REVIEW

GUIDELINES ON HOW TO ASSESS THE VALIDITY OF RESULTS PRESENTED IN SUBGROUP ANALYSIS OF CLINICAL TRIALS

Edson Duarte Moreira[1] and Ezra Susser[2]

RHCFAP/3074

MOREIRA ED et al. - Guidelines on how to assess the validity of results presented in subgroup analysis of clinical trials. **Rev. Hosp. Clín. Fac. Med. S. Paulo 57** (2):83-88, 2002.

In observational studies, identification of associations within particular subgroups is the usual method of investigation. As an exploratory method, it is the bread and butter of epidemiological research. Nearly everything that has been learned in epidemiology has been derived from the analysis of subgroups. In a randomized clinical trial, the entire purpose is the comparison of the test subjects and the controls, and when there is particular interest in the results of treatment in a certain section of trial participants, a subgroup analysis is performed. These subgroups are examined to see if they are liable to a greater benefit or risk from treatment. Thus, analyzing patient subsets is a natural part of the process of improving therapeutic knowledge through clinical trials. Nevertheless, the reliability of subgroup analysis can often be poor because of problems of multiplicity and limitations in the numbers of patients studied. The naive interpretation of the results of such examinations is a cause of great confusion in the therapeutic literature. We emphasize the need for readers to be aware that inferences based on comparisons between subgroups in randomized clinical trials should be approached more cautiously than those based on the main comparison. That is, subgroup analysis results derived from a sound clinical trial are not necessarily valid; one must not jump to conclusions and accept the validity of subgroup analysis results without an appropriate judgment.

DESCRIPTORS: **Randomized clinical trial. Subgroup analysis. Epidemiological methods. Interaction. Effect modification.**

Many scientific publications include, in addition to the main results, a subgroup or subset analysis. In observational studies, identification of associations within particular subgroups is the usual method of investigation, with the data being presented by sex, by age group, and so on. As an exploratory method, it is the bread and butter of epidemiological research. Nearly everything that has been learned in epidemiology has been derived from the analysis of subgroups.

In a randomized clinical trial (RCT), the entire purpose is expressed in the comparison of 2 subgroups—the test subjects and the controls. And quite commonly, when there is particular interest in the results of treatment in a certain section of trial participants (for example, the women, the very young or very old, or those with a specific pattern of disease), a subgroup analysis is performed. These subgroups are examined to see if they are liable to a greater benefit or risk from treatment. Thus, analyses of subgroup data from RCTs are undertaken to identify "effect modifiers", characteristics of the patients or treatment that modify the effect of the intervention under study.

Situations in which a particular subgroup does not benefit at all from treatment, or conversely, in which the benefit is extremely large are certainly important to discover. Unfortunately, such situations in which the treatment seems highly effective in one subgroup and not at all in others happen quite easily just by chance[1]. As shown by Peto[2], if patients are divided into 2 subgroups of similar size and if the treatment is just significant overall ($P < 0.05$), there is a 1 in 3 chance that the treatment effect will be large and sig-

[1] From the Epidemiology and Statistic Nuclei of the "Centro de Pesquisas Gonçalo Muniz, Fundação Oswaldo Cruz", "Associação Obras Sociais Irmã Dulce", "Escola Bahiana de Medicina e Saúde Pública" and "Hospital São Rafael", Bahia. [2] Division of Epidemiology, School of Public Health, Columbia University, New York.

nificant in one of the subgroups and negligible in the other. Thus, looking at treatment effect in individual subgroups is a sure way to be misled by the play of chance. This is particularly true when, as is almost always the case, there are several sensible ways to partition the patients. Not surprisingly then, the literature is fraught with contradictory observations because the subgroup that "clearly" benefits from treatment varies at random[3,4]. Nonetheless, if we always insisted on focusing all our attention on the overall results, we would clearly not exploit the data available to their full potential. Even if we know that we would make fewer mistakes, on average, by believing the overall results and not the subgroup results, we still need to see the latter as urgently as we need to see the former.

The results of subgroup analyses have had major effects, sometimes harmful, on treatment recommendations. For example, many patients with suspected myocardial infarction who could have benefited from thrombolytic therapy may not have received this treatment as a result of subgroup analyses based on the duration of symptoms before treatment and the conclusion that streptokinase was only effective in patients treated within 6 hours after the onset of pain[5-7]. A later, larger trial showed that streptokinase was effective up to 24 hours after the onset of symptoms[8]. Conversely, in the European Working Party on High Blood Pressure trial, a reduction in cardiovascular mortality was shown among those using active treatment as compared with those taking placebo[9]. Further subgroup analysis went on to demonstrate that no case could be made for treatment in patients aged 80 and over.

