STATISTICS IN PRACTICE

COMPARING THE MEANS OF SEVERAL GROUPS

KATHERINE GODFREY, A.M.

Abstract This article discusses statistical methods for comparing the means of several groups and focuses on examples from 50 Original Articles published in the *Journal* in 1978 and 1979. Although medical authors often present comparisons of the means of several groups, the most common method of analysis, multiple t-tests, is usually a poor choice. Which method of analysis is appropriate depends on what questions the investigators wish to ask. If the investigators want to identify which of the groups under study are different from the rest, they will need a different method from the one required if they wish simply to decide whether or not the groups share a common mean. More complicated questions about the group means call for more sophisticated techniques. Of the 50 *Journal* articles examined, 27 (54 per cent) used inappropriate statistical methods to analyze the differences between group means. Investigators need to become better acquainted with statistical techniques for making multiple comparisons between group means. (N Engl J Med 1985; 313:1450-6.)

IN one simple form of experiment or observational study, the investigator compares sets of measurements taken from two groups to decide whether the group means differ. Emerson and Colditz[1] have reported that the t-test, the standard analysis for such an experiment, is the most commonly used statistical procedure in the *Journal.*

When an experiment includes more than two groups, the choice of an appropriate statistical method for comparing group means depends on the experimental design and on the questions asked. Although investigators have many options, few are appropriate, and the most frequently used method, multiple t-tests, is usually a poor choice. Moreover, even appropriate methods of analysis may not directly answer the investigators' questions. A poorly chosen analysis may generate misleading results by giving incorrect answers to the investigators' questions, by giving correct answers to the wrong questions, or by failing to use all the information available from the experiment. Care in formulating the study questions and in choosing the method of analysis can prevent such costly mistakes.

Because computer programs offer ready numerical solutions to formerly obscure and difficult analytical problems, investigators now need only to match the type of data and the goal of the study to the method in order to obtain an adequate analysis. Using examples from the *Journal,* this article discusses some standard statistical methods for comparing group means and describes which questions about group means each method is designed to answer. Some standard methods of statistical analysis include multiple t-tests, analysis of variance, and multiple comparisons. More sophisticated techniques are available to answer complicated questions about group means.

In this article, I do not consider the mechanics of the calculations involved in the various techniques. Standard textbooks, such as those by Snedecor and Cochran,[2] Armitage,[3] Dixon and Massey,[4] Bliss,[5] Sokal and Rohlf,[6] Winer,[7] Kleinbaum and Kupper,[8] and Brownlee,[9] provide details of the calculations.

METHODS

The 332 Original Articles published in Volumes 298 to 301 (calendar years 1978 and 1979) of the *Journal* provided the examples used in this article. In these four volumes, 50 Original Articles[2,10-59] included a comparison of several group means. One article[52] contained 3 separate analyses, making a total of 52 analyses. I have not included in this survey articles comparing observed proportions and have considered only articles dealing with means of measurements. Table 1 gives information on the types of analysis reported in these articles.

I examined all Original Articles in the four volumes of the *Journal* to determine which contained comparisons of group means. To make sure no articles were omitted, this list of articles was then checked against a similar list prepared for Emerson and Colditz.[1] I then read each article to compile the information presented in Table 1.

An expanded version of Table 1, with details on the statistical analyses performed in each of the 50 articles, appears elsewhere.[60]

WHY NOT t-TESTS?

Investigators comparing three or more group means at once frequently examine each possible pair of groups separately, using the t-test to examine each pair. For example, Toft et al.[10] examined estimated thyroid-remnant weight in patients in four postoperative states: euthyroidism with normal serum thyrotropin levels; euthyroidism with raised serum thyrotropin levels; temporary hypothyroidism; and permanent hypothyroidism. Examining each pair of groups would involve six tests: Groups 1 versus 2, 1 versus 3, 1 versus 4, 2 versus 3, 2 versus 4, and 3 versus 4. In general, any experiment with k groups has $k(k - 1)/2$ different pairs available for testing.

The statistical methods most often used, such as the t-test, are designed for use with a single comparison or test; thus, when they are used for several tests at once, as in the case just described, multiplicity is a factor that needs to be considered. Clearly, the more tests are made, the more chances exist for unusually high or low values to occur. If the testing system fails to take account of this multiplicity, then the investigator will tend to interpret extreme results without the necessary grain of salt.

Investigators often label as "statistically signifi-

From the Department of Statistics, Harvard University, Cambridge, MA 02138, where reprint requests should be addressed.

