Impact of Multiple Comparisons in Randomized Clinical Trials

DAVID GARY SMITH, M.D.
Philadelphia, Pennsylvania
JOHN CLEMENS, M.D.
WILLIAM CREDE, M.D.
MARY HARVEY, M.D.
New Haven, Connecticut
EDWARD J. GRACELY, Ph.D.
Philadelphia, Pennsylvania

The randomized clinical trial is the preferred research design for evaluating competing diagnostic and therapeutic alternatives, but confidence in the conclusions from a randomized clinical trial depends on the authors' attention to acknowledged methodologic and statistical standards. This survey assessed the level of attention to the problem of multiple comparisons in the analyses of contemporary randomized clinical trials. Of the 67 trials surveyed, 66 (99 percent) performed multiple comparisons with a mean of 30 therapeutic comparisons per trial. When criteria for statistical impairment were applied, 50 trials (75 percent) had the statistical significance of at least one comparison impaired by the problem of multiple comparisons, and 15 (22 percent) had the statistical significance of all comparisons impaired by the problem of multiple comparisons. Although some statistical techniques are available, there still exists a great need for future work to clarify further the problem of multiple comparisons and determine how the impact of this problem can best be minimized in subsequent research.

The randomized clinical trial is the preferred research design for evaluating the efficacy of therapeutic interventions [1]. However, the use of random allocation should not minimize the attention to other required standards before inferences are drawn from the analyzed data. For example, recent work has suggested that a large proportion of randomized clinical trials may have reported false-negative findings due to inadequate statistical power [2]. Subsequently, investigators and readers have become more cognizant of the need to ensure adequate sample size and to evaluate the statistical power of results showing no significant differences [3]. Conversely, however, little attention has been given to the potential for false-positive results in randomized clinical trials [4].

Theoretically, a false-positive finding occurs when the analysis of research data from a randomized clinical trial yields significant statistical differences between differently treated groups when, in reality, none exists. However, the determination that a significant finding is falsely positive is often impossible in practice. Evaluation of published data often attends to possible biases due to methodologic problems, for example, unequal susceptibility of comparison groups, non-blinded treatment, and assessment of the outcomes [5]. If biases are found and if they impact directly on the conclusions, then it could be concluded that the finding may have been falsely positive [6,7].

An additional source of false-positive findings, however, stems from the probabilistic basis of most data analyses. Inferential reasoning requires some arbitrary standards to be defined before any statistical judgment can be made on the observed data [8,9]. An arbitrary point is

From the Departments of Medicine, Temple University School of Medicine, Philadelphia, Pennsylvania, Yale University School of Medicine, New Haven, Connecticut, and University of Alabama School of Medicine, Birmingham, Alabama, and the Department of Community and Preventive Medicine, Medical College of Pennsylvania, Philadelphia, Pennsylvania. This work was part of the Robert Wood Johnson Clinical Scholars Program at Yale University School of Medicine. This work was presented in part at the American Federation of Clinical Research, Washington, DC, 1983. Requests for reprints should be addressed to Dr. David Gary Smith, Temple University School of Medicine, 3401 North Broad Street, Philadelphia, Pennsylvania 19140. Manuscript submitted January 15, 1987, and accepted April 2, 1987.

TABLE I Probability of Finding at Least One False-Positive Result When Performing Multiple Comparisons

Number of Independent Comparisons	Probability of at Least One False-Positive Result
1	0.05
2	0.10
5	0.23
10	0.40
20	0.64

selected (the alpha value), which is the maximum probability of a result occurring by chance alone that a researcher is willing to accept. The usual alpha value has been chosen as $p = 0.05$. Thus, if the actual detected difference between differently treated groups would occur with a probability equal to or less than 0.05, then it is concluded that the experimental intervention and not random chance was the cause for this difference, assuming adherence to other methodologic standards. However, the caveat remains that one could have found a difference due to chance alone that would have fallen in the probability range equal to or less than 0.05, and in any given study, there is no way to determine if this has occurred. The likelihood of finding at least one false-positive difference increases with the number of comparisons made within the analysis of one research project. Data dredging or "fishing expeditions" are two of the current metaphors for making many comparisons in one research project; multiple comparison is the more neutral term for this problem [10,11].

