clinical research programs within the medical center or to contribute to a national pool linked with support from industry to establish a national endowment for funding translational research and drug or device development within academic medical centers. Such promotion of later-phase research within academic medical centers could enhance the value of the intellectual property derived from it, financial benefits from which could, in turn, be used to establish research endowments within the medical centers.

The federal government might also consider alternative ways to fund the NIH budget that are independent of allocations from the tax base. One approach might include seeking support from industries whose products contribute to the burden of disease, providing tax credits as an incentive for their contribution. These resources could be used to establish an independently managed national fund, which could be used to ensure adequate support for biomedical research without the funding gaps or oscillations that currently plague the process. In this scenario, unused money from any fiscal year would be retained in the fund, with the goal of achieving self-sustained growth.

Whatever mechanisms are ultimately chosen, it seems clear that new methods of support must be developed if biomedical research is to continue to thrive in the United States. The goal of a durable, steady stream of support for research in the life sciences has never been more pressing, since the research derived from that support has never promised greater benefits. The fate of life-sciences research should not be consigned to the political winds of Washington.

Dr. Loscalzo is the physician-in-chief and chair of medicine at Brigham and Women's Hospital and a professor of medicine at Harvard Medical School — both in Boston.

1. Office of Budget. FY 2007 budget in brief: advancing the health, safety, and well-being of our people. Washington, D.C.: Department of Health and Human Services, 2006.
2. Press release of the Pharmaceutical Research and Manufacturers of America, February 13, 2006. (Accessed March 30, 2006, at http://www.phrma.org/news_room/press_releases/r%26d_investments_by_america%92s_pharmaceutical_research_companies_nears_record_%2440_billion_in_2005/.)
3. Korn D, Rich RR, Garrison HH, et al. The NIH budget in the "postdoubling" era. Science 2002;296:1401-2.

STATISTICS AND MEDICINE

The Challenge of Subgroup Analyses — Reporting without Distorting

Stephen W. Lagakos, Ph.D.

Related article, page 1706

Subgroup analyses are an important part of the analysis of a comparative clinical trial. However, they are commonly overinterpreted[1-4] and can lead to further research that is misguided or, worse, to suboptimal patient care.

Consider a randomized, clinical trial designed to determine whether a new treatment is more effective than an established treatment and assessed with a test, based on all randomized patients, of the null hypothesis that the treatments have equal efficacy, as measured in terms of the primary end point. Then, subgroup analyses are conducted to assess whether different types of patients respond differently to the new treatment. This sounds simple enough, but there are several important sources of confusion and uncertainty regarding such subgroup analyses.

A single subgroup analysis may be conducted in which patients are classified according to sex. If the overall trial results fail to demonstrate that the new treatment is better than the conventional treatment, it may still be better in certain patients (say, women). And if the new treatment is demonstrated to be superior, the magnitude of the benefit may vary according to sex. Both scenarios should be formally investigated by means of an "interaction test" of the null hypothesis that the relative efficacy of the two treatments is the same in women and in men. An interaction is called quantitative[1,4] when the new treatment is superior for both subgroups but its relative benefit differs between the subgroups. The clinical implications are usually more important for a qualitative[1,4] interaction, in which the new treatment is superior in one subgroup but no different from or inferior to conventional treatment in another subgroup.

An alternative, but problematic,[1,3,4] approach to investigating subgroups is to test the hypothesis that there is no treatment dif-

ference separately in women and in men. However, even if both sex-specific treatment differences are statistically significant, this approach does not address the question of whether the magnitude of benefit depends on sex. Moreover, subdividing the data into subgroups reduces the study's power to detect treatment differences, because not only are the sample sizes reduced, but the number of statistical tests needed is also more than double that required to test for an interaction.

In practice, multiple subgroup analyses are frequently performed. For example, in this issue of the *Journal*, Bhatt et al. (pages 1706–1717) report having performed 20 prespecified analyses in subgroups defined according to different baseline variables. When multiple interaction tests are conducted, each using a nominal criterion (say, P=0.05) to assess statistical significance, the probability of a false positive result — that is, of appearing to find an interaction when none exists — can be greatly inflated. For example, when treatments have identical efficacy, the probability of finding at least one "statistically significant" interaction test when 10 independent interaction tests are undertaken is 40 percent (see graph). The more subgroup analyses conducted, the higher the probability of one or more chance findings that may be misinterpreted as clinically directive.

One way to correct for the inflated false positive rate when multiple subgroup analyses are conducted is to apply a stricter criterion than the usual P=0.05 for judging the significance of each interaction test.[1,2] If K independent tests are conducted, one way to ensure that the overall chances of a false positive result are no greater than 5 percent (0.05) is for each test to use a criterion of $(1–0.95)^{1/K}$, or about 0.05÷K, to assess statistical significance. For example, if 10 tests are conducted, each one should use 0.005 as the threshold for significance. False positive rates are also inflated when the multiple interaction tests are not independent of one another; since corrections for this problem require information about the correlation among the tests,[2] the criteria for statistical significance used for independent tests are commonly applied, even though these criteria may be conservative.

