice among these methods will depend on the circumstances surrounding the study and the desirability of stopping the study early versus the desirability of making a smaller adjustment in the final analysis if early termination does not occur. (Of course, the choice of a stopping rule must be made before the data are collected.)

SPECIFIC EXAMPLES

Example 1.—We consider a clinical trial conducted by G. S. Gilchrist, M.D., at the Mayo Clinic (personal communication) in which prednisone was compared with prednisone plus vincristine for the treatment of leukemia. Success was defined as remission, which occurs relatively soon after treatment or not at all. As indicated in Table 6-2, remissions (responses) occurred in 38 of 42 patients who received prednisone plus vincristine and in 14 of 21 patients who received prednisone only. This study was not designed to include interim testing, and a conventional χ^2 test at the conclusion of the study yielded $P = 0.0095$.

Table 6-1.—**Group Sequential Stopping Boundaries***

Test no.	*P* value, by boundary		
(k)	Pocock	O'Brien-Fleming	Haybittle/Peto
		One interim test	
1	0.029	0.005	0.010
2	0.029	0.049	0.050
		Two interim tests	
1	0.022	0.0006	0.010
2	0.022	0.0151	0.010
3	0.022	0.0471	0.010
		Three interim tests	
1	0.018	5×10^{-5}	0.001
2	0.018	0.004	0.001
3	0.018	0.018	0.001
4	0.018	0.042	0.050
		Four interim tests	
1	0.016	5×10^{-6}	0.001
2	0.016	0.001	0.001
3	0.016	0.009	0.001
4	0.016	0.023	0.001
5	0.016	0.042	0.050

*Per-experiment error rate = 0.05.

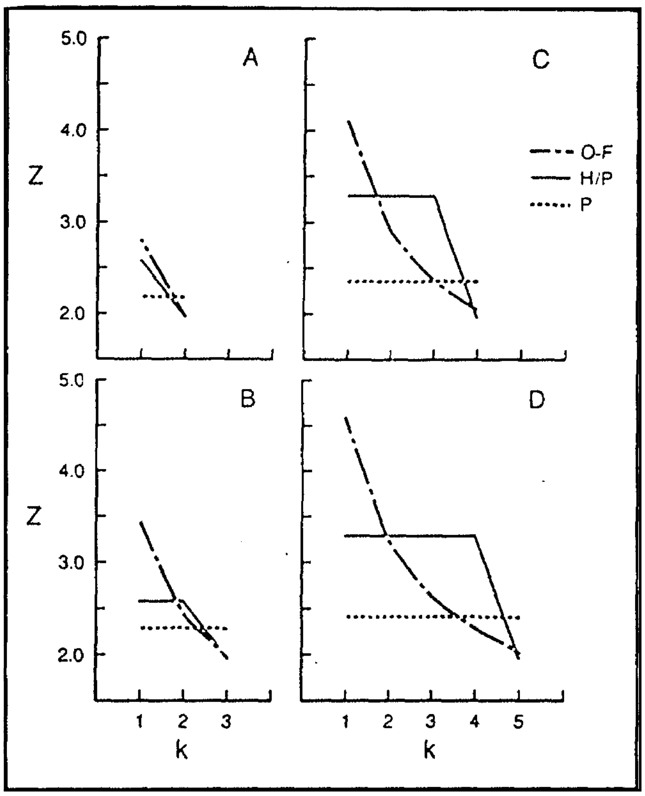

Fig. 6-1. Group sequential boundaries, indicating value of test statistic (*Z*) required at each test (*k*) to achieve overall statistical significance at $P = 0.05$ level. Boundaries are shown when maximum of two (*A*), three (*B*), four (*C*), or five (*D*) tests is planned. *O-F* = O'Brien-Fleming boundary; *H/P* = Haybittle/Peto boundary; *P* = Pocock boundary.

We now consider what would have occurred had the possibility for interim testing been incorporated into the study design. We will assume that the investigator had planned for two interim analyses after one-third and two-thirds of the patients had entered the trial and that the Pocock boundary was used with an overall significance level of 0.05. An unadjusted *P* value of 0.022 or less would have been needed at any of the analyses to achieve statistical significance. At the initial analysis, the unadjusted *P* value was 0.2159. Because this value exceeds 0.022, the investigator would have continued to enter patients into the trial. At the next interim analysis, the unadjusted *P* value was 0.0259. Because this value is again greater than 0.022, the investigator would have completed the trial and obtained an unadjusted *P* value of 0.0095 at the final analysis. Because this value is less than 0.022, the results would have been judged significant at $P<0.05$, adjusted for multiple testing. If the O'Brien-Fleming boundary or the Haybittle/Peto boundary had been used, essentially the same results would have been obtained; however, the adjusted *P* value would have been smaller (approximately 0.01).

