SPECIAL REPORT

Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials

Rui Wang, M.S., Stephen W. Lagakos, Ph.D., James H. Ware, Ph.D., David J. Hunter, M.B., B.S., and Jeffrey M. Drazen, M.D.

Medical research relies on clinical trials to assess therapeutic benefits. Because of the effort and cost involved in these studies, investigators frequently use analyses of subgroups of study participants to extract as much information as possible. Such analyses, which assess the heterogeneity of treatment effects in subgroups of patients, may provide useful information for the care of patients and for future research. However, subgroup analyses also introduce analytic challenges and can lead to overstated and misleading results.[1-7] This report outlines the challenges associated with conducting and reporting subgroup analyses, and it sets forth guidelines for their use in the *Journal.* Although this report focuses on the reporting of clinical trials, many of the issues discussed also apply to observational studies.

SUBGROUP ANALYSES AND RELATED CONCEPTS

SUBGROUP ANALYSIS

By "subgroup analysis," we mean any evaluation of treatment effects for a specific end point in subgroups of patients defined by baseline characteristics. The end point may be a measure of treatment efficacy or safety. For a given end point, the treatment effect — a comparison between the treatment groups — is typically measured by a relative risk, odds ratio, or arithmetic difference. The research question usually posed is this: Do the treatment effects vary among the levels of a baseline factor?

A subgroup analysis is sometimes undertaken to assess treatment effects for a specific patient characteristic; this assessment is often listed as a primary or secondary study objective. For example, Sacks et al.[8] conducted a placebo-controlled trial in which the reduction in the incidence of coronary events with the use of pravastatin was examined in a diverse population of persons who had survived a myocardial infarction. In subgroup analyses, the investigators further examined whether the efficacy of pravastatin relative to placebo in preventing coronary events varied according to the patients' baseline low-density lipoprotein (LDL) levels.

Subgroup analyses are also undertaken to investigate the consistency of the trial conclusions among different subpopulations defined by each of multiple baseline characteristics of the patients. For example, Jackson et al.[9] reported the outcomes of a study in which 36,282 postmenopausal women 50 to 79 years of age were randomly assigned to receive 1000 mg of elemental calcium with 400 IU of vitamin D_3 daily or placebo. Fractures, the primary outcome, were ascertained over an average follow-up period of 7.0 years; bone density was a secondary outcome. Overall, no treatment effect was found for the primary outcome; that is, the active treatment was not shown to prevent fractures. The effect of calcium plus vitamin D supplementation relative to placebo on the risk of each of four fracture outcomes was further analyzed for consistency in subgroups defined by 15 characteristics of the participants.

HETEROGENEITY AND STATISTICAL INTERACTIONS

The heterogeneity of treatment effects across the levels of a baseline variable refers to the circumstance in which the treatment effects vary across the levels of the baseline characteristic. Heterogeneity is sometimes further classified as being either quantitative or qualitative. In the first case, one treatment is always better than the other, but by various degrees, whereas in the second case, one treatment is better than the other for one subgroup of patients and worse than the other for

another subgroup of patients. Such variation, also called "effect modification," is typically expressed in a statistical model as an interaction term or terms between the treatment group and the baseline variable. The presence or absence of interaction is specific to the measure of the treatment effect.

The appropriate statistical method for assessing the heterogeneity of treatment effects among the levels of a baseline variable begins with a statistical test for interaction.[10-13] For example, Sacks et al.[8] showed the heterogeneity in pravastatin efficacy by reporting a statistically significant (P=0.03) result of testing for the interaction between the treatment and baseline LDL level when the measure of the treatment effect was the relative risk. Many trials lack the power to detect heterogeneity in treatment effect; thus, the inability to find significant interactions does not show that the treatment effect seen overall necessarily applies to all subjects. A common mistake is to claim heterogeneity on the basis of separate tests of treatment effects within each of the levels of the baseline variable.[6,7,14] For example, testing the hypothesis that there is no treatment effect in women and then testing it separately in men does not address the question of whether treatment differences vary according to sex. Another common error is to claim heterogeneity on the basis of the observed treatment-effect sizes within each subgroup, ignoring the uncertainty of these estimates.

MULTIPLICITY

It is common practice to conduct a subgroup analysis for each of several — and often many — baseline characteristics, for each of several end points, or for both. For example, the analysis by Jackson and colleagues[9] of the effect of calcium plus vitamin D supplementation relative to placebo on the risk of each of four fracture outcomes for 15 participant characteristics resulted in a total of 60 subgroup analyses.

