16. Bonde-Petersen F, Norsk P, Suzuki Y. A comparison between freon and acetylene rebreathing for measuring cardiac output. Aviat Space Environ Med 1980; 51:1214-21.
17. Rowell LB, Blackmon JR, Bruce RA. Indocyanine green clearance and estimated hepatic blood flow during mild to maximal exercise in upright man. J Clin Invest 1964; 43:1677-90.
18. Christensen NJ, Vestergaard P, Sørensen T, Rafaelsen OJ. Cerebrospinal fluid adrenaline and noradrenaline in depressed patients. Acta Psychiatr Scand 1980; 61:178-82.
19. Heding LG, Kasperska-Czyzykowa T. C-peptide and proinsulin after oral glucose. Acta Med Scand [Suppl] 1980; 639:33-6.
20. Nakagawa S, Nakayama H, Sasaki T, et al. A simple method for the determination of serum free insulin levels in insulin-treated patients. Diabetes 1973; 22:590-600.
21. Hilsted J, Madsbad S, Krarup T, et al. No response of pancreatic hormones to hypoglycemia in diabetic autonomic neuropathy. J Clin Endocrinol Metab 1982; 54:815-9.
22. Tronier B. Radioimmunological determination of pancreatic polypeptide. Acta Endocrinol [Suppl] (Copenh) 1979; 227:72. abstract.
23. de Bodo RC, Steele R, Altszuler N, Dunn A, Bishop JS. On the hormonal regulation of carbohydrate metabolism: studies with C^{14} glucose. Recent Prog Horm Res 1963; 19:445-88.
24. Hetenyi G Jr, Norwich KH. Validity of the rates of production and utilization of metabolites as determined by tracer methods in intact animals. Fed Proc 1974; 33:1841-8.
25. Radziuk J, Norwich KH, Vranic M. Experimental validation of measurements of glucose turnover in nonsteady state. Am J Physiol 1978; 234:E84-E93.
26. Sokal RR, Rohlf FJ. Biometry: the principles and practice of statistics in biological research. 2nd ed. New York: W.H. Freeman, 1981.
27. Best JD, Halter JB. Release and clearance rates of epinephrine in man: importance of arterial measurements. J Clin Endocrinol Metab 1982; 55:263-8.
28. Shamoon H, Hendler R, Sherwin RS. Altered responsiveness to cortisol, epinephrine, and glucagon in insulin-infused juvenile-onset diabetics: a mechanism for diabetic instability. Diabetes 1980; 29:284-91.
29. Berk MA, Clutter WE, Skor D, et al. Enhanced glycemic responsiveness to epinephrine in insulin-dependent diabetes mellitus is the result of the inability to secrete insulin: augmented insulin secretion normally limits the glycemic, but not the lipolytic or ketogenic, response to epinephrine in humans. J Clin Invest 1985; 75:1842-51.
30. Rizza RA, Cryer PE, Haymond MW, Gerich JE. Adrenergic mechanisms of catecholamine action on glucose homeostasis in man. Metabolism 1980; 29:Suppl 1:1155-63.
31. Clutter WE, Bier DM, Shah SD, Cryer PE. Epinephrine plasma metabolic clearance rates and physiologic thresholds for metabolic and hemodynamic actions in man. J Clin Invest 1980; 66:94-101.
32. Engelman K, Mueller PS, Horwitz D, Sjoerdsma A. Denervation hypersensitivity of adipose tissue in idiopathic orthostatic hypotension. Lancet 1964; 2:927-30.

SPECIAL ARTICLE

STATISTICAL PROBLEMS IN THE REPORTING OF CLINICAL TRIALS

A Survey of Three Medical Journals

STUART J. POCOCK, PH.D., MICHAEL D. HUGHES, M.SC., AND ROBERT J. LEE, M.SC.

Abstract Reports of clinical trials often contain a wealth of data comparing treatments. This can lead to problems in interpretation, particularly when significance testing is used extensively. We examined 45 reports of comparative trials published in the *British Medical Journal*, the *Lancet*, or the *New England Journal of Medicine* to illustrate these statistical problems.

