AJNR

The Problem of Subgroup Analyses: An Example from a Trial on Ruptured Intracranial Aneurysms

O. Naggara, J. Raymond, F. Guilbert and D.G. Altman

AJNR Am J Neuroradiol 2011, 32 (4) 633-636
doi: https://doi.org/10.3174/ajnr.A2442
http://www.ajnr.org/content/32/4/633

RESEARCH PERSPECTIVES

The Problem of Subgroup Analyses: An Example from a Trial on Ruptured Intracranial Aneurysms

O. Naggara
J. Raymond
F. Guilbert
D.G. Altman

SUMMARY: The randomized ISAT demonstrated the superiority of endovascular treatment in patients with ruptured intracranial aneurysms considered suitable for either clipping or coiling. A later publication proposed a second look at the results, demonstrating that older patients with ruptured MCA aneurysms appeared to benefit from clipping, in disagreement with the general findings of the trial. Subgroup analyses in randomized trials and observational studies examine whether effects of interventions differ between subgroups according to the characteristics of patients. However, many apparent subgroup effects have been shown to be spurious. Misleading subgroup effects can result in withholding efficacious treatment from patients who would benefit or can encourage ineffective or potentially harmful treatments for patients who would fare better without. Some guidelines for the prudent interpretation of subgroup findings are reviewed.

ABBREVIATIONS: CHUM = Centre Hospitalier de l'Université de Montreal; CI = confidence interval; ISAT = International Subarachnoid Aneurysm Trial; MCA = middle cerebral artery; OR = odds ratio; RCT = randomized controlled trial; SAH = subarachnoid hemorrhage

The ISAT was a turning point in modern neurosurgical history.[1] It was a randomized trial in which neurosurgical clipping and endovascular coiling were compared in over 2000 patients with ruptured intracranial aneurysms considered suitable for either treatment. The trial showed a 7.6% absolute reduction in morbidity and mortality at 1 year in patients treated by coiling.[1] That finding was found to be consistent in most prespecified subgroups of patients.[1] In a further article, the authors reported results for a subgroup of 278 patients older than 65 years of age.[2] We will examine their suggestion that in older patients with small anterior circulation aneurysms, coiling should probably be favored for internal carotid and posterior communicating artery aneurysms, whereas patients with MCA aneurysms would benefit from clipping.

We all understand the reasons for post hoc analyses. Results of an RCT represent the estimated average effect of a medical intervention in a heterogeneous group of patients. Hence, results may not apply to a specific patient. Clinicians treat individuals and are trained to recognize and differentiate categories of patients according to certain characteristics and their combinations (eg, age, sex, clinical presentation, and size and location of the lesion). Therefore, many investigators would like to know treatment effects in specific subpopulations of patients based on patient characteristics measured before randomization. That desire is generally addressed by subgroup analyses. Treatment effects in subgroups might offer clinicians more insight into treating individual patients. However, if sometimes they are informative, subgroup analyses are more frequently misleading. Subgroup analyses have been characterized as "a scientific challenge and a methodological trap,"[3] and many authors have cautioned against overinterpreting them.[4-7]

We will contrast the rigorous rules that regiment the planning and conduct of analyses in randomized trials in general against the dangers of doing post hoc subgroup analyses, by using ISAT as an example. What can reasonably be inferred from subgroup analyses and what would be required of a subgroup finding to overturn the general results of an RCT are 2 questions we wish to address. Finally, some guidelines are proposed to limit the damage that can be caused by an overenthusiastic reliance on subgroup findings. This work was presented at the Tenth Congress of World Federation of Interventional and Therapeutic Neuroradiology in Montreal in June 2009.[8]

Subgroup Analysis in Randomized Controlled Trials and Observational Studies

Treatment recommendations obtained from the overall results of RCTs do not necessarily apply equally to any particular individual. When both coil embolization and clip ligation are considered appropriate options, ISAT has shown that coil embolization, in general, leads to a better outcome at 1 year. Of course, this generalization was not and cannot be verified for all patients, even though the findings appeared reasonably consistent among subgroups of several prespecified variables (location, size, Fisher and World Federation of Neurological Societies grades at presentation, and age).[9]