Supported by a grant (RF 79026) from the Rockefeller Foundation.

Table 1. Type of Analysis Reported in 50 Articles Comparing Group Means in Volumes 298 to 301 of the *Journal.*

Analysis Performed	Figure Included Showing Results of Analysis	No Figure Provided	Total No. of Analyses*
Multiple-comparisons analysis	2	2	4
Analysis of variance	1	13	14
Both multiple comparisons and analysis of variance	1	2	3
Neither multiple comparisons nor analysis of variance	15	16	31
All analyses	19	33	52

*One article[52] included three different analyses, making a total of 52 for the 50 articles.

cant" differences so extreme that they would occur in less than 5 per cent of the tests when two groups were drawn from populations with identical means. When several tests are made, however, the chance of at least one such extreme result increases rapidly with the number of tests. The exact formula is not important. The main point is well illustrated by a case in which there are four groups and thus six comparisons; in this situation, the level of significance is closer to 21 per cent than to 5 per cent. Thus, unless account is taken of multiplicity, the investigator may be mistakenly impressed by the seemingly extreme (and thus seemingly rare) result.

Table 2 shows the increase in the probability that extreme values will occur as the number of groups, and of comparisons, increases. For seven groups, and 21 tests, the chance of finding at least one result that could be labeled "statistically significant" is nearly half, even if all populations have identical means. Such an extreme result is therefore about 10 times as likely as the 5 per cent significance level suggests.

Table 3 shows the numbers of groups studied in the 50 articles. Two thirds of the analyses (35 of 52) involved either three or four groups. Only two studies compared more than six groups.

Of the 50 articles comparing group means, more than half (27) used only multiple t-tests to make comparisons among groups and thus took no direct account of the problem of multiplicity. Although the technique of using many t-tests appears simple, the results are hard to interpret.

One method for taking account of multiplicity adjusts the testing so that the simultaneous risk of finding one or more spurious significant results in all the t-tests combined has the chosen significance level — for instance, 5 per cent. That is, we alter the criterion for individual comparisons so that the probability of finding at least one significant test result for the entire experiment (looking at all the t-tests simultaneously) is 0.05 when all the underlying group means are equal. This approach controls the chance of error per experiment instead of per test. A common way to approximate this method is to divide the desired simultaneous significance level for the experiment as a whole by the number of tests being made; the result is a new significance level to be used for each of the individual t-tests. This is called the Bonferroni method. For an experiment with five groups, there are 10 tests of pairs, so with a simultaneous significance level of 0.05, the Bonferroni method leads us to conduct each of the 10 individual tests with a significance level of 0.05/10 = 0.005.

The Bonferroni method changes the way we perceive the experiment, because the error rate for each t-test in the framework of the multiple t-tests is replaced by the error rate for the entire experiment of all questions posed. When we make more tests, we need to be more conservative in making each individual test if we want to control the error rate for the experiment as a whole.

The Bonferroni adjustment is deliberately conservative; that is, it keeps the significance level for the entire experiment at least as low as intended. Estimated confidence intervals for differences between group means may therefore be too wide, and the analysis may not identify a difference between two means that would be statistically significant if the simultaneous significance level were exactly the intended value. This approximation results in a lack of statistical power — that is, we may have a reduced chance of detecting small but real differences among the groups. It gives us a good way of thinking about the problem, but other methods are preferable.

Analysis of Variance

An analysis of variance answers the question whether there are differences among the population means of the groups being compared, but it does not pinpoint which populations, if any, differ from the others. For example, Perrin and Goodman[14] compared the way in which three groups of medical personnel (house officers, practicing pediatricians, and nurse practitioners) handled telephone calls from the parents of acutely ill children. Each worker received a score for completeness of telephone interviewing based on how much information he or she collected over the phone. The authors tested for differences among the three groups by using an analysis of variance, which indicated that differences in interviewing ability did exist.

Table 2. Probability of Finding at Least One Comparison Significant at the 0.05 Level When Testing All Possible Pairs of Groups.

No. of Groups	No. of Tests	Probability
2	1	0.05
3	3	0.11
4	6	0.21
5	10	0.30
6	15	0.39
7	21	0.47

Table 3. Distribution of the Numbers of Groups Studied in 50 Articles Comparing Group Means in Volumes 298 to 301 of the *Journal*.

No. of Groups Studied	No. of Analyses*
2	2
3	17
4	18
5	4
6	9
10	1
15	1

*One article[52] included three different analyses with different numbers of groups for each analysis; thus the total in column 2 is 52.