The issues we considered included the analysis of multiple end points, the analysis of repeated measurements over time, subgroup analyses, trials of multiple treatments, and the overall number of significance tests in a trial report. Interpretation of large amounts of data is complicated by the common failure to specify in advance the intended size of a trial or statistical stopping rules for interim analyses. In addition, summaries or abstracts of trials tend to emphasize the more statistically significant end points.

Overall, the reporting of clinical trials appears to be biased toward an exaggeration of treatment differences. Trials should have a clearer predefined policy for data analysis and reporting. In particular, a limited number of primary treatment comparisons should be specified in advance. The overuse of arbitrary significance levels (for example, $P<0.05$) is detrimental to good scientific reporting, and more emphasis should be given to the magnitude of treatment differences and to estimation methods such as confidence intervals. (N Engl J Med 1987; 317:426-32.)

OVER the years there has been a steady improvement in the clinical testing of new treatments. The randomized, controlled clinical trial has been increasingly accepted, along with higher standards in various aspects of trial design. We now have more precise definitions of patients' eligibility, treatment schedules, and outcome criteria; appropriate blinding and objectivity in assessments of patients; and better data collection and processing. In addition, methods of statistical analysis such as significance testing have become essential features in the reporting of trial findings, helping to ensure that any conclusions about the superiority of a treatment are based on evidence rather than opinion.

However, increased sophistication in the conduct of clinical trials, especially in regard to the use of computers for data processing and analysis, has produced more available data and more complex statistical analysis. For instance, trials often have several measures of patient outcome, some of which are assessed repeatedly during each patient's course of treatment; subgroups of patients may be analyzed for more specific differences between treatments; and more than two treatments may be compared. Because of this large amount of data, there is a danger that significance testing may be used excessively. This may lead to a loss of credibility of statistical methods since, inevitably, the risk of reporting some false positive

From the Department of Clinical Epidemiology and General Practice, Royal Free Hospital School of Medicine, Rowland Hill St., London NW3, England, where reprint requests should be addressed to Dr. Pocock.

findings increases. Also, if all findings are not reported, either in the article or its summary or abstract, treatment differences can be exaggerated or underplayed. Thus, there is a need to formulate in advance a strategy for handling complex trial data.

Statistical planning should also enter into other aspects of trial design, particularly the determination of the trial size required to detect treatment differences of interest. Any policy for undertaking interim analyses of the accumulating data should incorporate appropriate decision rules for stopping the trial and reporting its results early if important treatment differences are observed.

The tendency for significance testing to dominate reports of clinical trials is unfortunate, since indications for subsequent clinical practice depend more on the observed magnitude of any treatment difference, together with its statistical confidence intervals.

This paper reviews current practice regarding statistical aspects of clinical trial reports by examining a representative sample of recent reports in three medical journals. Methods for assessing the general quality of a trial's conduct and reporting have previously been discussed,[1,2] and several evaluations of trials reported in the medical literature have been undertaken.[3-6] Our aim here is to show how various problems in the design, conduct, analysis, and reporting of clinical trials lead to potential biases toward presenting findings that can exaggerate the overall perception of progress in clinical research. We also discuss guidelines on how to avoid the more common pitfalls.

Methods

We examined 45 reports of controlled clinical trials published in the *British Medical Journal*, the *Lancet*, or the *New England Journal of Medicine*. Fifteen consecutive reports published after July 1, 1985, were included from each journal.

The reports were published within three months in the *Lancet*, four months in the *British Medical Journal*, and six months in the *New England Journal of Medicine*. The survey included only comparative trials, both randomized (n = 38) and nonrandomized (n = 7). The median number of patients in the trials was 102; nine trials had fewer than 25 patients and two had more than 5000. The trials involved many diseases (including eight infectious, seven cardiovascular, four respiratory, and three digestive disorders) and were considered to be representative of clinical trials reported in major medical journals.

For each trial, one of us (R.J.L.) completed a standardized evaluation of all relevant aspects of trial design, analysis, and reporting. As a reliability check, a random group of 10 trials was reevaluated (by M.D.H.); no substantial differences were found.

Results

Multiple End Points

Trials often evaluate several different aspects of patients' responses. Such multiple end points allow a fuller comparison of the merits of different treatments, but the consequent increase in statistical analyses (for example, multiple significance tests) can lead to problems in interpretation.

