BRIEF COMMUNICATION

False-Positive Results in Clinical Trials: Multiple Significance Tests and the Problem of Unreported Comparisons

Ian F. Tannock*

At least three factors increase the probability that the results in a report of a clinical trial will be false positive:

1) Publication bias—the tendency for investigators to submit and editors to accept selectively reports of trials that appear to give positive results (1-4), some of which will occur by statistical chance.

2) The low probability that new treatments will lead to therapeutic advances (5), implying a low prevalence of true-positive trials (6). Just as a screening procedure (e.g., fecal occult blood test for colorectal cancer) will give a high rate of false-positive results when applied to a population with low prevalence of a disease, so a clinical trial will have a high false-positive rate (and a low false-negative rate) if there is a low prevalence of true differences in outcome. This effect is illustrated by the following example, provided by Dr. M. Parmar (personal communication). Suppose 200 trials are undertaken with a significance level of 5% (type I error of \( \alpha = 0.05 \)) and with a 0.9 power (1 - \( \beta \) [type II error = 0.1]). If the prevalence of trials with a real difference is 10% (20 trials), 18 (20 \( \times 0.9 \)) trials will be reported correctly as true positives; in contrast, there will be nine (180 \( \times 0.05 \)) false positives or one in three (i.e., nine of 27) trials reported as positive. Overall, 173 of the 200 trials will be reported as negative (171 true negatives and two false negatives) and 27 trials will be reported as positive (18 true positives and nine false positives).

3) The performance of multiple significance tests, only some of which may be reported in an article. One in 20 \( P \) values for comparison of equivalent outcomes will be less than .05 because of chance alone. Comparisons may be made for multiple end points, for subgroups, and serially during accrual and follow-up. Investigators are often encouraged by editors and reviewers to perform multiple analyses but to report only the most “interesting” results, i.e., those which appear to be “significant.”

To determine the extent of multiple significance testing, I reviewed the 32 reports of randomized clinical trials relating to oncology that appeared in The New England Journal of Medicine (\( n = 9 \)) and the Journal of Clinical Oncology (\( n = 23 \)) in 1992. Characteristics of these reports are summarized in Table 1. The median number of therapeutic end points per trial was 5 (range, 2-19), giving opportunity for considerable latitude in the present analysis in selecting the most “impressive” result in the 13 articles that did not define a primary end point. Most reports failed to indicate a planned sample size or stopping criteria, thus facilitating termination if any of several interim analyses suggested a positive result. Subgroup analyses were common and usually not indicated as planned, and they were sometimes used to draw conclusions rather than suggest hypotheses.

I estimated the number of statistical comparisons that were reported in each article and the number that were not reported but were very likely to have been undertaken (Table 2). A comparison was considered to be reported if a \( P \) value was stated, if there was a confidence interval for a difference or ratio, or if there was a statement of lack of statistical significance. Statistical comparisons were implied if some subgroups or categories in a group were compared but others (e.g., those listed in a table) were not. If a test was described in the methods with no result given, or if a two-way comparison among three groups was reported. If there was reasonable doubt in the current analysis, a conservative viewpoint was taken and a possible test was assumed not to have been undertaken; alternatives to an arbitrary cutoff in a continuous variable such as age were not assumed. Each article was reviewed twice, the second time several weeks after the first, and any discrepancies in estimates of the number of comparisons were resolved. Additional reports of 22 of these trials were identified from their bibliography, from meeting abstracts, and from a Medline search, and additional statistical significance tests from serial analyses were identified (Table 2).

The number of statistical comparisons undertaken during the analysis of these trials was large, and the number of unreported comparisons often exceeded the number in the article. Even though comparisons of end points of efficacy may not be independent (i.e., end points

---

*Correspondence to: Ian F. Tannock, M.D., Ph.D., Department of Medicine, Princess Margaret Hospital and University of Toronto, 610 University Ave., Toronto, ON M5G 2M9, Canada.

See "Notes" section following References.
Table 2. Number of statistical comparisons, reported or implied to have been undertaken, in 32 randomized clinical trials

<table>
<thead>
<tr>
<th>Original article</th>
<th>Total</th>
<th>Major outcome parameters</th>
<th>All comparisons of efficacy*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Reported</td>
<td>Reported plus implied</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>Range</td>
<td>Median</td>
</tr>
<tr>
<td>Original article</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>32</td>
<td>5-86</td>
<td>86</td>
</tr>
<tr>
<td>Major outcome parameters</td>
<td>6</td>
<td>1-31</td>
<td>13</td>
</tr>
<tr>
<td>All comparisons of efficacy*</td>
<td>9.5</td>
<td>1-49</td>
<td>20</td>
</tr>
<tr>
<td>Original article plus serial reports</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>8-112</td>
<td>95</td>
</tr>
<tr>
<td>All comparisons of efficacy</td>
<td>14.5</td>
<td>2-84</td>
<td>27</td>
</tr>
</tbody>
</table>

*Includes subgroup analyses.

such as response rate, duration of response, and survival may be correlated), the number of comparisons is sufficient to give a high probability that there will be spurious positive results because of chance alone in many of these trials.²

The current survey extends the results of Pocock et al. (7), who documented the problem of multiple significance testing in reports of clinical trials. There is disagreement about the usefulness of correcting for multiple tests (8,9); it is important, however, that investigators be aware of the problem of spurious positive results. Exploratory comparisons are important in establishing new hypotheses, but only comparisons of the major predefined end point(s) should be used to draw conclusions.

Multiple significance testing, publication bias, and the low expectation of therapeutic advances each contribute to the probability of falsely declaring trial results to be positive. More information about the analysis of clinical trials should be provided in their reports, including the number of statistical comparisons that were undertaken, and all protocols should define a primary end point. Important trials should be repeated, and low credence should be given to apparently positive trials with borderline statistical significance in their major end point(s).

References

8 Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials [see comment citation in Medline]. Biometrics 1987;43:487-98.

Notes

References are available from the author on request.

Twenty of the 32 studies reported statistically significant benefits of one treatment. For these 20 studies, a median total of 15 statistical tests (range, 4-48) were either reported or considered to have been undertaken. Of the 15 tests, a median of 4 (range, 2-15) showed apparently statistically significant differences in major outcome parameters between groups containing randomly assigned individuals.

Manuscript received June 28, 1995; revised November 7, 1995; accepted November 13, 1995.