Ideally, medical trials should yield reliable and precise predictions of clinical outcomes as a function of treatment and patient characteristics. However, even among randomized studies, conflicting reports exist of trials purporting to evaluate the relative benefits of the same treatments[10,11]. These inconsistencies are often the result of inappropriate subset analysis conducted in an effort to determine which treatment is preferable for what kind of patient. Analyzing patient subsets is a natural part of the process of improving therapeutic knowledge through clinical trials. But the naive interpretation of the results of such examination is a cause of great confusion in the therapeutic literature. The adopted study designs and methods of statistical analysis determine the extent to which erroneous conclusions are likely to result from the use of subset analysis[12-16].

The reliability of subgroup analysis can often be poor due to problems of multiplicity and limitations in numbers of patients studied[17,18]. First, examining the former condition, suppose that 2 treatments have been randomly assigned and subjects have been partitioned into N mutually exclusive subsets. For each subset we perform a statistical significance test with alpha set at 0.05. Even if the treatments are identical, the probability of obtaining at least one significant result for $N = 10$ is about 0.40 [Probability = $1-(1-\text{alpha})^N$]. That is, if the treatments are equivalent for all 10 subsets, there is a 40% chance that at least one difference will appear "significant" at the 0.05 level. For 5 subsets, the probability is about 23%. For many clinical trials, the number of subgroups that might be examined is far greater than 10. Moreover, very often authors will not report how many subsets were examined[12,19,20]. Thus, the problem of performing many comparisons is that it may not be reasonable to interpret individual results at face value. In a subgroup analysis, many tests are usually performed, and the overall error probability is much higher than the nominal significance level of the test[21,22]. A priori, the presence of important true interactions is unknown, whereas the spurious appearance of variability in relative efficacy among many subsets is almost a certainty.

Second, concerning the problems with subgroup analysis caused by limitations in numbers of patients enrolled in the study, it is usual that constraints of time and money restrict the sample size of clinical trials to the minimum number necessary to detect a difference between the treated and the untreated groups. As the total number of individuals studied in the experimental and control groups is further partitioned in subsets of particular interest, the numbers of study subjects in these subgroups become too small to support a demonstration of statistical significance. Hence, estimation of interactions between treatment and patients' characteristics is usually plagued by insufficient statistical power.

For these reasons, subgroup analyses are potentially misleading, and there is a tendency to over emphasize the results of such analyses. Therefore, it is not surprising that it has been argued that treatment recommendations based on subgroup analyses may do more harm than good[3,4,10,15,23-25]. Different classification schemes can generate several different sets of subgroups, resulting in numerous subset analyses in the search for more specific differences between treatments. Because of this large amount of data, there is the danger that significance testing may be used excessively, and the credibility of subgroup analysis be poor, since, inevitably, the chance of reporting some false-positive finding increases[26-29]; this is particularly true when the subgroup categories were not clearly defined by the original hypothesis. Thus, caution must be exercised when evaluating the results of subgroup analysis.

Although the best evidence for the efficacy of medical interventions comes from well-conducted, randomized controlled trials, unless such trials are reported adequately, it is impossible to assess that information. Results of a RCT cannot be appropriately interpreted without detailed information about the methods used in the study. Ideally, the report of such an evaluation needs to convey to the reader information concerning the design, conduct, and analysis of the trial. This information should provide the reader with the ability to make informed judgments regarding the validity of the results reported in the trial. However, such reports frequently omit important features of design and analysis[19,30,31].

In response to increasing evidence that reporting of RCTs is imperfect, there has been a concerted effort to set standards for reporting RCTs[12,19,32,33]. These guidelines attempt to help authors and editors improve the completeness of reporting RCTs. These publications recommend important aspects to be mentioned for improving the reporting of trials so editors can provide authors with a list of features to be used as a checklist for the report of trials. The Standards of Reporting Trials (SORT) group[34] and the Asilomar Working Group on Recommendations for Reporting of Clinical Trials in Biomedical Literature[35] met to produce a checklist of items that should be included when reporting a clinical trial, along with a suggestion that editors add it to the Instruction for Authors. This meeting resulted in the Consolidated Standards of Reporting Trials (CONSORT) statement (a checklist of 21 items and a flow diagram) that identifies key pieces of information necessary to evaluate the validity of a trial report[36].