Planning a study that would be powered to provide evidence for all kinds of patients would lead to eternal trials, with the evidence becoming nonconvincing each time the data are split into other interesting subgroups. However, all physicians, confronted with a treatment decision in a particular patient, would like to know the evidence that pertains most directly and most specifically to that individual. Thus, in both RCTs and observational studies, investigators, trying to meet clinicians' expectations for specific information, frequently conduct subgroup analyses that explore multiple hypotheses. In doing so, they risk confusion between real and chance findings and so may mislead rather than enlighten treatment decisions.[10] Despite repeated warnings and published recommen-

From the Department of Radiology (O.N., J.R., F.G.), International Consortium of Neuroendovascular Centres, Interventional Neuroradiology Research Unit, University of Montreal, CHUM, Notre-Dame Hospital, Montreal, Quebec, Canada; Department of Neuroradiology (O.N.), Paris-Descartes University, Centre Hospitalier Sainte-Anne, Paris, France; and Centre for Statistics in Medicine (D.G.A.), University of Oxford, Oxford, United Kingdom.

Please address correspondence to Jean Raymond, MD, CHUM-Notre-Dame Hospital, Interventional Neuroradiology, 1560 Sherbrooke East, Pavilion Simard, Room Z12909, Montreal, Quebec, Canada H2L 4M1; e-mail: jean.raymond@Umontreal.ca

DOI 10.3174/ajnr.A2442

dations, ill-advised subgroup analyses remain common: The prevalence of trial publications claiming at least 1 (statistically significant) subgroup effect has ranged from 25% to 60%.[10-14] In a large systematic review aiming to study the analysis, reporting, and claims of subgroup effects in a representative sample of recent RCTs, investigators showed that 45% of 139 RCTs reported subgroup effects for any outcome.[15] Advocates of subgroup analyses are alarmed by the risk of missing important differences in treatment effect, which could result in failure to detect important differences in heterogeneous populations.[16-19] Opponents describe subgroup analyses as "fishing expeditions" and "data dredging exercises."[20,21] Conventional subgroup analyses examine whether specific patient characteristics modify the effects of treatment, by considering each in turn. That approach leads to the danger of multiple comparisons. The analyses are underpowered, and they do not account for the fact that patients have multiple characteristics simultaneously that affect the likelihood of treatment benefit.[22]

An Example of Subgroup Analysis from the ISAT Trial

The ISAT was initiated in 1997 and aimed to recruit 2500 patients to achieve a 90% power at the 1% level of significance to detect a 25% relative reduction in the proportion of patients dependent or dead at 1 year. Recruitment was stopped in 2002 following a recommendation of the independent monitoring committee who judged that differences in the primary clinical outcome events between patients treated with clip ligation and those treated with coil embolization were too large for the trial to continue. They found that 190 of 801 (23.7%) patients allocated to endovascular treatment were dependent or dead at 1 year compared with 243 of 793 (30.6%) allocated to neurosurgical treatment ($P = .002$), a relative risk reduction of 23%. A second article from the ISAT group claimed generalizability, with results consistent in all prespecified subgroups.[9] However, another publication in 2008[3] concerned a post hoc (nonprespecified) subgroup analysis of 278 (13% of the entire ISAT population) patients 65 years of age or older at the time of the SAH. One can notice that the cutoff is now at 65 years, compared with 60 or 70 years in previous publications.[2,9] Investigators found that in good-grade elderly patients with SAH with small anterior circulation aneurysms, endovascular treatment led to better outcomes for internal carotid and posterior communicating artery aneurysms ($n = 134$, 6.3%; coiling, 72; clipping, 62), whereas patients with MCA aneurysms ($n = 37$, 1.7%; coiling, 22; clipping, 15) appeared to benefit from neurosurgical clipping rather than endovascular treatment ($P < .05$).