In our survey, an end point was defined broadly as any measure of a patient's state assessed after assignment to treatment and intended to be obtained for all patients. However, we excluded data on adverse events, since they are often not formally analyzed and are kept separate from evaluations of efficacy. We found that 37 trials included qualitative end points (i.e., a response criterion in two or more categories), 32 trials included quantitative end points, and 10 included survival data (i.e., time to death or other disease-related event). In recent years biostatisticians have made great progress in developing methods for analyzing survival data, but in practice most trials are confined to other types of data, for which there are more established analytic techniques.

Table 1 shows the number of end points mentioned in each trial report, as well as the number for which a significant test comparing treatments was performed. The median number of end points mentioned was six. Five reports discussed only one end point but three mentioned more than 15. A difference between journals was apparent: the median numbers of end points were three, six, and nine, respectively, in the *Lancet*, the *British Medical Journal*, and the *New England Journal of Medicine*. The differences may have been due to the relative length of reports in the journals.

Significance tests were performed for most end points. The median number of such tests per trial was four. Three trials used no significance tests for comparing treatments, although in one, the response within each treatment group was tested before and after treatment.

For trials with only one end point, the standard interpretation of a significance test is straightforward: the significance level truly represents the probability of rejecting the null hypothesis of no treatment difference when it is in fact true (the Type I error). Eight trials reported only one end point (one hopes that this was not a post hoc selection from a wider choice of possible end points). However, most trials reported several end points, thereby increasing the risk of a Type I error. For instance, for a trial with five end points, the chance under the null hypothesis of at least one treatment difference achieving a significance level of $P<0.05$ is about 20 percent, provided that the end points are not highly correlated. Since many people's interpretations of P values do not take into account

Table 1. Distribution of the Number of End Points in 45 Reports of Clinical Trials.

No. of End Points Mentioned	No. of Trials	No. of End Points Tested for Significance	No. of Trials
		0	3
1	5	1	8
2	5	2	5
3	5	3	7
4–5	6	4–5	5
6–9	15	6–9	12
10–14	6	10–14	4
≥15	3	≥15	1
Total	45		45

their repeated use in this way, the increased danger of a Type I error goes largely unnoticed.

One way to overcome this problem is to identify a single primary end point when designing a trial and to draw valid conclusions from its statistical analysis. Other secondary end points are then interpreted in a more exploratory manner. Among the trials in this survey, 12 (27 percent) did this; however, 5 of them mentioned only one end point anyway.

For instance, one study of extracranial–intracranial bypass[7] identified a primary end point — the occurrence of stroke and stroke-related death in a patient — but also listed three secondary end points. The interpretation of the results is clear and valid. In contrast, a trial studying the suppression of secondary hyperparathyroidism in children[8] included 12 end points, 11 of which were subjected to significance tests. No priorities were specified, and the trial included only 12 children. The multiple P values and low statistical power cast doubt on the results.

On some occasions a single end point cannot be identified in advance. Stricter nominal significance levels might then be employed by, for instance, using Bonferroni correction methods.[9,10] This emphasizes the importance of the size of P values. For example, a trial with 10 or more significance tests and one P value less than 0.001 still yields good evidence of a true treatment effect. Alternatively, methods of simultaneous assessment of closely related end points[11,12] might be employed.

Subgroup Analyses

In addition to performing an overall comparison of treatment groups, it is often relevant to inquire whether treatment differences are more (or less) pronounced in any particular subgroup of patients classified according to prognostic factors. Twenty-three trials in our survey (51 percent) had at least one subgroup analysis that compared the response to treatment in different categories of patients, and 10 trials included more than one prognostic factor in their subgroup analyses.

A study of ribavirin in the treatment of lassa fever[13] illustrates the problems in interpreting subgroup analyses. Attention is focused on treatment differences based on three prognostic factors (levels of serum aspartate aminotransferase, the extent of viremia, and the time since the onset of fever), with the results presented as P values within the subgroups. Many findings were statistically significant, but no clear idea was given of the real effects of the prognostic factors on the value of treatment.