Example 2.—Thus far, we have assumed that the response to therapy was known immediately after treatment. In practice, one usually conducts follow-up of patients over time, and the endpoint is survival time after treatment. We will illustrate how the methods described thus far may be used in these situations. We consider a clinical trial done at the Mayo Clinic by Lininger and associates.[7] The goal of the study was to compare two regimens of chemotherapy for extensive small cell lung cancer. Regimen A consisted of cyclophosphamide, vincristine, VP-16, and cisplatin alternated with doxorubicin hydrochloride (Adriamycin) and dacarbazine. Regimen B consisted of doxorubicin hydrochloride, vincristine, VP-16, and cisplatin alternated with cyclophosphamide and dacarbazine. Sixty-six patients were randomized equally between the two regimens; one patient on regimen A was declared ineligible for the study shortly after randomization because of incorrect cell typing. The data were reported by Fleming and associates,[8] and survival distributions are shown in Figure 6-2.

Table 6-2.—Group Sequential Analysis of Data in Example 1 (See Text)

Test no.	Outcome with prednisone		Outcome with prednisone + vincristine		Unadjusted
(k)	Response	No response	Response	No response	*P* value
1	5	2	12	2	0.2159
2	4	3	13	1	0.0259
3	5	2	13	1	0.0095

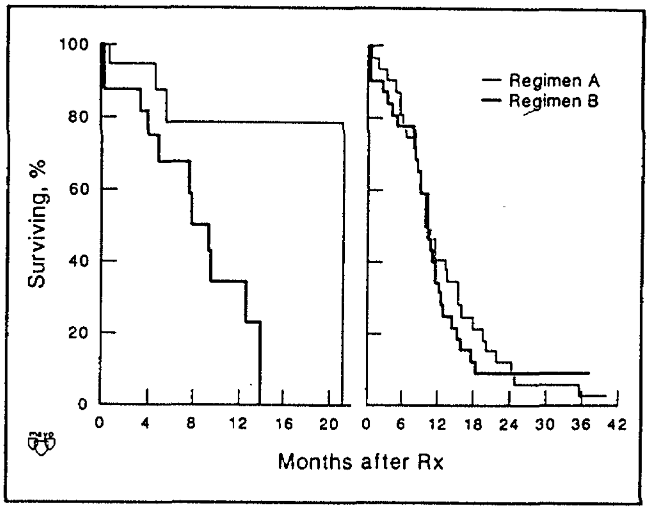

Fig. 6-2. Group sequential analyses. Estimated survival distributions for patients with extensive small cell lung cancer treated with two different regimens. Regimen A was cyclophosphamide, vincristine, VP-16, and cisplatin alternated with doxorubicin hydrochloride (Adriamycin) and dacarbazine. Regimen B was doxorubicin hydrochloride, vincristine, VP-16, and cisplatin alternated with cyclophosphamide and dacarbazine. *Left Panel,* Survival distributions on 9/12/77. *Right Panel,* Survival distributions on 7/15/79. (From Fleming and associates.[8] By permission of Elsevier Science Publishing Company.)

We next suppose that the investigators had incorporated provisions for three interim analyses into the study design. The results at the time of each of the analyses are shown in Table 6-3. With use of the Pocock boundary, statistical significance would have been achieved at the first analysis; hence, the trial would have been terminated and regimen A would have been declared superior. Conversely, with use of the modified O'Brien-Fleming boundary (using 0.001 for the initial analysis in place of 0.00005) or the Haybittle/Peto boundary, the investigators would have appropriately continued with the study and eventually would have found no evidence of a difference.

DISCUSSION

We have described three techniques for performing multiple statistical tests on accumulating data. The goal has been to assist investigators in deciding when to terminate a study early and how to make appropriate probability statements that will help determine whether one therapy is more efficacious than another. When any strategy is formulated for early termination of a study, there is one central question: Is it desirable to terminate a study early? On the surface, the obvious answer to this question may seem to be an unqualified "yes." Strong motivations exist for monitoring accumulating data. There is no question about the ethical need to terminate a study when the welfare of the patient is at stake and the superiority of one therapy over another has been clearly demonstrated. The opportunity to terminate a study early also may result in cost savings that, in some instances, may be substantial. Furthermore, an early answer to the scientific question at issue may facilitate the initiation of additional studies, as a result of the completed study either indicating new areas of investigation or releasing patients for participation.

From a scientific standpoint, however, often powerful incentives exist for continuing a trial. Because the ultimate goal of medical research is patient care, these incentives also translate into ethical arguments.

1. One of the disadvantages of early termination is that it may preclude obtaining satisfactory

Table 6-3.—**Group Sequential Analysis of Data in Example 2 (See Text)**

	No. of patients randomized				*P* value required for early termination		
Date	Regimen A	Regimen B	No. of deaths	Two-sided *P* value	Pocock	Modified O'Brien-Fleming	Haybittle/Peto
9/12/77	19	17	15	0.013	0.018	0.001	0.001
5/5/78	30	32	30	0.214	0.018	0.004	0.001
11/12/78	32	33	45	0.701	0.018	0.018	0.001
7/15/79	32	33	60	0.785	0.018	0.042	0.050

From Fleming and associates.[8] By permission of Elsevier Science Publishing Company.

answers to secondary (but nonetheless important) questions pertaining to drug safety and efficacy. Typically, many such questions could be posed, and the answers to these questions determine how the new therapy is implemented in clinical practice.