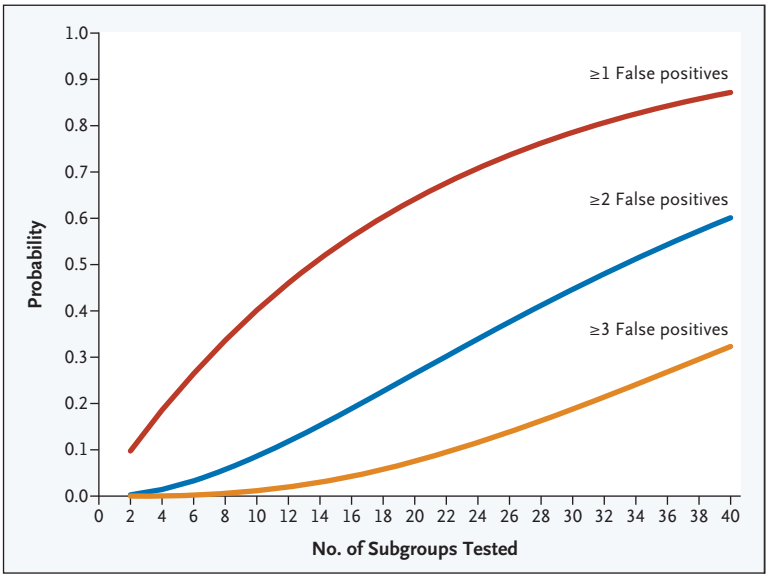
In the 20 subgroup analyses conducted by Bhatt et al., only one interaction test, for symptomatic versus asymptomatic patients (see the article for the precise definitions), gives an uncorrected P value smaller than 0.05 (0.045). Had the interaction tests been assessed with a criterion of 0.05÷20 (0.0025) to account for the fact that 20 were conducted, none would have come close to reaching statistical significance.

Instead of assessing an uncorrected P value against a stricter criterion for significance to account for multiple subgroup analyses, one can sometimes correct the P value so that it can be compared with the usual criterion of P=0.05. When K independent interaction tests are performed, the appropriate correction for the smallest of the resulting P values — say, P* — is $1–(1–P^*)^K$. This formula can be modified for correlated tests, and if applied without modification, it will usually be conservative. Its application to the analyses by Bhatt et al. gives a corrected P value of 0.60 for the interaction test of whether the relative efficacy of clopidogrel depends on symptomatic status.

The inflation of false positive rates by the application of multiple statistical tests applies to both prespecified and post hoc subgroup analyses. The important distinction is that the number of prespecified subgroup analyses is known and determined before the data are examined (though in some cases, important details such as how variables such as age will be categorized are not specified in advance). In contrast, when a report presents the results of post hoc subgroup analyses, it may be unclear why and how the subgroups were selected and how many other subgroups were analyzed. Post hoc subgroup analyses undertaken because of an intriguing trend seen in the results or selective reporting of certain subgroup analyses can be especially misleading.[1]

Authors and medical journals have a responsibility to ensure that the reporting of subgroup analyses is transparent. Ignorance of the total number of subgroup analyses, which ones were prespecified and which were post hoc, and whether any were suggested by the data makes it very difficult to interpret the reported results. When an interaction test for a baseline variable fails to reach the appropriate threshold for significance, conclusions about a differential treatment benefit related to this variable should be avoided or presented with caution.

When subgroup analyses are properly conducted, presentation of their results can be informative, especially when the treatments being compared are used in practice. When reporting subgroup analyses, it is best not to present P values for within-subgroup comparisons, but rather to give an estimate of the magnitude of the treatment difference and a cor-



Probability That Multiple Subgroup Analyses Will Yield at Least One (Red), Two (Blue), or Three (Yellow) False Positive Results.

responding confidence interval. This information can be presented concisely in a figure, along with other summary information, as illustrated by Antman et al. in a recent issue of the *Journal*.[5] These confidence intervals should not be used to infer whether a treatment difference in a subgroup is statistically significant, on the basis of whether the interval excludes the hypothesis of equality between treatment groups, since such analyses suffer from the same problems as the use of multiple statistical tests. Rather, they should be interpreted as providing a plausible range of treatment differences consistent with the trial results.

Overstating the results of subgroup analyses can misinform future research and lead to suboptimal clinical practice. Yet avoiding any presentation of subgroup analyses because of their history of being overinterpreted is a steep price to pay for a problem that can be remedied by more responsible analysis and reporting. Ultimately, medical research and patients are best served when subgroup analyses are well planned and appropriately analyzed and when conclusions and recommendations about clinical practice are guided by the strength of the evidence.

Dr. Lagakos is a professor of biostatistics at the Harvard School of Public Health, Boston, and a statistical consultant to the *Journal*.

1. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991;266:93-8.
2. Bailar JC III, Mosteller F, eds. Medical uses of statistics. 2nd ed. Waltham, Mass.: NEJM Books, 1992.
3. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000;355:1064-9.
4. Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. Am Heart J 2000;139:952-61.
5. Antman EM, Morrow DA, McCabe CH, et al. Enoxaparin versus unfractionated heparin with fibrinolysis for ST-elevation myocardial infarction. N Engl J Med 2006;354:1477-88.

CORRECTION

The Challenge of Subgroup Analyses — Reporting without Distorting

The Challenge of Subgroup Analyses — Reporting without Distorting . On page 1668, the formula on line 3 of the middle column should have read ``1–$0.95^{1/K}$'' rather than ``$(1–0.95)^{1/K}$,'' as printed. We regret the error.

N Engl J Med 2006;355:533-b