The problems here are (1) the overall treatment comparison is not sufficiently prominent, so that the reader gets lost in a morass of subgroup analyses; (2) subgroups were not selected a priori, so that post hoc findings may present a selective and distorted picture of the role of prognostic factors; and (3) P values for subgroups are inappropriate for assessing whether the treatment difference varied between subgroups. Instead, statistical tests of interaction[14] assess directly whether a prognostic factor affects the difference in treatments.

None of the trial reports that we surveyed specified that subgroups had been defined in advance, although a few stated that such analyses were exploratory. Of the 23 trials with subgroup analyses, 16 focused on subgroup P values, 4 used descriptive statistics only (perhaps an appropriate way of playing down subgroup analyses), and only 3 used statistical tests of interaction. One of these trials, the Medical Research Council's study of treatments for mild hypertension,[15] showed a significant sex–treatment interaction ($P = 0.05$) for all-cause mortality, when the difference in the numbers of deaths among men receiving active treatment and men receiving placebo (157 and 181, respectively) was in the opposite direction from that among women (91 and 72, respectively). This interaction test needs cautious interpretation, since many such tests could have been performed with the extensive data from this trial.

Authors can be too eager to explore subgroup analysis in trials containing too few patients for such an evaluation. A more important use of data on prognostic factors is to check that treatment groups are comparable.[16] Indeed, 38 trials (84 percent) did report such a base-line comparison of treatment groups. If any noncomparability exists, one can use regression methods to adjust for imbalances in prognostic factors when making an overall comparison of treatments.[17] In particular, logistic regression and proportional-hazard models are useful for binary and survival end points, respectively. The importance of a prognostic factor in such regression models may be a useful prerequisite to assessing its possible interaction with treatment. Only nine trials reported such an adjustment for base-line differences. It should be noted that base-line differences should not be detected by significance testing, since the effect of a prognostic factor on the overall difference in treatment results depends both on its effect on response and the magnitude of the imbalance between groups.[16,17]

Repeated Measurements over Time

In trials with quantitative measurements of response, such measurements are often made before treatment begins and several times during treatment. Eighteen trials in our survey (40 percent) had such repeated measurements over time. Ten of the trials used descriptive statistics only, but eight reported the results of significance tests at several time points. This repeated testing seriously increases the risk of a Type I error, since authors focus attention on time points with the most significant differences.

For example, one trial[18] compared two active treatments for onchocerciasis and a placebo with use of a clinical-reaction score over six months. Pairwise significance tests were performed daily for 8 days and then at 10 days, four weeks, three months, and six months. This produced 36 P values: the tests were not

independent, the risk of a Type I error was large, and interpretation is difficult. Two main features appeared to be of interest — the patient's maximal score and the interval during which the score remained above the value obtained with placebo. Appropriate analyses could have been defined a priori — perhaps a test comparing maximal scores and a survival analysis of the times needed to fall below a given score. Other aspects of the data could simply have been displayed graphically.

In general, trials with repeated measurements need an overall prespecified strategy for statistical analysis.[19] Unfortunately, this did not appear to have been present in the trials we surveyed. Several possible strategies might be considered, depending on the clinical objectives. First, for each patient, the mean value for observations over a specified time could be taken as the summary measure of response. One is then comparing the average effect of treatments over time. Second, one or two time points for a formal treatment comparison could be specified in advance. Third, time periods to attain a specified (threshold) value could be compared. More complex techniques for analyzing repeated measurements can be used,[20] but they are difficult to communicate to nonstatisticians. A graphic display of the time trends is a valuable supplement to any analysis.

Number of Treatment Groups

Twenty-six trials in our survey had two treatment groups, and seven others used a two-period crossover design. However, nine trials involved three types of treatments, and three trials involved four. The analysis of such multitreatment trials needs careful consideration. Applying significance tests to all pairwise differences, as in the trial described above, will increase the risk of a Type I error and can make interpretation difficult.

There are methods for comparing more than two treatment groups — e.g., analysis of variance and studentized range methods for quantitative data.[21] However, many multitreatment trials compare two active treatments with either placebo or their combination (or both). A priori hypotheses could then be formulated to compare two groups defined by appropriate amalgamation. For instance, the Medical Research Council's trial of three treatments for mild hypertension (bendrofluazide, propranolol, and placebo)[15] had a primary hypothesis comparing active treatment with placebo. A subsidiary analysis compared the two active drugs. The specification of such priorities in advance helps to secure a valid interpretation of multitreatment trials, from which the derivation of recommendations may otherwise be difficult and controversial.