The problem of multiple comparisons can be appreciated by using the formula, $P(FP) = 1 - (1 - \text{alpha value})^n$, where P(FP) is the probability of at least one false-positive finding, the alpha value is the arbitrary value used to protect against false-positive results, and n is the number of independent comparisons [4]. With use of 0.05 for the alpha value (the usual standard), the P(FP) can be estimated for different numbers of independent comparisons (**Table I**). For one independent comparison, the P(FP) is 0.05; however, as n increases, the P(FP) quickly increases until the likelihood of finding at least one false-positive difference becomes almost certain. For 20 comparisons, the P(FP) is 0.64, or simply stated, there would be a 64 percent chance of finding at least one statistically significant positive difference if, in fact, all 20 comparisons were in truth not different. This $p = 0.64$ is far greater than the usual standard of 0.05 (the alpha value) that is required to protect against false-positive results. The purpose of this survey is to document how frequently the problem of multiple comparisons could impair the statistical significance of therapeutic differences reported by contemporary randomized clinical trials.

METHODS

From January through June of 1982, all human research projects using random allocation of comparison regimens were selected from four weekly general medical journals: *Lancet*, *New England Journal of Medicine*, *Journal of the American Medical Association*, and the *British Medical Journal*. A list of the articles is available upon request from the authors.

Sources of Multiple Comparisons. The basic architecture of a randomized clinical trial involves the selection of an eligible sample from which persons are then allocated randomly to each of the comparison regimens. The actual assessment of the allocated regimen effects is the outcome comparison between two or more designated groups. Basically, the comparison can be directed either at the actual effects due to the different therapeutic regimens (i.e., therapeutic comparisons) or at characteristics that are unrelated to the impact of the allocated therapies (i.e., a non-therapeutic comparison). For the purposes of our study, only the impact of multiple therapeutic comparisons will be analyzed.

Multiple therapeutic comparisons often arise when different subgroups are defined on the basis of pertinent clinical characteristics. For example, the effect of antiplatelet therapy on the prevention of stroke can be assessed in treated versus placebo groups by comparing how all treated patients did versus control patients; then, the treated men can be contrasted with the control men, and the same comparison can be made with treated versus control women.

Another source of multiple therapeutic comparisons stems from the contrast of more than two regimens, e.g., comparison of three groups receiving amantidine, rimantidine, and placebo, respectively, thus resulting in at least four cogent pairwise comparisons: (1) rimantidine versus placebo; (2) amantidine versus placebo; (3) amantidine versus rimantidine; and (4) combined amantidine and rimantidine groups versus placebo.

Last, multiple therapeutic comparisons occur more frequently when investigators assess more than one outcome. For example, a randomized clinical trial evaluating cancer chemotherapy may evaluate the impact of therapy on mortality, tumor size, and quality of life.

Emphasized Comparisons. Since most randomized clinical trials analyze multiple therapeutic comparisons, we needed to define the most important analyzed comparisons relevant to the research questions addressed by the randomized clinical trial. To accomplish this, we designated an emphasized comparison as one that appeared in the abstract of the report. This special designation will permit assessment of the impact of the multiple comparison problem on only those reported comparisons that served as the primary basis for the authors' conclusions.

With use of standardized forms, all articles were reviewed and the following data collected: the number of reported therapeutic comparisons and the involved subgroups, regimens, and outcomes. Statistically significant therapeutic comparisons and all emphasized therapeutic comparisons were listed, and their associated p values were recorded. Any notation or statistical adjustment for the problem of multiple therapeutic comparisons was noted. Any disagreements between reviewers were discussed at weekly meetings, and the most conservative estimate was used for the disputed variable.

Criteria for Impairment of Statistical Significance. The statistical significance of a therapeutic comparison was judged "impaired" if all three of the following criteria were met: (1) the comparison was supported by a p value that was considered significant, but the p value was greater than 0.01; (2) more than five total therapeutic comparisons were made; and (3) the authors used no statistical technique to adjust the p values for having analyzed multiple therapeutic comparisons.

RESULTS

Sixty-seven trials were published in the reviewed medical journals during the interval from January 1982 to June 1982: *Lancet* (n = 33), *New England Journal of Medicine* (n = 12), *British Medical Journal* (n = 17), and *Journal of the American Medical Association* (n = 7). Fifty-six of the 67 trials were conducted within the Departments of Medicine; there were 14 multicenter trials. The mean number of patients per trial was 188, and there were seven trials with a crossover design.

In 66 of the 67 trials (99 percent), the analyses contained more than one therapeutic comparison, and there was a mean of 30.0 therapeutic comparisons per trial. In 52 (78 percent) of the reported trials, the analyses contained more than 10 therapeutic comparisons (**Figure 1**).