Even though the CONSORT statement recommendations can lead to more accurate and complete reporting of the main comparisons between treatment groups in RCTs, the quality of reporting subgroup analysis of RCT will not necessarily improve through the use of these guidelines alone. That is, these suggestions do not guarantee that results from comparisons of treatment responses in different categories of patients are valid. While there have been some reports on guidelines for assessing the strength of inferences based on subgroup analyses that can assist clinicians in making decisions regarding whether to base a treatment decision on the results of such analyses[14,18,24,37], neither the CONSORT statement nor the Instruction for Authors in leading medical journals address specific aspects of reporting subgroup analyses in RCTs.

Recently, we reviewed the current practice of reporting methodological aspects of subgroup analysis in randomized controlled clinical trials in 4 leading scientific journals[38] and developed a list of important items to be included in reports of subgroup analysis of RCTs (Table 1). Our data show that inaccurate and incomplete reports occur frequently. When authors fail to mention essential features of the design and methods used in a subgroup analysis, they are omitting much needed information for the readers to assess the validity of the results presented. At the same time, readers need to be more aware of the important distinctions between inferences based on the main comparison of a RCT, and those based on results of subgroup analyses. In addition to information on the design and methods of the trial, specific aspects pertaining the conduct of the subset analysis must be reported in order to make informed judgments on the strength of inferences drawn from such analyses of RCTs.

CONCLUSIONS

Current practices of reporting subgroup analysis of RCTs are far from ideal. Incomplete and inaccurate reports are the norm, rather than the exception. This incomplete and inaccurate reporting imposes severe limitations on the correct interpretation of the results; it almost always precludes correct interpretation. We emphasize the need for readers to be aware that inferences based on comparisons between subgroups in RCTs should be approached more cautiously than those based on the main comparison. That is, subgroup analysis results derived from a sound clinical trial are not necessarily valid. Unless specific methodologic aspects of the subgroup analysis are explicitly mentioned, accurate evaluation by the reader is not possible. One must not jump to conclusions and accept the validity of subgroup analysis results without an appropriate judgment.

Table 1 - List of important methodological items in reporting subgroup analysis in randomized clinical trials.

1. *A priori/post-hoc subgroup analysis* (information on whether the subgroup analysis preceded or followed the analysis);
2. *Number of subgroups examined* (information about how many subgroups were examined);
3. *Justification for subset definition* (information explaining how the subgroups were stratified);
4. *When subgroups were delineated* (information on whether the subgroups were defined after or before randomization);
5. *Statistical methods* (the names of specifics tests or techniques used for statistical analyses);
6. *Power* (information describing the determination of sample size or the size of detectable differences);
7. *Clinical significance* (information on the clinical significance of the interaction); and
8. *Overall treatment comparison* (information about the significance of the comparison between the main treatment groups).

We need to enforce recommendations and guidelines for reporting results of subgroup analysis in clinical trials. This enforcement would improve the quality of published trials, helping clinicians make more informed decisions about the best treatment choices. As readers find these methodological features repeatedly mentioned in RCTs, they will become more aware of the relevance of these items for the interpretation of the results, improving their assessment of the validity of the study findings. In addition, more structured reports will help editors and reviewers in their deliberations regarding submitted manuscripts.

The decision of whether to base clinical practice on the average estimate of effect from the overall analysis or on a subgroup analysis depends on the careful assessment of the criteria described throughout this text. It might be tempting to take one extreme or the other: to base decisions either on the overall estimate of effect or on the most applicable subgroup analysis. However a thoughtful approach is more likely to result in the most benefit and the least harm for patients. It is essential that this information is completely and accurately reported in clinical trials. A better report of methods of subgroup analysis will ultimately benefit patients, since more comprehensive and complete reports are more likely to lead clinicians to make correct therapeutical decisions.

We acknowledge that applying these standards for reporting subgroup analysis presents intrinsic difficulties and challenges. The list of recommendations for reporting subgroup analysis of clinical trials that we developed will be modified and improved as it is disseminated and made available to wider audiences. We invite all interested editors and readers to join us in using and improving this checklist.