Analysis of variance generalizes the t-test from two groups to three or more groups. It replaces multiple t-tests with a single F test of the assumption that the underlying group population means are all equal; the single test takes proper account of the multiple-comparison problem. The logic of the F test is as follows: If all the groups have a common population mean, M, then the observed group means should all lie near M. If the means are sufficiently dispersed, then the critical quantity called the F statistic is significant, and we conclude that at least one of the population means for the groups differs from the others. By itself, however, analysis of variance does not tell us which groups differ from which others.

In some cases, analysis of variance is a very useful technique. When comparing several groups to evaluate treatment effects, for example, investigators want to make sure that the groups are similar in "covariates" — characteristics other than treatment that may affect the outcome. For example, age and sex often affect the way patients react to treatment, and investigators should assure themselves that the groups being compared are similar in age and sex when these factors may affect the outcome. Analysis of variance is well suited for testing the usual assumption that a number of treatment groups have comparable population means for variables other than the main variable of interest. For instance, in the study of medical personnel described above,[14] we may wish to see whether the three groups had comparable mean years of experience. Analysis of variance provides a simple way to test for differences of this sort.

Analysis of variance offers a standard method for comparing various groups when there is no presumption beforehand that they differ. We can use this method to compare drug therapies or disease groups that we consider to be comparable or "about the same." For example, Griffiths et al.[50] performed an analysis of variance to test whether the mean ages of patients in three endoscopic categories were similar. When this test proved nonsignificant, they went on to make another analysis of variance to look for differences among the three groups in transfusion requirements, without adjusting for age.

The authors used analysis of variance to compare several group means in about 30 per cent (14 of 50) of the *Journal* articles.

Multiple Comparisons

An investigator may well have an idea what the results will be, or which questions are most important, before the study begins. For example, when comparing the efficacy of several treatments with that of a placebo, he or she may expect that the active treatments will have similar effects but that each will be more effective than the placebo. An example is an experiment by Thadani et al.,[47] who compared the performance of five beta-adrenergic–blocking drugs in the treatment of angina with that of a placebo. The authors found that the responses of the patients given the placebo seemed quite different from those of each of the other treatment groups, but that the drug-treated groups were all similar.

What is an appropriate analysis for such an experiment, with one placebo group and several treated groups? In the context of analysis of variance, we ask a single question of the data: Do the groups have equal means? An ordinary analysis of variance simultaneously compares all six groups and, in the instance just described, confirms differences among the groups, but it does not tell us which groups differ from which others. If we want more information than analysis of variance can provide, we need to ask more questions. Multiple-comparison methods allow us to do this and still control the overall significance level. We can find all the pairwise differences among the group means and test them individually for significance, or we can separate the group means into clusters of "like" means, so that all the means in one cluster differ significantly from all the means in any other cluster. Multiple-comparison methods can also test certain weighted sums of the group means; using this technique, we multiply each group mean by its own constant (a weight) and then add these products together. A difference between two group means is only one example of such a weighted sum; the weights for such a difference are 1 and −1 for the two group means in the difference, and 0 for each of the other means. More complicated examples are described below.

Multiple-Comparison Techniques

As described above, the Bonferroni method adjusts t-tests to make each test more stringent, but it is conservative and may have low statistical power. For this reason, methods that are more nearly correct in their significance levels have been developed for making multiple comparisons among pairs of group means. These methods attempt to keep the overall significance level of the experiment at the intended limit (for instance, 0.05), while making it less likely that a true population difference between two groups will be missed. Several of these methods are discussed briefly below: those of Scheffé, Tukey, Newman–Keuls, and Duncan. Both the SAS[61] and SPSS[62,63] statistical computing packages provide calculations for each of these methods. Miller[64] discusses methods for making multiple comparisons in detail. Each method assumes that the data are normally distributed and that the true (but unknown) variance within each group is the same. Games et al.[65] survey recent developments in multiple-comparison techniques. Of the 50 articles I studied, 7 used a multiple-comparison technique.[13,21,25,30,51,53,58]

Scheffé's test allows the investigator to examine the data and then to choose one or more weighted combinations of means to test, without bias to the results. To do this, the test must allow for many different weighted combinations, because investigators may be imaginative in choosing a weighted sum. The weights must add up to zero, so that the value of the sum will be zero when all the group means are equal. Such sums are called contrasts. For example, in a study of motor ability in children in grades one, two, and three, we might ask whether the second-graders' performance differed from that of the average of the children in the first and third grades. Then, in the usual notation, we are asking about

$$\bar{x}_2 - \tfrac{1}{2}(\bar{x}_1 + \bar{x}_3).$$

The weights for this contrast are $-\frac{1}{2}$ for $\bar{x}_1$, 1 for $\bar{x}_2$, and $-\frac{1}{2}$ for $\bar{x}_3$. The method must be able to look at any contrast, because it must allow for the variability of all possible contrasts, even one, such as $0.2\bar{x}_1 + 0.7\bar{x}_2 - 0.9\bar{x}_3$, that few investigators are likely to consider.