When multiple subgroup analyses are performed, the probability of a false positive finding can be substantial.[7] For example, if the null hypothesis is true for each of 10 independent tests for interaction at the 0.05 significance level, the chance of at least one false positive result exceeds 40%. Thus, one must be cautious in the interpretation of such results. There are several methods for addressing multiplicity that are based on the use of more stringent criteria for statistical significance than the customary P<0.05.[7,15] A less formal approach for addressing multiplicity is to note the number of nominally significant interaction tests that would be expected to occur by chance alone. For example, after noting that 60 subgroup analyses were planned, Jackson et al.[9] pointed out that "Up to three statistically significant interaction tests (P<0.05) would be expected on the basis of chance alone," and then they incorporated this consideration in their interpretation of the results.

PRESPECIFIED ANALYSIS VERSUS POST HOC ANALYSIS

A prespecified subgroup analysis is one that is planned and documented before any examination of the data, preferably in the study protocol. This analysis includes specification of the end point, the baseline characteristic, and the statistical method used to test for an interaction. For example, the Heart Outcomes Prevention Evaluation 2 investigators[16] conducted a study involving 5522 patients with vascular disease or diabetes to assess the effect of homocysteine lowering with folic acid and B vitamins on the risk of a major cardiovascular event. The primary outcome was a composite of death from cardiovascular causes, myocardial infarction, and stroke. In the Methods section of their article, the authors noted that "Prespecified subgroup analyses involving Cox models were used to evaluate outcomes in patients from regions with folate fortification of food and regions without folate fortification, according to the baseline plasma homocysteine level and the baseline serum creatinine level." Post hoc analyses refer to those in which the hypotheses being tested are not specified before any examination of the data. Such analyses are of particular concern because it is often unclear how many were undertaken and whether some were motivated by inspection of the data. However, both prespecified and post hoc subgroup analyses are subject to inflated false positive rates arising from multiple testing. Investigators should avoid the tendency to prespecify many subgroup analyses in the mistaken belief that these analyses are free of the multiplicity problem.

SUBGROUP ANALYSES IN THE *JOURNAL* — ASSESSMENT OF REPORTING PRACTICES

As part of internal quality-control activities at the *Journal*, we assessed the completeness and quality of subgroup analyses reported in the *Journal* during the period from July 1, 2005, through June 30, 2006. A detailed description of the study methods can be found in the Supplementary Appendix, available with the full text of this article at www.nejm.org. In this report, we describe the clarity and completeness of subgroup-analysis reporting, evaluate the authors' interpretation and justification of the results of subgroup analyses, and recommend guidelines for reporting subgroup analyses.

Among the original articles published in the *Journal* during the period from July 1, 2005, through June 30, 2006, a total of 95 articles reported primary outcome results from randomized clinical trials. Among these 95 articles, 93 reported results from one clinical trial; the remaining 2 articles reported results from two trials. Thus, results from 97 trials were reported, from which subgroup analyses were reported for 59 trials (61%). Table 1 summarizes the characteristics of the trials. We found that larger trials and multicenter trials were significantly more likely to report subgroup analyses than smaller trials and single-center trials, respectively. With the use of multivariate logistic-regression models, when ranked according to the number of participants enrolled in a trial and compared with trials with the fewest participants, the odds ratio for reporting subgroup analyses for the second quartile was 1.38 (95% confidence interval [CI], 0.45 to 4.20), for the third quartile was 1.98 (95% CI, 0.62 to 6.24), and for the fourth quartile was 8.90 (95% CI, 2.10 to 37.78) (P=0.02, trend test). The odds ratio for reporting subgroup analyses in multicenter trials as compared with single-center trials was 4.33 (95% CI, 1.56 to 12.16).

Among the 59 trials that reported subgroup analyses, these analyses were mentioned in the Methods section for 21 trials (36%), in the Results section for 57 trials (97%), and in the Discussion section for 37 trials (63%); subgroup analyses were reported in both the text and a figure or table for 39 trials (66%). Other characteristics of the reports are shown in Figure 1. In general, we are unable to determine the number of subgroup analyses conducted; we attempted to count the number of subgroup analyses reported in the article and found that this number was unclear in nine articles (15%). For example, Lees et al.[17] reported that "We explored analyses of numerous other subgroups to assess the effect of baseline prognostic factors or coexisting conditions on the

Table 1. Characteristics and Predictors of Reporting Subgroup Analyses in 97 Clinical Trials.*