ACKNOWLEDGMENTS

We are very grateful to Dr. Zena Stein for reviewing the manuscript and her helpful comments, Mrs. Isabella Arapiraca for her secretarial assistance and Dr. Beatriz Helena Carvalho Tess for her wise suggestions that ultimately made it possible to accomplish this work.

RESUMO

RHCFAP/3074

MOREIRA ED e col. - Critérios para avaliar a validade dos resultados apresentados em análises de subgrupo em ensaios clínicos. **Rev. Hosp. Clín. Fac. Med. S. Paulo 57** (2):83-88, 2002.

Em estudos de observação, a identificação de associações dentro de subgrupos particulares é o método habitual de investigação. Como um método exploratório, faz parte do dia-a-dia da pesquisa epidemiológica. Quase tudo o que se sabe hoje em Epidemiologia foi derivado da análise de subgrupo. Em ensaios clínicos randomizados, o propósito principal é a comparação dos indivíduos sob experimentação e os controles, e quando nós estamos particularmente interessados nos resultados do tratamento em uma certa seção de participantes do ensaio, uma análise de subgrupo é executada. Estes subgrupos são examinados para verificar se eles foram objeto de um maior benefício ou malefício secundário ao tratamento. Assim, analisar subconjuntos de pacientes é uma parte natural do processo de melhorar o conhecimento terapêutico através de ensaios clínicos randomizados. Não obstante, a confiança na análise de subgrupo pode ser pobre devido a problemas de multiplicidade de comparações e limitações em números de pacientes estudados. A interpretação simplista dos resultados de tal técnica é uma causa de grande confusão na literatura terapêutica. Nós enfatizamos a necessidade de chamar a atenção dos leitores que as conclusões baseadas em comparações entre subgrupos em ensaios clínicos randomizados sejam examinadas com mais cautela do que aquelas baseadas na comparação principal. Ou seja, resultados de análise de subgrupo provenientes de ensaios clínicos corretamente desenhados não são necessariamente válidos, portanto, não devemos tirar conclusões precipitadas e aceitar a validade dos resultados sem um julgamento judicioso.

DESCRITORES: **Ensaio clínico randomizado. Análise de subgrupo. Métodos epidemiológicos. Interação. Modificação de efeito.**

REFERENCES

1. ARMITAGE P - Controversies and achievements in clinical trials. **Control Clin Trials** 1984; **5** (1): 67.

2. PETO R - Statistical aspects of cancer trials. In: HALNAN KE. **Treatment of cancer**. London, Chapman, 1982. p. 94-115.

3. CUZICK J - The assessment of subgroups in clinical trials. In: BAUM M, KAY R & SCHEURLEN H, ed. - **Clinical trials in early breast cancer**. Birkhausen, Verlag, 1982. p. 1-13.

4. GELBER RD & GOLDHIRSCH - A Interpretation of results from subset analyses within overviews of randomized clinical trials**. Stat Med** 1987; **6** (3): 371.

5. EFFECTIVENESS of intravenous thrombolytic treatment in acute myocardial infarction. Gruppo Italiano Per Lo Studio Della Streptochinasi Nell'infarto Miocardico (GISSI). **Lancet** 1986; **1** (8478): 397.

6. LEE TH, WEISBERG MC, BRAND DA, et al. - Candidates for thrombolysis among emergency room patients with acute chest pain. Potential true- and false-positive rates. **Ann Intern Med** 1989; **110** (12): 957.

7. TATE DA & DEHMER GJ - New challenges for thrombolytic therapy. **Ann Intern Med** 1989; **110** (12): 953.

8. RANDOMISED trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study Of Infarct Survival) Collaborative Group. **Lancet** 1988; **2** (8607): 349.

9. AMERY A, BIRKENHAGER W, BRIXKO R et al. - Efficacy of antihypertensive drug treatment according to age, sex, blood pressure, and previous cardiovascular disease in patients over the age of 60**. Lancet** 1986; **2** (8507): 589.

10. HORWITZ RI - Complexity and contradiction in clinical trial research**. Am J Med** 1987; **82** (3): 498.

11. LEVENSTEIN MJ & BISHOP YM - Analysis and reporting as causes of controversies. In: ROSENOER VM, ROTHSCHILD M, editors. **Controversies in clinical care**. New York, Spectrum, 1981. p. 1-23.