The main source of these multiple therapeutic comparisons was multiple outcomes with a mean of 21.7 different analyzed outcomes per trial. There was a mean of 1.6 subgroups and 2.7 regimens per trial. Despite the fact that 66 of 67 trials performed multiple therapeutic comparisons, only reports from two trials contained any statistical adjustments for multiple therapeutic comparisons.

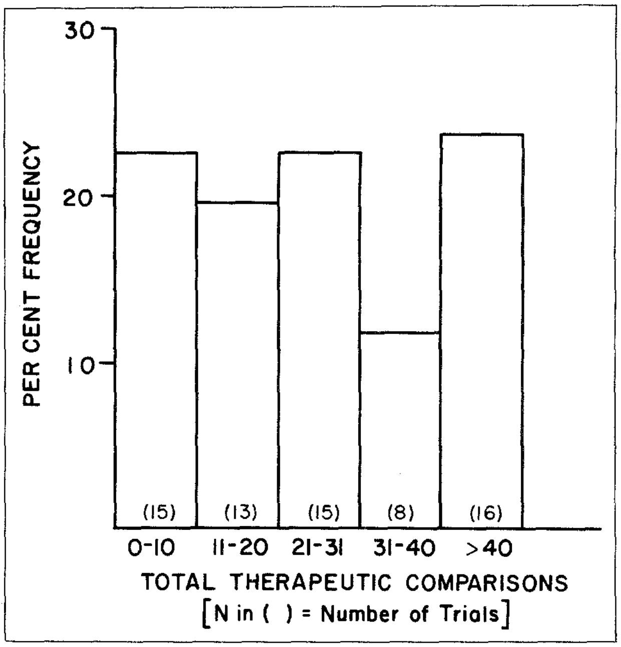

***Figure 1.** Percent frequency distribution of randomized clinical trials by total therapeutic comparisons.*

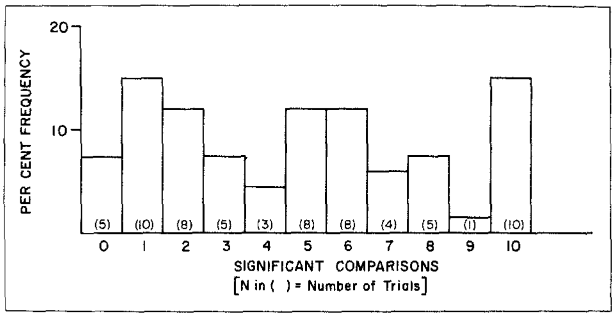

***Figure 2.** Percent frequency distribution of randomized clinical trials by total number of significant comparisons.*

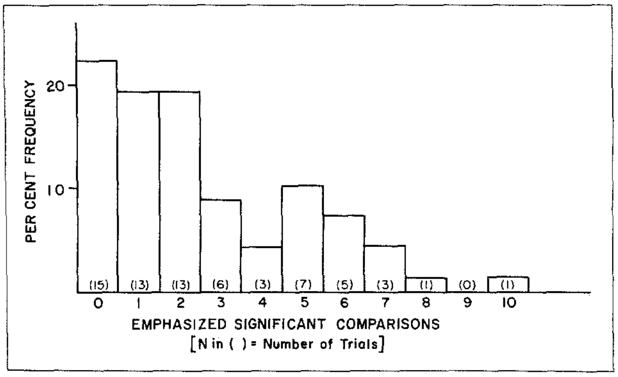

***Figure 3.** Percent frequency distribution of randomized clinical trials by total number of emphasized significant comparisons.*

Sixty-two of the 67 trials (93 percent) reported at least one statistically significant comparison, and 51 of these 67 trials (76 percent) reported an emphasized comparison that was supported by a p value less than 0.05. The mean number of statistically significant comparisons was 2.87 per trial for trials with 10 or less total therapeutic comparisons. The mean number of statistically significant comparisons increased to 4.69, 4.93, 5.38, and 11.88 in the following categories of total therapeutic comparisons: 11 to 20, 21 to 30, 31 to 40, and more than 40, respectively. Twenty-three of the 67 trials (34 percent) had fewer than three significant therapeutic comparisons (**Figure 2**), and 41 of the trials (61 percent) had fewer than three emphasized comparisons with statistical significance (**Figure 3**). Eight of the 67 trials (12 percent) had only one statistically significant comparison despite having analyzed more than five therapeutic comparisons. Although an average of 30 therapeutic comparisons was analyzed per trial, most trials had very few statistically significant comparisons and even fewer emphasized comparisons that were statistically significant.