The Extent of Significance Testing

Significance testing is a valuable and established tool for interpreting medical data. It is most clearly interpretable when there is a single prespecified hypothesis. However, most clinical trial reports make more extensive use of significance tests. Table 2 shows the number of significance tests for treatment comparisons reported in the 45 trials we examined. It includes all significance tests, whether performed on multiple end points, in subgroup analyses, or on repeated measurements over time. The median number of tests per trial was 8; six trials reported more than 20 tests.

Table 2. Extent of Significance Tests Comparing Treatments in 45 Reports of Clinical Trials.

No. of Tests	No. of Trials
0	3
1–5	11
6–10	13
11–20	12
21–50	5
>50	1
Total	45

This may underrepresent the actual use of significance testing, since authors may select which tests to mention. For instance, some nonsignificant comparisons may not be explicitly reported as having been tested or may be excluded from the report altogether.

The excessive and unstructured use of significance testing in medical articles casts doubt on its credibility, especially when people erroneously interpret $P<0.05$ as "proof" of a treatment difference. The quality of statistical reporting would be improved if $P<0.05$ had no special relevance and authors presented actual P values, so that $P = 0.04$ and $P = 0.06$ were seen as similar findings of moderate evidence against the null hypothesis.

Trial reports ought to focus on a small number of hypotheses that are specified in advance, so that the principal significance tests can be interpreted without concern about post hoc selection. Subsidiary analyses (of secondary end points and patient subgroups) may still use significance tests but in the spirit of cautious and exploratory data analysis. This is not to deny the value of exploratory analyses, since any unexpected findings should be reported so that they can be confirmed or refuted in subsequent trials. Thus, exploratory analyses can be useful for formulating new hypotheses, but not for testing them.

Confidence Intervals

Because of the obsession with significance testing in the medical literature, authors often give insufficient attention to estimating the magnitude of treatment differences. Confidence intervals express the uncertainty inherent in any trial by presenting upper and lower bounds for the anticipated true treatment difference. Several authors have described the use of confidence intervals for various types of data.[22,23] They are closely related to significance testing: a treatment difference that is significant at the 5 percent level has a 95

percent confidence interval for the difference that is wholly on one side of zero.

Unfortunately, only six trials in our survey (13 percent) made use of confidence intervals. The *British Medical Journal* now requires more extensive use of confidence intervals (or other estimation methods) — a policy that other journals might wish to follow. For example, the Medical Research Council's trial of treatments for mild hypertension[15] reported an observed 45 percent reduction in strokes among patients receiving active treatment as compared with those receiving placebo. Instead of simply quoting $P<0.001$, the report gave 95 percent confidence limits of 25 and 60 percent for this reduction, a helpful indication of the uncertainty inherent in a comparison of 60 and 109 strokes. A trial of oral magnesium treatment for hypertension[24] reported no statistically significant effect. However, the trial had only 17 patients, and its low power to detect realistic changes could have been demonstrated by a wide confidence interval.

Intended Size of Trial and Stopping Rules

The intended number of patients in a clinical trial should be determined in advance, and statistical power calculations are valuable.[14,25,26] Only five trial reports in our survey (11 percent) mentioned the intended number of patients; in each case, this was supported by a statement of statistical power. With most trial reports, the reader has no idea whether the investigators (1) had no preset trial size and reported the results at an arbitrary time, with the magnitude (or significance) of the treatment difference possibly affecting the decision to report; (2) failed to achieve the intended trial size and decided to report the trial anyway; (3) extended the trial beyond its intended size in order to achieve better statistical power; or (4) reported the trial before the intended trial size was achieved, because interim results showed a substantial treatment difference.