2. A related and equally important concern is the need to obtain a sample size sufficient for subgroup analyses because the efficacy and toxicity of a drug often will vary among different types of patients.

3. The use of early stopping strategies necessitates a corresponding strategy for obtaining unbiased estimates of drug efficacy. Because the study will be stopped early only at a time when one therapy seems to be considerably superior to the other, the usual estimates (for example, the proportion of patients who improved) will be biased. Because the special methods required to estimate efficacy are often extremely complicated, the need for such methods increases the difficulties of communicating study results to nonstatisticians.

4. The criterion that may be appropriate for addressing the immediate ethical question (which therapy should the physician select for the next patient?) may be less satisfactory for obtaining a definitive answer to the scientific questions. Although $P<0.05$ or $P<0.10$ may suffice for the ethical question, $P<0.01$ may be desirable for the scientific questions.

5. As illustrated in the second example, when the primary endpoint is survival, the long-term effect of therapy may not become apparent until later in the trial and may differ substantially from the early effects. For example, some types of therapy (operation or chemotherapy) may be sufficiently hazardous that early mortality (for example, at 1 year) might actually be increased, and a benefit may become apparent only much later (for example, at 5 to 10 years). The converse, an early transient benefit with no subsequent difference in survival, is also a possibility.

6. The interpretation of "sequentially adjusted" P values is often difficult to communicate to medical investigators. In fact, statisticians often disagree on this point.[9-15] For example, interpretation of the results of a group sequential trial may be controversial if the trial did not terminate early and did not achieve statistical significance with use of a group sequential boundary but the unadjusted P value was less than 0.05.

7. Finally, one must consider the substantial effort and cost involved in launching a well-designed clinical trial and the practical need of providing stable financing for both the investigators and their supporting staff. Uncertainty about continuity of funding for individual studies raised by the possibility of early termination may jeopardize the prospects for long-term research programs.

CONCLUSION

Is early termination of a study desirable? Rather than attempt a simple, universally acceptable answer to this question, we recommend that group sequential testing be incorporated into the study design but that a stopping boundary be chosen that appropriately reflects the perceived desirability of early termination. In practice, stopping rules are only guidelines to be used by those making the decision to stop a trial. Rarely, if ever, is a trial terminated solely on the basis of a statistical stopping rule.

ACKNOWLEDGMENT

We thank Gerald S. Gilchrist, M.D., and Thomas R. Lininger, M.D., and their colleagues for the use of their data.

REFERENCES

1. O'Brien PC, Shampo MA: Statistics for clinicians. 8. Comparing two proportions: the relative deviate test and chi-square equivalent. Mayo Clin Proc 56:513-515, 1981
2. Armitage P, McPherson CK, Rowe BC: Repeated significance tests on accumulating data. J R Stat Soc [A] 132:235-244, 1969
3. Pocock SJ: Group sequential methods in the design and analysis of clinical trials. Biometrika 64:191-199, 1977
4. O'Brien PC, Fleming TR: A multiple testing procedure for clinical trials. Biometrics 35:549-556, 1979
5. Haybittle JL: Repeated assessment of results in clinical trials of cancer treatment. Br J Radiol 44:793-797, 1971
6. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 34:585-612, 1976
7. Lininger TR, Fleming TR, Eagan RT: Evaluation of alternating chemotherapy and sites and extent of disease in extensive small cell lung cancer. Cancer 48:2147-2153, 1981
8. Fleming TR, Harrington DP, O'Brien PC: Designs for group sequential tests. Controlled Clin Trials 5:348-361, 1984

9. Dupont WD: Sequential stopping rules and sequentially adjusted *P* values: does one require the other? Controlled Clin Trials 4:3-10, 1983
10. Brown BW Jr: Comments on the Dupont manuscript. Controlled Clin Trials 4:11-12, 1983
11. Canner PL: Comment on "statistical inference from clinical trials: choosing the right *P* value." Controlled Clin Trials 4:13-17, 1983
12. Greenhouse SW: Discussion of early stopping. Controlled Clin Trials 4:19-21, 1983
13. Royall RM: Discussion of Dupont's "statistical inference from clinical trials: choosing the right *P* value." Controlled Clin Trials 4:23-25, 1983
14. Dupont WD: Rejoinder: statistical inference from trials with sequential stopping rules. Controlled Clin Trials 4:27-33, 1983
15. O'Brien PC: Group sequential methods in clinical trials. *In* Statistical Methodology in the Pharmaceutical Science. Edited by DA Berry. New York, Marcel Dekker (in press)

END OF STATISTICS SERIES