Scheffé's method is the least likely of the multiple-comparison techniques to identify differences, because the investigator is allowed to look at many more differences, in the form of contrasts, than the other methods allow. The invention of the method was a milestone in statistics; it allowed the investigator to peek at the data before choosing the contrast to report on and still report an honest significance level. The price of this advantage was that the test had to allow for every possible contrast, even though few were actual candidates for study.

One of the 50 articles in my survey used Scheffé's method for multiple comparisons.[25]

Tukey's method exemplifies those that test for differences among group means by using the difference between the largest and smallest means, often called the range, as a measure of their dispersion. Although Tukey's method, like Scheffé's, can be used to obtain confidence limits for all contrasts, it can also be used to set limits only on differences. This second use reduces the impact of multiplicity on significance levels when contrasts other than simple differences have no interest for the investigator. Tukey's method gives narrower confidence limits than does Scheffé's method. When used for all possible contrasts, however, Tukey's method has the disadvantage of giving wider confidence intervals than Scheffé's method.

Tukey's method uses special tables, comparable to but different from the F tables. These tables are available in many textbooks, including Snedecor and Cochran's.[2] None of the 50 *Journal* articles I examined used Tukey's method.

The Newman–Keuls and the Duncan methods take a different approach. They create clusters of group means that might reasonably be drawn from populations with identical means and that may overlap. For a five-group study, if we numbered the observed means from 1 through 5 in order of increasing size, it might turn out that two clusters would be Groups 1, 2, and 3 and Groups 3, 4, and 5. Thus, the largest three means or the smallest three means might reasonably form clusters from the same population, but in this instance no four groups would form such a cluster. The actual process of constructing the clusters requires a sequential set of comparisons, which I shall not describe. Both methods make use of the same tables used for Tukey's method.

Of the two, the Newman–Keuls method is more conservative than Duncan's test. Newman–Keuls was used once[23] and Duncan four times[13,30,51,58] in the 50 *Journal* articles comparing group means.

Although multiple-comparison methods usually involve considering all possible pairs of differences among the groups, they are well suited to examining a few selected differences. In the experiment reported by Thadani et al.,[47] we might be interested only in the 5 differences between each drug and the placebo and not in all 15 possible differences among the six study groups. We can use multiple-comparison techniques to test these preselected differences without calculating the others. For example, we can use Bonferroni-adjusted t-tests, taking as our divisor for the significance level the number of tests we make, five. When we test only a few differences, the Bonferroni method works fairly well. Note that we are testing for differences chosen for study before the experiment begins, not those suggested by the collected data. The latter case is discussed below.

Questions Suggested by the Data

Ideally, investigators should decide which statistical tests they will perform, including which groups to compare, before they examine the data in even a cursory fashion. In practice, however, the data in hand may suggest comparisons that were not originally planned. If groups that the investigators expected to be similar have quite different means, for instance, there may be reasons to test this apparent difference. Investigators who are careful in choosing multiple-comparison methods can make this new comparison without changing the overall significance level for the experiment.

If we choose to test only the significance of the larger observed mean differences in an experiment, the probability of finding apparently significant differences will be greater than if we test pairs of means chosen at random before the experiment began. In such a case, the overall significance level for the experiment will actually be higher than we determined beforehand. The more conservative multiple-comparison methods allow investigators to select which group means to test after the experiment begins, by allowing for all possible comparisons. If group means are chosen for comparison testing on the basis of their apparent differences after the data have been collected, it is important to use these conservative methods. As mentioned above, Scheffé's method, in particular, was designed to permit this sort of "data dredging."

In summary, a more conservative method gives broader confidence intervals, is less likely to report a difference between means when none exists, and is more likely to miss a real difference. Thus, it has lower statistical power than a less conservative method. A less conservative method, in turn, has a better chance of detecting a small difference but also has a greater chance of reporting a difference that is not real.