Problems with Subgroup Analyses

Several empiric studies that have evaluated how trialists conduct and report subgroup analyses all revealed several problems, including the study of an excessive number of variables and outcomes, the use of inappropriate statistical methods, and insufficient a priori specification of variables.[11-15,23] In a clinical trial, it is usual to collect detailed information on patient characteristics and the specific outcome measures. Researchers can perform many separate analyses in the hope that "something will turn up" that has a P value lower than .05. Conducting multiple tests is associated with a raised risk of false-positive results due to chance alone.[7] Clinicians have to keep in mind that that when a treatment is ineffective, there is still a probability of 5% of observing a significant effect ($P < .05$) due to chance. Suppose we split the study population into 20 mutually exclusive subgroups and examine the difference between the treatments in each group when, in fact, there is no true effect. The probability of at least 1 significant but false-positive result is 0.64.

When multiple subgroup analyses are conducted, 1 way to ensure that the overall chances of a false-positive result are no greater than 5% (0.05) is for each test to use a criterion of $0.05/n$, to assess statistical significance (the Bonferroni correction).[24] For example, if 20 tests are conducted, each should use $P = .0025$ as the threshold for significance.[24]

The power of chance to mislead is particularly high when investigators perform numerous post hoc subgroup analyses seeking statistical significance. The situation may be further complicated by the use of dichotomization or categorization of continuous variables according to various cutoff values (such as age categorized in decades[2,9] or as >65 years or <65 years),[3] a process that allows the production of various, sometimes diverging, results.[25] The reluctance of investigators to acknowledge that a specific hypothesis was post hoc can prevent readers from being cautious in interpreting these findings. These problems can only be prevented by requiring precise specification and publication of detailed protocols of clinical trials. Good practice for RCTs is to prespecify the research questions with precision, defining the primary outcome, the error rates, the sample size, and any subgroup of interest (preferably few) to minimize the risks of spurious effects and wrong conclusions.

The risks are not theoretic. Medical history is replete with examples of misleading findings, such as "aspirin reduced stroke risk in men but not in women"[26] or "β blockers are ineffective in patients with inferior (as opposed to other) myocardial infarctions,"[27] driven by over-reliance on subgroup results from otherwise well-designed RCTs. For this reason, many recommend that subgroup analysis should be seen as generating hypotheses for further testing.[22] A study comparing trial protocols with subsequent publications found that few prespecified subgroup analyses were published, and most of the published subgroup analyses had not been prespecified.[28] Discrepancies between the study protocol and the publication were found in all cases.[28]

The key question when examining subgroup differences is the following: If the true effect was the same in all patients, how likely was it that the differences occurred by chance alone. The wrong way is to test whether the effect was significant in each subgroup of interest, exploring the null hypothesis (no treatment effect) in each instance. A claim of subgroup effect is then made if a significant effect is observed in 1 subgroup but not in the others. In the example we have chosen, Ryttlefors et al[3] emphasized the superiority of clipping over coiling for MCA aneurysms (OR, 7.8; 95% CI, 1.4–43.1; $P = .02$) and the superiority of coiling over clipping for internal carotid artery–posterior communicating artery aneurysms (OR, 0.4; 95% CI, 0.2–1.0; $P = .04$) in patients older than 65 years, in effect comparisons of subgroups based on location within subgroups based on a dichotomy according to age, an analysis that raises numerous concerns.

The issue is not whether the treatment effect is significant

How credible is the subgroup finding?
Questions
Is the subgroup variable a characteristic specified at the time of the design of the study?
Was the correct direction of the subgroup effect (favorable or unfavorable to treatment for example) specified a priori?
Is the significant interaction effect independent of other potential subgroup effects?
Does the interaction test suggest a low likelihood that chance can explain the apparent subgroup effect?
Was the subgroup effect 1 of a small number of hypothesized effects tested? (The greater the number of hypotheses tested, the greater the number of subgroup effects one will discover by chance.)
Is the magnitude of the subgroup effect large?
Is the interaction consistent across closely related outcomes within the study?
Is the interaction consistent across studies?
Are the subgroup effects plausible? Is there a biologic rationale?

in 1 subgroup and not in another but rather whether the differences between subgroups can be readily explained by chance alone. A better way to conduct the analysis is to perform an interaction test that compares the treatment effect across complementary subgroups.[29,30] Another approach is to see whether the treatment effect is related to risk of the outcome as predicted by a multivariable risk-prediction tool.[22]

What Can Give Support to a Subgroup Finding?