12. POCOCK SJ, HUGHES MD & LEE RJ - Statistical problems in the reporting of clinical trials. A survey of three medical journals. **N Engl J Med** 1987; **317** (7): 426.

13. BUYSE ME - Analysis of clinical trial outcomes: some comments on subgroup analyses. **Control Clin Trials** 1989; **10** (4 Suppl): 187S.

14. STALLONES RA - The use and abuse of subgroup analysis in epidemiological research. **Prev Med** 1987; **16** (2): 183.

15. BYAR DP - Assessing apparent treatment—covariate interactions in randomized clinical trials. **Stat Med** 1985; **4** (3): 255.

16. SHUSTER J & VAN EYS J - Interaction between prognostic factors and treatment. **Control Clin Trials** 1983; **4** (3): 209.

17. BROWN BW - Statistical controversies in the design of clinical trials. **Controlled Clin Trials** 1980; **1**: 13.

18. SIMON R - Patients subsets and variation in therapeutic efficacy. **Br J Clin Pharmac** 1982; **14** (1): 473.

19. DERSIMONIAN R, CHARETTE LJ, MCPEEK B et al. - Reporting on methods in clinical trials. **N Engl J Med** 1982; **306** (22): 1332.

20. MOSTELLER F, GILBERT JP & MCPEEK B - Reporting standards and research strategies for controlled trials. **Controlled Clin Trials** 1980; **1**: 37.

21. ARMITAGE P & PAMAR M - Some approaches to the problem of multiplicity in clinical trials. In: INTERNATIONAL BIOMETRIC CONFERENCE, 13th, Seatle, 1986. *Proceedings*. Seatle, 1986. p. 49.

22. MILLER RG - Simultaneous statistical interference. New York, Springer, 1981.

23. GOLDMAN L & FEINSTEIN AR - Anticoagulants and myocardial infarction. The problems of pooling, drowning, and floating. **Ann Intern Med** 1979; **90** (1): 92.

24. FURBERG CD & MORGAN TM - Lessons from overviews of cardiovascular trials. **Stat Med** 1987; **6** (3): 295.

25. BULPITT CJ - Subgroup analysis. **Lancet** 1988; **2** (8601): 31.

26. SCHNEIDER B - Analysis of clinical trial outcomes: alternative approaches to subgroup analysis. **Control Clin Trials** 1989; **10** (4 Suppl): 176S.

27. BEACH ML & MEIER P - Choosing covariates in the analysis of clinical trials. **Control Clin Trials** 1989; **10** (4 Suppl): 161S.

28. DAVIS CE & LEFFINGWELL DP - Empirical Bayes estimates of subgroup effects in clinical trials. **Control Clin Trials** 1990; **11** (1): 37.

29. HILL AB - Principles of medical statistics. **Lancet**, 1971: 312.

30. MOSTELLER F - Problems of omission in communications. **Clin Pharmacol Ther** 1979; **25**: 761.

31. REIFFENSTEIN RJ, SCHILTROTH AJ & TODD DM - Current standards in reported drug trials. **Can Med Assoc J** 1968; **99** (23): 1134.

32. CHALMERS TC, SMITH H, BLACKBURN B et al. - A method for assessing the quality of a randomized control trial. **Control Clin Trials** 1981; **2** (1): 31.

33. LIBERATI A, HIMEL HN & CHALMERS TC - A quality assessment of randomized control trials of primary treatment of breast cancer**. J Clin Oncol** 1986; **4** (6): 942.

34. A PROPOSAL for structured reporting of randomized controlled trials. The Standards of Reporting Trials Group. **Jama** 1994; **272** (24): 1926.

35. CALL FOR comments on a proposal to improve reporting of clinical trials in the biomedical literature. Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature. **Ann Intern Med** 1994; **121** (11): 894.

36. BEGG C, CHO M, EASTWOOD S et al. - Improving the quality of reporting of randomized controlled trials. The Consort statement. **Jama** 1996; **276** (8): 637.

37. OXMAN AD & GUYATT GH - A consumer’s guide to subgroup analyses. **Ann Intern Med** 1992; **116** (1): 78.

38. MOREIRA EDJ & STEIN Z ES - Reporting on methods of subgroup analysis in clinical trials: A Survey of Four Scientific Journals. **Braz J Med Biol Res** 2001; **34**(11): 1441-1446.

Received for publication on May 08, 2001.