Two-Way Analysis of Variance

A one-way analysis of variance assumes that the groups are at the same "level" in some hierarchy. In other words, the groups should be distinct, comparable, and of equal stature. This is not always true for the groups we wish to compare, however. There may be several different levels within a group. For instance, three different drug treatments, A, B, and C, administered at dosages A1, B1, and C1, respectively, may be taken as being at the same level, but an experiment that includes three additional treatments — drug A at dosage A2, drug B at dosage B2, and drug C at dosage C2 — is more complicated in that it includes two different dosages of each of the drugs in the design. A one-way analysis of variance will not allow us to test the effect of different dosages of the same drug; it allows us only to compare all six groups at once. A multiple-comparison method may help answer questions about how the effect of a single drug varies with dosage, but the comparability of these six treatment groups remains unclear. A better method is an analysis of variance that permits the dose level to be included as a variable in the analysis. This method allows for one F test of the effect of differing dosage in all the groups simultaneously, as well as a separate F test comparing the three drugs. An analysis of variance that compares group means in this way, across two separate categorizing variables, is called a two-way analysis of variance.

In a study by the Veterans Administration Cooperative Study Group on Antihypertensive Agents,[51] each of two antihypertensive drugs (ticrynafen and hydrochlorothiazide) was administered at two different dosages. The authors conducted a one-way analysis of variance comparing all four treatment groups. A two-way analysis of variance would have allowed the authors to look at the effects of dosage as well as of drug.

Carmel and Johnson[17] considered three categories of anemic patients (European, black, and Latin American) and also examined the effect of sex. The authors used more than a dozen t-tests to try to get at the effects of ethnic group and sex. A two-way analysis of variance would have allowed them to estimate the effects of sex and ethnic group separately, looking at all the data at once, as well as to measure the joint effect of sex and ethnic group over and above their separate effects.

Two of the 50 articles in my survey did use a two-way analysis of variance. McLellan et al.[59] analyzed patients in three drug-abuse groups by means of psychological testing and retesting. The two-way analysis permitted the investigators to use the scores on the test and the retest simultaneously in comparing the three groups. It also compared the test and retest scores themselves. Adams et al.[37] studied levels of female sexual activity according to contraceptive method and segment of the menstrual cycle (first or second half). This analysis could have been extended to a three-way analysis of variance if the authors had included another variable analyzed elsewhere in the article, type of sexual activity. This approach would have allowed them to examine the effects of all three variables simultaneously.

Other Approaches

Depending on what questions researchers wish to answer, they can design their experiments in different ways. In the thyroid experiment described earlier,[10] the groups might have fallen into two, three, or four separate categories, depending on the way the researchers viewed the data. Each group could have been considered separately, giving four categories. Another possible design would be two categories — euthyroid and hypothyroid — each with two subcategories. There could also be three categories — euthyroid with two subcategories, temporarily hypothyroid, and permanently hypothyroid. Each different design calls for its own analysis. We have discussed only three types of analysis-of-variance designs: one-way, two-way, and three-way. Brownlee[9] and Winer[7] discuss other analysis-of-variance designs and give examples of the calculations involved.

Possible Difficulties in Comparing Group Means

Both the multiple-comparison methods and analysis of variance assume equality of variances in the different groups. This assumption allows for a pooled estimate of the common group variance, using the variances of each group, that is more precise than any of the individual group variances. Although the required F test is not greatly affected by small differences in group variances, the data should be checked for large differences in group variances before the means are compared. Several tests have been devised for this purpose, including Hartley's F-max, a test based on the ratio of the largest group variance to the smallest group variance. Details are available in textbooks such as Snedecor and Cochran's.[2] The BMDP computing package[66] provides another such method, Levene's test, also described by Snedecor and Cochran.

The variances of groups compared in the *Journal* articles in my survey often differed substantially, raising questions about the accuracy of the analysis. If one group variance is much larger than the others, it will increase the estimate of the pooled variance, making it more difficult to detect differences among the groups with small variances. Sometimes converting, or "transforming," the measurements to another scale can make the variances more nearly equal. For example, if the variance of the group increases as the mean of the group increases, taking logarithms or square roots of the original data may make the variances more nearly equal in the transformed scale. The analysis then proceeds in the new scale. Adams et al.[37] used the F-max test to compare the variances of their study groups. The disparity in variances led them to take square roots of the data before performing an analysis; this technique corrected the disparity.