Although many subgroup findings should not be trusted, there are clues that enable readers to give some credence to subgroup effects. These are summarized in the Table (inspired from published literature[15,20,31]). Subgroup effects gain credibility when a small number of plausible groups, perhaps identified in previous studies, has been prespecified in a detailed published trial protocol. The direction of the subgroup effect should been specified a priori.[31] The larger the difference is between the 2 treatments in particular subgroups, the more plausible it is that the difference is real, provided that the sample size is large enough. When sample sizes are small (such as for patients older than 65 years with ruptured MCA aneurysms in our example), one will see large differences in the apparent effect simply by chance. We are all more ready to believe an interaction if additional internal and external evidence sustains the hypothesis. Readers should look for an interaction consistent across closely related outcomes within the same study (internal confirmation). Evidence from intermediary outcomes (at 1 month, 6 months) is the strongest type. Readers should also look for external evidence: studies of different populations, observation of subgroup differences for similar interventions, and results of studies of intermediary or surrogate outcomes. Replication of the subgroup difference in other studies increases the credibility of the finding; and the extent to which a rigorous systematic review of the relevant literature finds such difference, again in the same subgroup, is a good index of credibility.[23]

Conclusions

Randomized trials provide the best evidence as to whether a treatment is, in general, beneficial. To use trial results with confidence in the treatment of future patients, we need reassurance that the same treatment benefited a diversity of patients, with varying prognostic factors of clinical interest that might have an impact on the treatment effect. However, if we allow ourselves to explore many ways of grouping patients by using numerous characteristics (and even their combinations), some discrepant heterogeneous "results" can nearly always be found. A prudent interpretation of trial results is to limit findings that will affect clinical decisions to overall treatment effects regarding primary end points that have been carefully planned, powered, and controlled for errors. Hence subgroup findings should generally be considered as just exploratory results. They can be given some credence when they have been limited to a small number of prespecified groups and when effects are large, consistent, duplicated in other studies, and clinically plausible. When subgroup effects are in the opposite direction of the overall results, the most prudent approach is to consider subgroup findings as hypotheses for another trial. Until then, the best estimator of the treatment effect for any subgroup is the overall treatment effect. Because the subgroup effects described by Ryttlefors et al[3] meet none of these criteria, their use in clinical decision-making is considered ill-advised.