Variable	Trials Reporting Subgroup Analyses	P Value†	
	No. of Trials/ Total No. (%)	Univariate Odds Ratio	Multivariate Odds Ratio
No. of subjects		0.002†	0.02†
≤218	11/25 (44)		
219–429	13/25 (52)		
430–1012	14/23 (61)		
>1012	21/24 (88)		
Superiority trial		0.25	0.89
Yes	53/84 (63)		
No	6/13 (46)		
Trial sites		0.005	0.05
Single-center	7/21 (33)		
Multicenter	52/76 (68)		
Type of disease studied		0.18	0.37
Cardiovascular	16/20 (80)		
Infectious	2/7 (29)		
Oncologic	9/11 (82)		
Respiratory	7/10 (70)		
Pediatric	5/10 (50)		
Psychiatric or neurologic	6/10 (60)		
Metabolic, endocrine, or gastrointestinal	5/10 (50)		
Gynecologic	3/6 (50)		
Other	6/13 (46)		
Statistically significant primary end point		0.24	0.38
Yes	35/62 (56)		
No	24/35 (69)		

* A total of 59 trials reported subgroup analyses.
† P values were determined with the use of trend tests.

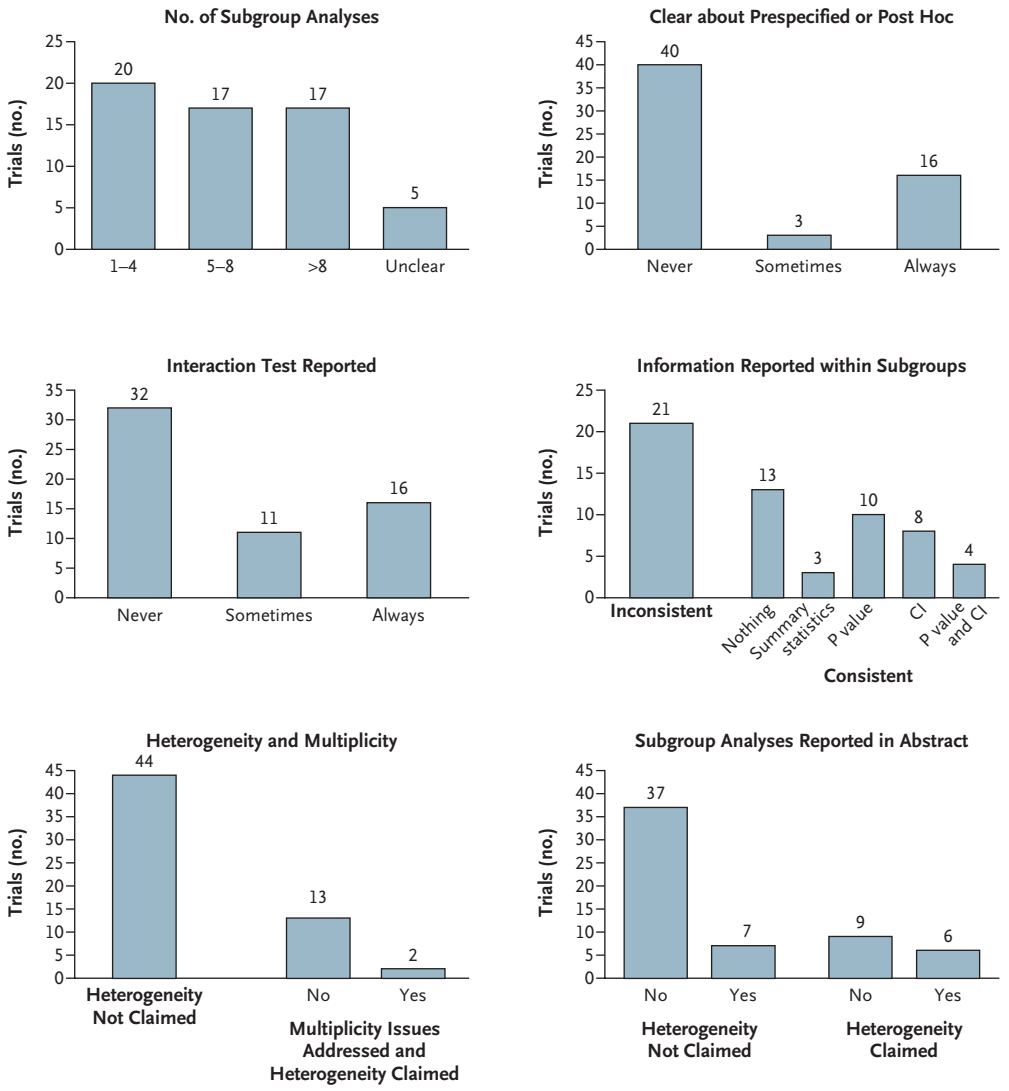

Figure 1. Reporting of Subgroup Analyses from 59 Clinical Trials.