One possible drawback to the use of such transformations is that all the results of the analysis refer to the transformed data, not the original data. Sometimes the units of the transformed data cannot be interpreted clearly. For example, reciprocals of data on the time required to complete some action give the speed of completion, but the square roots of such data have a less obvious meaning.

When we have evidence that the assumptions about normality and equal variance do not hold, we may want to use nonparametric techniques to test for differences among means. Nonparametric methods do not require the usual assumption that the data are normally distributed and so are appropriate when it is likely that the data do not meet that assumption. One such method is the Kruskal–Wallis test, available on SAS, SPSS, and BMDP. Sokal and Rohlf[6] give an example.

No matter how the data are analyzed, it is almost always useful to provide a plot of the results. Often, a plot speaks so clearly that the message is obvious regardless of the method. Differences among groups are instantly apparent, and outlying values are easily recognized.

References

1. Emerson JD, Colditz GA. Use of statistical analysis in *The New England Journal of Medicine.* N Engl J Med 1983; 309:709-13.
2. Snedecor GW, Cochran WG. Statistical methods. 7th ed. Ames, Iowa: Iowa State University Press, 1980.
3. Armitage P. Statistical methods in medical research. New York: John Wiley, 1971.
4. Dixon WJ, Massey FJ. Introduction to statistical analysis. 3rd ed. New York: McGraw-Hill, 1969.
5. Bliss C. Statistics in biology: statistical methods for research in the natural sciences. Vol. 1. New York: McGraw-Hill, 1967.
6. Sokal RR, Rohlf FJ. Biometry: the principles and practice of statistics in biological research. San Francisco: WH Freeman, 1969.
7. Winer BJ. Statistical principles in experimental design. 2nd ed. New York: McGraw-Hill, 1971.
8. Kleinbaum DG, Kupper LL. Applied regression analysis and other multivariable methods. North Scituate, Mass.: Duxbury Press, 1978.
9. Brownlee KA. Statistical theory and methodology in science and engineering. 2nd ed. New York: John Wiley, 1965.
10. Toft AD, Irvine WJ, Sinclair I, McIntosh D, Seth J, Cameron EHD. Thyroid function after surgical treatment of thyrotoxicosis: a report of 100 cases treated with propranolol before operation. N Engl J Med 1978; 298:643-7.
11. Reisin E, Abel R, Modan M, Silverberg DS, Haskel HE, Modan B. Effect of weight loss without salt restriction on the reduction of blood pressure in overweight hypertensive patients. N Engl J Med 1978; 298:1-6.
12. Potkin SG, Cannon HE, Murphy DL, Wyatt RJ. Are paranoid schizophrenics biologically different from other schizophrenics? N Engl J Med 1978; 298:61-6.
13. Ibels LS, Alfrey AC, Haut L, Huffer WE. Preservation of function in experimental renal disease by dietary restriction of phosphate. N Engl J Med 1978; 298:122-6.
14. Perrin EC, Goodman HC. Telephone management of acute pediatric illnesses. N Engl J Med 1978; 298:130-5.
15. Cohn WJ, Boylan JJ, Blanke RV, Fariss MW, Howell JR, Guzelian PS. Treatment of chlordecone (kepone) toxicity with cholestyramine: results of a controlled clinical trial. N Engl J Med 1978; 298:243-8.