The reader's interpretation of trial findings depends on which of the four circumstances occurred. For trials with a "negative" conclusion — that is, with no evidence of a treatment difference — possibilities 1 and 2 may reflect a lack of statistical power and premature publication. Confidence intervals would then be valuable in conveying whether clinically important differences may exist. For trials with a "positive" conclusion — that is, in which there were some statistically significant treatment effects — possibilities 1, 3, and 4 should instill some caution in the reader. For instance, the authors may have taken repeated looks at the accumulating data and chosen to report the analysis that best highlighted the treatment difference they were hoping to see. In such cases, neither confidence limits nor P values protect sufficiently against the biased timing of publication. These statistical methods allow only for the effects of random variability and cannot correct for injudicious timing or other biases in reporting.

It is ethically desirable to perform interim analyses so that clear evidence of a treatment difference can lead to stopping the trial early. The danger is that repeated use of significance testing increases the risk of a Type I error. For instance, with 10 interim analyses of one end point, there is a 20 percent chance of reaching $P<0.05$ even if the null hypothesis is true. This problem becomes more serious if there are several end points or more than two treatments. Several statistical stopping rules exist,[27,28] their essential feature being that allowance for repeated significance testing requires greater treatment differences in order for a trial to be stopped early.

Only five trials in our survey (11 percent) mentioned any such stopping rules, and no trial was actually stopped early. Most trial reports mentioned neither an intended trial size nor any policy on stopping and publication. This leaves enormous scope for bias in reporting, which in turn affects the credibility of results, since reported P values make no allowance for the selective timing of publication.

Selection of Results for the Summary

Since the summary or abstract of a clinical trial report receives the greatest attention, it is important that it provide a fair reflection of the trial's findings. We were concerned about the selection of results for inclusion in the summary. For simplicity, we restricted our evaluation of abstracts or summaries to the 33 reports on trials of two treatments. Each end-point treatment difference in the main text of the article was classified as either statistically significant at the 5 percent level, or not significant at the 5 percent level or not tested. We then noted which end points were mentioned in the summary. Overall, 25 percent of the 130 nonsignificant comparisons were included in the summary, whereas 70 percent of the 91 significant comparisons were included.

Table 3 shows that in seven trials, all reported end points were statistically significant, whereas in another seven, no end points were statistically significant. The other 19 trial reports contained a mixture of significant and nonsignificant end points. For these, a stratified Mantel–Haenszel procedure estimates a within-trial relative odds ratio of 9.2:1 for the inclusion of significant:nonsignificant end points in the summary.

This tendency to favor statistically significant results when writing a trial summary means that such summaries may exaggerate the true extent of treatment differences. It is also noteworthy that 15 summaries (33 percent) provided only the statistical significance, thus failing to describe the magnitude of the treatment differences.

Recommendations

This survey highlights some important statistical issues inherent in the reporting of clinical trials. A critical attitude has been adopted in order to em-

phasize the severe bias toward exaggeration of treatment differences that can arise as a result of current practice. We now offer a set of recommendations that could go some way toward avoiding distortions in reports of clinical trials. Of course, deficiencies in reporting often arise because of earlier failings in planning, conduct, and statistical analysis. Hence, our recommendations bear on the whole process of a clinical trial, from the development of a study protocol to the publication of the report.

(1) Although it is valuable for trials to evaluate several aspects of patients' responses, it is important to identify a small set of primary end points in advance. In many trials the design should specify a single primary end point. The results for primary end points (including any nonsignificant findings) should be fully reported in both the trial report and its summary or abstract.

(2) Results for secondary end points should be presented as exploratory findings. Any significance testing of these end points requires cautious interpretation, since there may be many of them and because, by definition, they are not considered a priori to be of principal importance.

(3) Subgroup analysis should be confined to a limited number of prespecified hypotheses concerning the interaction between treatment and a prognostic factor. Statistical tests for interaction should be used, rather than subgroup P values. Subgroup findings should be interpreted cautiously, in a spirit of exploratory data analysis. If a trial has limited statistical power (i.e., not enough patients), subgroup analyses should be avoided.

(4) Trials with repeated measurements of a quantitative end point over time require a prespecified policy for statistical analysis. This should be aimed toward a single specific hypothesis of interest, and repeated significance tests at each time point should be avoided.

(5) For trials with more than two treatments, the primary treatment contrasts should be specified beforehand and emphasized in the report.