The specific reporting characteristics examined in this quality-improvement exercise are indicated in each panel. CI denotes confidence interval.

treatment effect but found no evidence of nominal significance for any biologically likely factor." For four of these nine articles, we were able to determine that at least eight subgroup analyses were reported. In 40 trials (68%), it was unclear whether any of the subgroup analyses were prespecified or post hoc, and in 3 others (5%) it was unclear whether some were prespecified or post hoc. Interaction tests were reported to have been used to assess the heterogeneity of treatment effects for all subgroup analyses in only 16 trials (27%), and they were reported to be used for some, but not all, subgroup analyses in 11 trials (19%).

We assessed whether information was provided about treatment effects within the levels of each subgroup variable (Fig. 1). In 25 trials (42%), information about treatment effects was reported consistently for all of the reported subgroup analyses, and in 13 trials (22%), nothing was reported. Investigators in 15 trials (25%), all using superiority designs,[10] claimed heterogeneity of treatment effects between at least one subject sub-

group and the overall study population (see Table 1 of the Supplementary Appendix). For 4 of these 15 trials, this claim was based on a nominally significant interaction test, and for 4 others it was based on within-subgroup comparisons only. In the remaining seven trials, significant results of interaction tests were reported for some but not all subgroup analyses. When heterogeneity in the treatment effect was reported, for two trials (13%), investigators offered caution about multiplicity, and for four trials (27%), investigators noted the heterogeneity in the Abstract section.

ANALYSIS OF OUR FINDINGS AND GUIDELINES FOR REPORTING SUBGROUPS

In the 1-year period studied, the reporting of subgroup analyses was neither uniform nor complete. Because the design of future clinical trials can depend on the results of subgroup analyses, uniformity in reporting would strengthen the foundation on which such research is built. Furthermore, uniformity of reporting will be of value in the interval between recognition of a potential subgroup effect and the availability of adequate data on which to base clinical decisions.

Problems in the reporting of subgroup analyses are not new.[1-6,18] Assmann et al.[2] reported shortcomings of subgroup analyses in a review of the results of 50 trials published in 1997 in four leading medical journals. More recently, Hernández et al.[4] reviewed the results of 63 cardiovascular trials published in 2002 and 2004 and noted the same problems. To improve the quality of reports of parallel-group randomized trials, the Consolidated Standards of Reporting Trials statement was proposed in the mid-1990s and revised in 2001.[19] Although there has been considerable discussion of the potential problems associated with subgroup analysis and recommendations on when and how subgroup analyses should be conducted and reported,[19,20] our analysis of recent articles shows that problems and ambiguities persist in articles published in the *Journal*. For example, we found that in about two thirds of the published trials, it was unclear whether any of the reported subgroup analyses were prespecified or post hoc. In more than half of the trials, it was unclear whether interaction tests were used, and in about one third of the trials, within-level results were not presented in a consistent way.

Guidelines for Reporting Subgroup Analysis.

In the Abstract:

Present subgroup results in the Abstract only if the subgroup analyses were based on a primary study outcome, if they were prespecified, and if they were interpreted in light of the totality of prespecified subgroup analyses undertaken.

In the Methods section:

Indicate the number of prespecified subgroup analyses that were performed and the number of prespecified subgroup analyses that are reported. Distinguish a specific subgroup analysis of special interest, such as that in the article by Sacks et al.,[8] from the multiple subgroup analyses typically done to assess the consistency of a treatment effect among various patient characteristics, such as those in the article by Jackson et al.[9] For each reported analysis, indicate the end point that was assessed and the statistical method that was used to assess the heterogeneity of treatment differences.

Indicate the number of post hoc subgroup analyses that were performed and the number of post hoc subgroup analyses that are reported. For each reported analysis, indicate the end point that was assessed and the statistical method used to assess the heterogeneity of treatment differences. Detailed descriptions may require a supplementary appendix.

Indicate the potential effect on type I errors (false positives) due to multiple subgroup analyses and how this effect is addressed. If formal adjustments for multiplicity were used, describe them; if no formal adjustment was made, indicate the magnitude of the problem informally, as done by Jackson et al.[9]

In the Results section:

When possible, base analyses of the heterogeneity of treatment effects on tests for interaction, and present them along with effect estimates (including confidence intervals) within each level of each baseline covariate analyzed. A forest plot[21,22] is an effective method for presenting this information.