Impact of Multiple Therapeutic Comparisons. When our criteria for impaired statistical significance were used, 50 of the 67 trials had at least one comparison whose claim of statistical significance was impaired. Of the 51 trials with emphasized comparisons, 40 had at least one impaired emphasized comparison, and all emphasized comparisons were impaired in 15 trials. In only one of the eight trials with only one statistically significant comparison was the comparison judged to be impaired by our criteria. Therefore, 74 percent of all trials were judged as having at least one impaired comparison, and 60 percent had at least one emphasized comparison that was impaired by the statistical problem of multiple comparisons. None of the trials judged as having impaired comparisons discussed the potential impact of the multiple comparison problem on their conclusions.

COMMENTS

This survey documents that multiple therapeutic comparisons are commonly presented in published reports of contemporary randomized clinical trials and that these multiple comparisons are primarily due to the many measures of outcomes used to determine the effect of the intervention. Only one of the published trials, however, employed statistical adjustments to prevent false-positive statistical appraisals despite the fact that a substantial number of the trials both reported and emphasized a "statistically significant" therapeutic difference whose "significance" is rendered questionable due to multiple comparisons. Although a recent survey found a similar inattention to this problem [12], our survey demonstrates the impact of this inattention on the conclusions from a significant proportion of contemporary randomized clinical trials.

Clinical Significance versus Statistical Significance. Much attention has been focused upon the distinction between clinically significant and statistically significant therapeutic differences in randomized clinical trials. The size of the therapeutic difference, large enough to be of clinical significance, varies with the particular patient subgroups, therapies, and outcomes under study. However, regardless of how large a difference is required for clinical significance, little credence can be placed in statistically significant therapeutic differences that could easily have arisen by chance alone. Because statistical significance is a prerequisite for the credibility of therapeutic differences large enough to be of clinical significance, and because statistical significance may be greatly overestimated when multiple therapeutic comparisons are analyzed, we focused exclusively on statistical significance in this survey.

Factors Influencing the Results of the Survey. Several factors could have influenced our estimate of the extent of the problem of multiple comparisons. Most important among these factors are the manner in which we assessed the total number of comparisons in each trial, the criteria by which we judged statistical significance to be impaired by the analysis of multiple comparisons, and the problem of non-independent therapeutic comparisons.

Assessment of the total number of comparisons: In order to formulate criteria for determining whether statistically significant therapeutic differences were impaired by the analysis of multiple comparisons, it was necessary to decide how many statistical comparisons were analyzed in each trial. The greater the total number of such comparisons, the more likely it was that a trial would detect statistically significant comparisons merely as a function of multiple analyses. To arrive at the total number of comparisons for each trial, we counted only therapeutic comparisons (i.e., comparisons involving contrasts of the therapeutic outcomes of at least two groups receiving different therapies), and we ignored other analyzed comparisons that did not contrast the outcomes of groups receiving different therapies (e.g., comparisons of the baseline characteristics of groups receiving different therapies). Although all analyses in a given study contribute to the likelihood of falsely finding a significant therapeutic difference, we omitted the non-therapeutic comparisons from our assessment so that the total number of comparisons that we used for our evaluation would be conservative. In addition, we evaluated only those therapeutic comparisons documented in the report rather than all possible therapeutic comparisons given the extensive data collection described in the methods section of each report. Since trials with greater numbers of therapeutic comparisons were more likely to receive negative ratings for inadequate attention to the problem of multiple comparisons, our conservative estimates of therapeutic comparisons undoubtedly caused an underestimation of the prevalence of trials having comparisons with impaired statistical significance.

Criteria for judging therapeutic comparisons to be impaired: We used a modification of the Bonferroni criterion to judge whether the alleged statistical significance of a therapeutic comparison was impaired by the analysis of multiple comparisons [5]. According to this criterion, when a study performs n independent comparisons and accepts $p = 0.05$ as the threshold of statistical significance, it is necessary for the difference to be significant at $p = 0.05/n$ to declare statistical significance if the study is to avoid an excess of falsely statistically significant comparisons. In our survey, we judged a therapeutic comparison to be impaired by multiple analyses if $p > 0.01$ and if the study analyzed more than five therapeutic comparisons. This criterion was in accord with Bonferroni limits, which would dictate that a study with five comparisons should accept only therapeutic differences with $p = 0.01$ in order to prevent at $p = 0.05$ the overall probability of accepting a false therapeutic difference [13].