16. Dreisin RB, Schwarz MI, Theofilopoulos AN, Stanford RE. Circulating immune complexes in the idiopathic interstitial pneumonias. N Engl J Med 1978; 298:353-7.
17. Carmel R, Johnson CS. Racial patterns in pernicious anemia: early age at onset and increased frequency of intrinsic-factor antibody in black women. N Engl J Med 1978; 298:647-50.
18. Fillit HM, Read SE, Sherman RL, Zabriskie JB, van de Rijn I. Cellular reactivity to altered glomerular basement membrane in glomerulonephritis. N Engl J Med 1978; 298:861-8.
19. Koster F, Levin J, Walker L, et al. Hemolytic-uremic syndrome after shigellosis: relation to endotoxemia and circulating immune complexes. N Engl J Med 1978; 298:927-33.
20. Avram MM, Feinfeld DA, Huatuco AH. Search for the uremic toxin: decreased motor-nerve conduction velocity and elevated parathyroid hormone in uremia. N Engl J Med 1978; 298:1000-3.
21. Cosio M, Ghezzo H, Hogg JC, et al. The relations between structural changes in small airways and pulmonary-function tests. N Engl J Med 1978; 298:1277-81.
22. Jensen DM, McFarlane IG, Portmann BS, Eddleston ALWF, Williams R. Detection of antibodies directed against a liver-specific membrane lipoprotein in patients with acute and chronic active hepatitis. N Engl J Med 1978:299:1-7.
23. Aronow WS. Effect of passive smoking on angina pectoris. N Engl J Med 1978; 299:21-4.
24. Canadian Cooperative Study Group. A randomized trial of aspirin and sulfinpyrazone in threatened stroke. N Engl J Med 1978; 299:53-9.
25. Foster RS Jr, Lang SP, Constanza MC, Worden JK, Carleton RH, Yates JW. Breast self-examination practices and breast-cancer stage. N Engl J Med 1978; 299:265-70.
26. Trentham DE, Dynesius RA, Rocklin RE, David JR. Cellular sensitivity to collagen in rheumatoid arthritis. N Engl J Med 1978; 299:327-32.
27. Opelz G, Terasaki PI. Absence of immunization effect in human-kidney retransplantation. N Engl J Med 1978; 299:369-74.
28. Raskin P, Unger RH. Hyperglucagonemia and its suppression: importance in the metabolic control of diabetes. N Engl J Med 1978; 299:433-6.
29. Bilezikian JP, Canfield RE, Jacobs TP, et al. Response of 1α,25-dihydroxyvitamin D_3 to hypocalcemia in human subjects. N Engl J Med 1978; 299:437-41.
30. Siber GR, Weitzman SA, Aisenberg AC, Weinstein HJ, Schiffman G. Impaired antibody response to pneumococcal vaccine after treatment for Hodgkin's disease. N Engl J Med 1978; 299:442-8.
31. Loes MW, Singh S, Lock JE, Mirkin BL. Relation between plasma and red-cell electrolyte concentrations and digoxin levels in children. N Engl J Med 1978; 299:501-4.
32. Schussler GC, Schaffner F, Korn F. Increased serum thyroid hormone binding and decreased free hormone in chronic active liver disease. N Engl J Med 1978; 299:510-5.
33. Hoffman PM, Robbins DS, Nolte MT, Gibbs CJ Jr, Gajdusek DC. Cellular immunity in Guamanians with amyotrophic lateral sclerosis and Parkinsonism-dementia. N Engl J Med 1978; 299:680-5.
34. Opelz G, Terasaki PI. Improvement of kidney-graft survival with increased numbers of blood transfusions. N Engl J Med 1978; 299:799-803.