In the Discussion section:

Avoid overinterpretation of subgroup differences. Be properly cautious in appraising their credibility, acknowledge the limitations, and provide supporting or contradictory data from other studies, if any.

When properly planned, reported, and interpreted, subgroup analyses can provide valuable information. With the availability of Web supplements, the opportunity exists to present more detailed information about the results of a trial. The purpose of the guidelines (see box) is to encourage more clear and complete reporting of subgroup analyses. In some settings, a trial is conducted with a subgroup analysis as one of the primary objectives. These guidelines are directly applicable to the reporting of subgroup analyses in the primary publication of a clinical trial when the subgroup analyses are not among the primary objectives. In other settings, including observational studies, we encourage complete and thorough reporting of the subgroup analyses in the spirit of the guidelines listed.

The editors and statistical consultants of the *Journal* consider these guidelines to be important in the reporting of subgroup analyses. The goal is to provide transparency in the statistical meth-

ods used in order to increase the clarity and completeness of the information reported. As always, these are guidelines and not rules; additions and exemptions can be made as long as there is a clear case for such action.

No potential conflict of interest relevant to this article was reported.

We thank Doug Altman, John Bailar, Colin Begg, Mohan Beltangady, Marc Buyse, David DeMets, Stephen Evans, Thomas Fleming, David Harrington, Joe Heyse, David Hoaglin, Michael Hughes, John Ioannidis, Curtis Meinert, James Neaton, Robert O'Neill, Ross Prentice, Stuart Pocock, Robert Temple, Janet Wittes, and Marvin Zelen for their helpful comments.

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. 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.
3. Pocock SJ, Assmann SF, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 2002; 21:2917-30.
4. Hernández A, Boersma E, Murray G, Habbema J, Steyerberg E. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? Am Heart J 2006;151:257-64.
5. Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. Am Heart J 2000;139:952-61.
6. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005; 365:176-86.
7. Lagakos SW. The challenge of subgroup analyses — reporting without distorting. N Engl J Med 2006;354:1667-9. [Erratum, N Engl J Med 2006;355:533.]
8. Sacks FM, Pfeffer MA, Moye LA, et al. The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. N Engl J Med 1996;335:1001-9.
9. Jackson RD, LaCroix AZ, Gass M, et al. Calcium plus vitamin D supplementation and the risk of fractures. N Engl J Med 2006; 354:669-83. [Erratum, N Engl J Med 2006;354:1102.]
10. Pocock SJ. Clinical trials: a practical approach. Chichester, England: John Wiley, 1983.
11. Halperin M, Ware JH, Byar DP, et al. Testing for interaction in an I×J×K contingency table. Biometrika 1977;64:271-5.
12. Simon R. Patient subsets and variation in therapeutic efficacy. Br J Clin Pharmacol 1982;14:473-82.
13. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics 1985;41:361-72.
14. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters T. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004;57:229-36.
15. Bailar JC III, Mosteller F, eds. Medical uses of statistics. 2nd ed. Waltham, MA: NEJM Books, 1992.
16. Lonn E, Yusuf S, Arnold MJ, et al. Homocysteine lowering with folic acid and B vitamins in vascular disease. N Engl J Med 2006;354:1567-77. [Erratum, N Engl J Med 2006;355:746.]
17. Lees KR, Zivin JA, Ashwood T, et al. NXY-059 for acute ischemic stroke. N Engl J Med 2006;354:588-600.
18. Al-Marzouki S, Roberts I, Marshall T, Evans S. The effect of scientific misconduct on the results of clinical trials: a Delphi survey. Contemp Clin Trials 2005;26:331-7.
19. Moher D, Schulz KF, Altman DG, et al. The CONSORT Statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. (Accessed November 1, 2007, at http://www.consort-statement.org/.)
20. International Conference on Harmonisation (ICH). Guidance for industry: E9 statistical principles for clinical trials. Rockville, MD: Food and Drug Administration, September 1998. (Accessed November 1, 2007, at http://www.fda.gov/cder/guidance/ICH_E9-fnl.PDF.)
21. Cuzick J. Forest plots and the interpretation of subgroups. Lancet 2005;365:1308.
22. Wactawski-Wende J, Kotchen JM, Anderson GL, et al. Calcium plus vitamin D supplementation and the risk of colorectal cancer. N Engl J Med 2006;354:684-96.