(6) Authors should use as few significance tests as possible, so that the risk of a Type I error is limited. Exact P values should be presented, rather than references to arbitrary levels (e.g., $P<0.05$). The magnitude of treatment differences for primary end points should be stated, along with the confidence limits.

(7) The intended size of a trial and the mathematical justification of the intended size (e.g., power calculations) should be specified in the Methods section. Any discrepancy between the actual and intended number of patients should be explained.

(8) If interim analyses of the accumulating data will be undertaken, a policy for the frequency and content of such analyses should be defined in advance. In particular, there should be stopping rules for terminating and reporting a trial that are based on established statistical principles, and such rules should be stated in the trial report.

(9) The summary or abstract of a trial report should reflect the overall findings fairly. Authors should not emphasize the more statistically significant findings in the summary. The summary should mention the magnitude of treatment differences rather than their statistical significance.

(10) All the above recommendations imply that investigators must define a coordinated policy for the statistical aspects of a clinical trial, which reflects a consistency of intent from the design of a trial through its conduct, analysis, interpretation, and reporting. The problems described are widely recognized by professional biostatisticians and could be alleviated by closer collaboration between them and clinicians during all phases of trial development. Although clinical trials can provide fascinating data on many important aspects of a disease and its treatment, the funda-

Table 3. Significant and Nonsignificant End-Point Treatment Differences in Reports of 33 Trials of Two Treatments.*

Trial No.†	No. of Significant End Points		No. of Nonsignificant End Points	
	In paper	Also in summary	In paper	Also in summary
All end points significant				
B5	3	3	0	0
L2	1	1	0	0
L6	1	1	0	0
L8	1	1	0	0
N3	3	2	0	0
N5	6	6	0	0
N7	8	4	0	0
All end points nonsignificant				
B1	0	0	5	1
B12	0	0	3	2
B13	0	0	6	2
L3	0	0	2	1
L12	0	0	2	1
L13	0	0	1	1
N8	0	0	9	5
Both significant and nonsignificant end points				
B2	4	2	2	2
B4	2	2	12	2
B6	3	2	7	0
B9	1	1	10	2
B10	4	1	6	1
B11	7	5	11	1
B14	4	3	1	0
L1	2	2	7	1
L5	1	1	2	0
L7	1	0	7	2
L10	2	2	2	2
L11	3	2	2	0
L14	7	7	4	1
N4	3	2	8	1
N6	3	2	6	1
N12	1	0	8	3
N13	5	2	4	0
N14	2	2	1	0
N15	13	8	2	0
Total	91	64 (70%)	130	32 (25%)

*"Significant" means that the main overall treatment comparison for the end point was statistically significant at the 5 percent level.

†B denotes the *British Medical Journal*, L the *Lancet*, and N the *New England Journal of Medicine*.

mental hypotheses affecting the practitioner's future choice of treatment must not be compromised by a plethora of subsidiary information.