References

1. Molyneux A, Kerr R, Stratton I, et al. **International Subarachnoid Aneurysm Trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: a randomised trial.** *Lancet* 2002;360:1267–74
2. Ryttlefors M, Enblad P, Kerr RS, et al. **International Subarachnoid Aneurysm Trial of neurosurgical clipping versus endovascular coiling: subgroup analysis of 278 elderly patients.** *Stroke* 2008;39:2720–26. Epub 2008 Jul 31
3. Furberg CD, Byington RP. **What do subgroup analyses reveal about differential response to beta-blocker therapy? The Beta-Blocker Heart Attack Trial experience.** *Circulation* 1983;67:I98–101
4. Fletcher J. **Subgroup analyses: how to avoid being misled.** *BMJ* 2007;335:96–97
5. Oxman AD, Guyatt GH. **A consumer's guide to subgroup analyses.** *Ann Intern Med* 1992;116:78–84
6. Schulz KF, Grimes DA. **Multiplicity in randomised trials. II. Subgroup and interim analyses.** *Lancet* 2005;365:1657–61
7. Yusuf S, Wittes J, Probstfield J, et al. **Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials.** *JAMA* 1991;266:93–98
8. Altman DG. **Critical methodology: clipping vs coiling of ruptured aneurysms.** In: *Proceedings of the 10th Congress of World Federation of Interventional and Therapeutic Neuroradiology*, Montreal, Quebec, Canada; June 29-July 3, 2009
9. Molyneux AJ, Kerr RS, Yu LM, et al. **International Subarachnoid Aneurysm Trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: A randomised comparison of effects on survival, dependency, seizures, rebleeding, subgroups, and aneurysm occlusion.** *Lancet* 2005;366:809–17
10. 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;317:426–32
11. Assmann SF, Pocock SJ, Enos LE, et al. **Subgroup analysis and other (mis)uses of baseline data in clinical trials.** *Lancet* 2000;355:1064–69
12. Bhandari M, Devereaux PJ, Li P, et al. **Misuse of baseline comparison tests and subgroup analyses in surgical trials.** *Clin Orthop Relat Res* 2006;447:247–51
13. Hernandez AV, Boersma E, Murray GD, et al. **Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading?** *Am Heart J* 2006;151:257–64
14. Wang R, Lagakos SW, Ware JH, et al. **Statistics in medicine: reporting of subgroup analyses in clinical trials.** *N Engl J Med* 2007;357:2189–94
15. Sun X, Briel M, Busse JW, et al. **Subgroup Analysis of Trials Is Rarely Easy (SATIRE): a study protocol for a systematic review to characterize the analysis, reporting, and claim of subgroup effects in randomized trials.** *Trials* 2009;10:101
16. Horwitz RI, Singer BH, Makuch RW, et al. **On reaching the tunnel at the end of the light.** *J Clin Epidemiol* 1997;50:753–55
17. Feinstein AR. **The problem of cogent subgroups: a clinicostatistical tragedy.** *J Clin Epidemiol* 1998;51:297–99
18. Goldman L, Feinstein AR. **Anticoagulants and myocardial infarction: the problems of pooling, drowning, and floating.** *Ann Intern Med* 1979;90:92–94

19. Horwitz RI. **Complexity and contradiction in clinical trial research.** *Am J Med* 1987;82:498–510
20. Guyatt G, Wyer P, Ioannidis J. **When to believe a subgroup analysis.** In: *User's Guide to the Medical Literature: A Manual for Evidence-Based Clinical Practice.* Chicago: American Medical Association; 2008:571–83
21. Altman DG. **Within-trial variation: a false trail?** *J Clin Epidemiol* 1998;51:301–03
22. Kent DM, Rothwell PM, Ioannidis JPA, et al. **Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal.** *Trials* 2010;11:85
23. Hernandez AV, Steyerberg EW, Taylor GS, et al. **Subgroup analysis and covariate adjustment in randomized clinical trials of traumatic brain injury: a systematic review**. *Neurosurgery*. 2005;57:1244–53, discussion 1244–53
24. Bland JM, Altman DG. **Multiple significance tests: the Bonferroni method.** *BMJ* 1995;310:170
25. Naggara O, Raymond J, Guilbert F, et al. **Critical methodology in neurovascular disease clinical research. I. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms.** *AJNR Am J Neuroradiol.* In press
26. Lee DC, Johnson RA, Bingham JB, et al. **Heart failure in outpatients: a randomized trial of digoxin versus placebo.** *N Engl J Med* 1982;306:699–705
27. **Reduction in mortality after myocardial infarction with long-term beta-adrenoceptor blockade: multicentre international study— supplementary report.** *BMJ* 1977;2:419–21
28. Chan AW, Hróbjartsson A, Jørgensen KJ, et al. **Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols.** *BMJ* 2008;337:a2299
29. Altman DG, Bland JM. **Interaction revisited: the difference between two estimates.** *BMJ* 2003;326:219
30. Lagakos SW. **The challenge of subgroup analyses: reporting without distorting.** *N Engl J Med* 2006;354:1667–69
31. Sun X, Briel M, Walter SD, et al. **Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses.** *BMJ* 2010;340:c117