Although statisticians have argued that Bonferroni adjustments for multiple comparisons are overly stringent, our adaptation of Bonferroni limits was lenient in two important ways. First, according to Bonferroni criteria, it is possible for the statistical significance of an individual comparison to be impaired even if five or fewer total comparisons are analyzed. For example, a therapeutic difference significant at $p = 0.04$ in the setting of four analyzed comparisons would exceed the Bonferroni limit of $p = 0.05/4 = 0.0125$. Thus, by considering only trials with more than five comparisons as being potentially impaired by multiple comparisons, we probably underestimated the number of impaired comparisons. Second, Bonferroni limits also suggest that it is possible to jeopardize the significance of an individual comparison even if the p value for that comparison is less than 0.01. This may happen if a large number of comparisons are statistically appraised. For example, if 20 comparisons are analyzed, the Bonferroni correction would be $p = 0.05/20$ or $p = 0.0025$. Again, we did not discredit therapeutic differences for which the associated p values were less than 0.01, despite the fact that the majority of trials in our survey analyzed more than 20 therapeutic differences. For these reasons, our survey probably underestimated rather than overestimated the impact of multiple comparisons upon statistical appraisals of therapeutic contrasts.

Nonindependence of therapeutic comparisons: When the p value threshold for "statistical significance" is lowered to correct for the analysis of therapeutic comparisons, the assumption is that these comparisons are independent of one another, i.e., the probability of obtaining statistical significance for one therapeutic contrast does not affect the probability of finding significance for other contrasts. Clearly, this assumption does not hold for many therapeutic comparisons in a clinical trial. For example, if the efficacy of two antibiotics in the treatment of bacterial pneumonia were compared, the comparisons of the antibiotics with respect to their efficacy in clearing radiographic infiltrates, in resolving fever, and in lowering the white blood cell count are highly interrelated. When therapeutic comparisons are interrelated, it is unnecessarily stringent to demand that each comparison meet a lower p value threshold calculated under the assumption of independent comparisons.

Although statisticians have developed methods for circumventing problems due to interrelated outcomes, none of the trials in our study used such techniques. Because it was impossible to determine from the reported data whether comparisons were independent of one another, we made no attempt to discriminate between independent and non-independent therapeutic comparisons. Consequently, we may have overestimated the number of statistically independent comparisons for some of the trials. However, it is unlikely that this problem caused us to exaggerate the frequency of therapeutic comparisons whose statistical significance was impaired by analyzing multiple comparisons, since, as we described earlier, we deliberately employed several tactics that underestimated the total number of comparisons in most of the trials. In addition, among the 40 trials reporting significant comparisons, 26 reported a total of two or fewer significant comparisons in the context of more than 20 analyzed therapeutic comparisons. We therefore do not believe that the problem of non-independence was serious enough to have substantially affected our analysis.

Need for Research and Consensus. Although a variety of statistical solutions are available to adjust for multiple comparisons, most of these techniques are suited only for outcome variables expressed on a dimensional or continuous scale (e.g., blood pressure) and are not useful for categoric outcome variables (e.g., therapeutic response versus no therapeutic response) that are typically studied in randomized clinical trials [14]. For this reason, further statistical work is needed to develop methods that handle multiple comparisons in the types of analyses performed in contemporary randomized clinical trials.

However, beyond improvements needed in statistical methodology, there remain several unresolved issues whose solutions will emerge only from scientific consensus and the development of scientific policy. Two dilemmas relevant to the analysis of clinical trial data are prominent. First, the definition of a "comparison" for the purpose of analyzing multiple comparisons must be made specific. The definition should clarify whether all comparisons or only therapeutic comparisons should be considered in statistical adjustments. The definition should also specify whether all comparisons, or merely those that have been statistically appraised and/or reported, need to be considered in the adjustments.