References

1. Chalmers TC, Smith H Jr, Blackburn B, et al. A method for assessing the quality of a randomized control trial. Controlled Clin Trials 1981; 2:31-49.
2. Antczak AA, Tang J, Chalmers TC. Quality assessment of randomized control trials in dental research. I. Methods. Periodontal Res 1986; 21:305-14.
3. DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on methods in clinical trials. N Engl J Med 1982; 306:1332-7.
4. Moskowitz G, Chalmers TC, Sachs HS, Fagerstrom RM, Smith H Jr. Deficiencies of clinical trials of alcohol withdrawal. Alcoholism (NY) 1983; 7:42-6.
5. Liberati A, Himel HN, Chalmers TC. A quality assessment of randomized control trials of primary treatment of breast cancer. J Clin Oncol 1986; 4:942-51.
6. Antczak AA, Tang J, Chalmers TC. Quality assessment of randomized control trials in dental research. II. Results: periodontal research. Periodontal Res 1986; 21:315-21.
7. EC/IC Bypass Study Group. Failure of extracranial-intracranial arterial bypass to reduce the risk of ischemic stroke: results of an international randomized trial. N Engl J Med 1985; 313:1191-200.
8. Mak RHK, Turner C, Thompson T, Powell H, Haycock GB, Chantler C. Suppression of secondary hyperparathyroidism in children with chronic renal failure by high dose phosphate binders: calcium carbonate versus aluminium hydroxide. Br Med J 1985; 291:623-7.
9. Miller RG Jr. Simultaneous statistical inference. New York: Springer-Verlag, 1981.
10. Armitage P, Parmar M. Some approaches to the problem of multiplicity in clinical trials. In: Proceedings of the XIII International Biometric Conference. Seattle: Biometric Society, 1986.
11. O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics 1984; 40:1079-87.
12. Pocock SJ, Geller NL. The analysis of multiple endpoints in clinical trials. Biometrics (in press).
13. McCormick JB, King IJ, Webb PA, et al. Lassa fever: effective therapy with ribavirin. N Engl J Med 1986; 314:20-6.
14. Pocock SJ. Clinical trials: a practical approach. Chichester, England: John Wiley, 1983.
15. Medical Research Council Working Party. MRC trial of treatment of mild hypertension: principal results. Br Med J 1985; 291:97-104.
16. Altman DG. A fair trial? Br Med J 1984; 289:336-7.
17. Lavori PW, Louis TA, Bailar JC III, Polansky M. Designs for experiments — parallel comparisons of treatment. N Engl J Med 1983; 309:1291-8.
18. Greene BM, Taylor HR, Cupp EW, et al. Comparison of ivermectin and diethylcarbamazine in the treatment of onchocerciasis. N Engl J Med 1985; 313:133-8.
19. de Klerk NH. Repeated warnings re repeated measures. Aust NZ J Med 1986; 16:637-8.
20. Healy MJR. Some problems of repeated measurements. In: Bithell JF, Coppi R, eds. Perspectives in medical statistics: proceedings of the European Symposium on Medical Statistics, Rome, 1980. London: Academic Press, 1981:155-71.
21. Armitage P. Statistical methods in medical research. Oxford: Blackwell, 1971:202-7.
22. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J 1986; 292:746-50.
23. Simon R. Confidence intervals for reporting results of clinical trials. Ann Intern Med 1986; 105:429-35.
24. Cappuccio FP, Markandu ND, Beynon GW, Shore AC, Sampson B, MacGregor GA. Lack of effect of oral magnesium on high blood pressure: a double blind study. Br Med J 1985; 291:235-8.
25. Altman DG. Statistics and ethics in medical research. III. How large a sample? Br Med J 1980; 281:1336-8.
26. Gore SM. Assessing clinical trials — trial size. Br Med J 1981; 282:1687-9.
27. Geller NL, Pocock SJ. Interim analyses in randomized clinical trials: ramifications and guidelines for practitioners. Biometrics 1987; 43:213-23.
28. Fleming TR, Harrington DP, O'Brien PC. Designs for group sequential tests. Controlled Clin Trials 1984; 5:348-61.

CASE RECORDS
OF THE
MASSACHUSETTS GENERAL HOSPITAL

Weekly Clinicopathological Exercises

CASE 33-1987

Presentation of Case

A 46-year-old man was admitted to the hospital because of pain in the right lower abdominal quadrant.

He was well until 9 or 10 days earlier, when fever and intermittent severe periumbilical pain developed, lasted for a day or two, and then became localized in the right lower abdominal quadrant. He came to the Emergency Ward of this hospital, where moderate anemia was found. An x-ray film of the chest was normal, and an x-ray film of the abdomen showed no abnormality. Oxycodone–acetaminophen was prescribed, without relief. During the week before entry the pain was mild and continuous and was increased when the patient palpated his abdomen. He observed bright-red blood in his stools on several occasions. Two days before admission the patient experienced chilliness. On the day of entry nausea developed, and he came to this hospital.

The patient was a police officer. There was a history of diverticulitis of the colon six years earlier, with similar pain in the right lower abdominal quadrant and hematochezia; the patient was admitted to another hospital for 10 days and received multiple antibiotics. The diagnosis was reportedly established by a barium-enema examination; the patient was considered in too much pain to endure colonoscopic examination. In the same year he was involved in a motor-vehicle accident that was followed by persistent low-back pain, for which he used ibuprofen. He was said to have hypertension but received no treatment for it. There was no history of abdominal trauma, surgical procedures, vomiting, constipation, diarrhea, hematemesis, hematuria, dysuria, hepati-