35. Soman V, Tamborlane W, DeFronzo R, Genel M, Felig P. Insulin binding and insulin sensitivity in isolated growth hormone deficiency. N Engl J Med 1978; 299:1025-30.
36. Felsher BF, Norris ME, Shih JC. Red-cell uroporphyrinogen decarboxylase activity in porphyria cutanea tarda and other forms of porphyria. N Engl J Med 1978; 299:1095-8.
37. Adams DB, Gold AR, Burt AD. Rise in female-initiated sexual activity at ovulation and its suppression by oral contraceptives. N Engl J Med 1978; 299:1145-50.
38. Rapoport J, Aviram M, Chaimovitz C, Brook JG. Defective high-density lipoprotein composition in patients on chronic hemodialysis: a possible mechanism for accelerated atherosclerosis. N Engl J Med 1978; 299:1326-9.
39. Wyatt R, Waschek J, Weinberger M, Sherman B. Effects of inhaled beclomethasone dipropionate and alternate-day prednisone on pituitary-adrenal function in children with chronic asthma. N Engl J Med 1978; 299:1387-92.
40. Urban MD, Lee PA, Migeon CJ. Adult height and fertility in men with congenital virilizing adrenal hyperplasia. N Engl J Med 1978; 299:1392-6.
41. Cox DW, Breckenridge WC, Little JA. Inheritance of apolipoprotein C-II deficiency with hypertriglyceridemia and pancreatitis. N Engl J Med 1978; 299:1421-4.
42. Dover GJ, Boyer SH, Charache S, Heintzelman K. Individual variation in the production and survival of F cells in sickle-cell disease. N Engl J Med 1978; 299:1428-35.
43. Piafsky KM, Borgå O, Odar-Cederlöf I, Johansson C, Sjöqvist R. Increased plasma protein binding of propranolol and chlorpromazine mediated by disease-induced elevations of plasma α_1 acid glycoprotein. N Engl J Med 1978; 299:1435-9.
44. Eaton LW, Weiss JL, Bulkley BH, Garrison JB, Weisfeldt ML. Regional cardiac dilatation after acute myocardial infarction: recognition by two-dimensional echocardiography. N Engl J Med 1979; 300:57-62.
45. Tamborlane WV, Sherwin RS, Genel M, Felig P. Reduction to normal of plasma glucose in juvenile diabetes by subcutaneous administration of insulin with a portable infusion pump. N Engl J Med 1979; 300:573-8.
46. Henderson WR, Shelhamer JH, Reingold DB, Smith LJ, Evans R III, Kaliner M. Alpha-adrenergic hyper-responsiveness in asthma: analysis of vascular and pupillary responses. N Engl J Med 1979; 300:642-7.
47. Thadani U, Davidson C, Singleton W, Taylor SH. Comparison of the immediate effects of five β-adrenoreceptor-blocking drugs with different ancillary properties in angina pectoris. N Engl J Med 1979; 300:750-5.
48. Toskes PP, Dawson W, Curington C, Levy NS, Fitzgerald C. Non-diabetic retinal abnormalities in chronic pancreatitis. N Engl J Med 1979; 300:942-6.
49. Schroeder SA, Showstack JA, Roberts HE. Frequency and clinical description of high-cost patients in 17 acute-care hospitals. N Engl J Med 1979; 300:1306-9.
50. Griffiths WJ, Neumann DA, Welsh JD. The visible vessel as an indicator of uncontrolled or recurrent gastrointestinal hemorrhage. N Engl J Med 1979; 300:1411-3.
51. Veterans Administration Cooperative Study Group on Antihypertensive Agents. Comparative effects of ticrynafen and hydrochlorothiazide in the treatment of hypertension. N Engl J Med 1979; 301:293-7.
52. Bloomfield CD, Gajl-Peczalska KJ, Frizzera G, Kersey JH, Goldman AI. Clinical utility of lymphocyte surface markers combined with the Lukes–Collins histologic classification in adult lymphoma. N Engl J Med 1979; 301:512-8.
53. Okada RD, Pohost GM, Kirshenbaum HD, et al. Radionuclide-determined change in pulmonary blood volume with exercise: improved sensitivity of multigated blood-pool scanning in detecting coronary-artery disease. N Engl J Med 1979; 301:569-76.
54. Gadek JE, Kelman JA, Fells G, et al. Collagenase in the lower respiratory tract of patients with idiopathic pulmonary fibrosis. N Engl J Med 1979; 301:737-42.
55. Pantely GA, Goodnight SH Jr, Rahimtoola SH, et al. Failure of antiplatelet and anticoagulant therapy to improve patency of grafts after coronary-artery bypass: a controlled randomized study. N Engl J Med 1979; 301:962-6.
56. Lang DA, Matthews DR, Peto J, Turner RC. Cyclic oscillations of basal plasma glucose and insulin concentrations in human beings. N Engl J Med 1979; 301:1023-7.
57. Leon MB, Borer JS, Bacharach SL, et al. Detection of early cardiac dysfunction in patients with severe beta-thalassemia and chronic iron overload. N Engl J Med 1979; 301:1143-8.
58. Packer M, Meller J, Medina N, Gorlin R, Herman MV. Rebound hemodynamic events after the abrupt withdrawal of nitroprusside in patients with severe chronic heart failure. N Engl J Med 1979; 301:1193-7.
59. McLellan AT, Woody GE, O'Brien CP. Development of psychiatric illness in drug abusers: possible role of drug preference. N Engl J Med 1979; 301:1310-4.
60. Godfrey K. Comparing the means of several groups. In: Bailar J, Mosteller F. Medical uses of statistics. Waltham, Mass.: New England Journal of Medicine (in press).
61. SAS users' guide: statistics — 1982 edition. Cary, N.C.: SAS Institute, 1982.
62. Nie NH, Hull CH, Jenkins JG, Steinbrunner K, Bent DH. SPSS: statistical package for the social sciences. 2nd ed. New York: McGraw-Hill, 1975.
63. Hull CH, Nie NH. SPSS update: new procedures and facilities for releases 7-9. New York: McGraw-Hill, 1981.
64. Miller R Jr. Simultaneous statistical inference. New York: Springer-Verlag, 1981.
65. Games PA, Keselman HJ, Rogan JC. A review of simultaneous pairwise multiple comparisons. Stat Neerland 1983; 37:53-8.
66. Dixon WJ, Brown MB, Engelman L, et al. BMDP statistical software 1981. Berkeley, Calif.: University of California Press, 1981.