Second, a consistent policy is needed for analyses of comparisons that are formulated before the research, and the analyses that yield unanticipated differences that arise after inspection of the data. It has been argued that only comparisons anticipated in designing the study should be appraised with p values [15,16]. However, much valuable information would be lost if analyses of randomized clinical trials are restricted only to anticipated comparisons. It is often important, for example, to explore a significant overall therapeutic difference by considering therapies within patient subgroups that were not anticipated in the design of the study. Conversely, even if no overall significant difference is detected, subgroup and other analyses may suggest hypotheses for future research. Strictly speaking, no statistical test can calculate the probabilities of false-positive results for these data-dependent analyses [17]. Nevertheless, since these analyses are scientifically useful, and since investigators require some means to evaluate the likelihood that findings in these analyses have arisen by chance, a uniform policy is required for the statistical management of these analyses and for the permissible ways in which the results of these analyses may be reported in publications.

The results of our survey suggest that contemporary randomized clinical trials are at risk of reporting false-positive therapeutic differences due to a lack of attention to problems created by multiple comparisons. It is not surprising that so few published trials are cognizant of the problem of multiple comparisons, since the nature of the problem is not widely known, appropriate statistical methods have not been extensively developed, and uniform policies have not been elucidated for defining and counting the number of comparisons in a particular study and for handling data-dependent analyses. The development and dissemination of improved statistical methods and the enunciation of a clear scientific policy for dealing with multiple comparisons should therefore assume great priority in the future of randomized clinical trials.

Clearly, the problem of multiple comparisons should be evaluated in other research designs, especially when the investigator is engaged in hypothesis-generating research. Any attempt to deal with the problem of multiple comparison must also attend to the strongly held perception that only "significant results" (i.e., $p < 0.05$) will be accepted for publication. An additional factor is the availability of statistical software, which greatly eases the performance of multiple comparisons with simple commands compared with the old days of laboring with a hand-held calculator or a lead pencil. Some software systems give explicit warnings to investigators, e.g.:

> . . . By making it more difficult for you to do a slew of uncritical automatic analyses, I am trying to help you preserve the meaning of probabilities. P values are a health hazard. The more of them you see on a computer printout, the less meaningful they are . . . [18].

This policy should be standard. Last, editorial policy will probably be the most effective way of sensitizing readers and purveyors of research to these important statistical problems.

REFERENCES

1. Spodick D: The randomized controlled clinical trial, scientific and ethical basis. Am J Med 1982; 73: 420–425.
2. Freiman J, Chalmers T, Smith H, Kuebler P: The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. N Engl J Med 1978; 299: 690–694.
3. Young M, Bresnitz E, Strom B: Sample size nomograms for interpreting negative clinical studies. Ann Intern Med 1983; 99: 248–251.
4. Tukey J: Some thoughts on clinical trials, especially problems of multiplicity. Science 1977; 198: 679–684.
5. Feinstein A: Clinical biostatistics. St. Louis: CV Mosby, 1977.
6. Braitman L: Statistical, clinical and experimental evidence in randomized controlled trials (editorial). Ann Intern Med 1983; 98: 407–408.
7. Diamond G, Forrester J: Clinical trials and statistical verdicts: probable grounds for appeal. Ann Intern Med 1983; 98: 385–394.
8. Lee K, McNeer J, Starmer C, Harris P, Rosati R: Clinical judgment and statistics. Circulation 1982; 61: 508–516.
9. Moesteller F, Gilbert J, McPeek B: Controversies in design and analysis of clinical trials. In: Shapiro S, Louis T, eds. Clinical trials, chap 1. New York: Dekker, 1983; 13–64.
10. Ryan T: Multiple comparisons in psychological research. Psychol Bull 1959; 56: 26–47.
11. Selvin H, Stuart A: Data-dredging procedures in survey analysis. Am Statistician 1966; 20: 20–23.
12. Godfrey K: Comparing the means of several groups. N Engl J Med 1985; 313: 1450–1456.
13. Jones D, Rushton L: Simultaneous inference in epidemiological studies. Int J Epidemiol 1982; 11: 276–282.
14. Cupples L, Herren I, Schatzkin S, Colton T: Multiple testing of hypotheses in comparing two groups. Ann Intern Med 1984; 100: 122–129.
15. Murphy E: Skepsis, dogma, and belief. Baltimore: Johns Hopkins University Press, 1981.
16. Murphy E: The analysis and interpretation of experiments: some philosophical issues. J Med Philos 1982; 7: 307–325.
17. Jekel J: Should we stop using p value in description studies? (Editorial). Pediatrics 1977; 60: 124.
18. Wilkinson L: SYSTAT: The system for statistics. Evanston, Illinois: SYSTAT, Inc., 1986.