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

**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.\(^2\) That finding was found to be consistent in most prespecified subgroups of patients.\(^3\) In a further article, the authors reported results for a subgroup of 278 patients older than 65 years of age.\(^4\) 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,”\(^5\) and many authors have cautioned against overinterpreting them.\(^6,7\)

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-
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%.\textsuperscript{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.\textsuperscript{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.\textsuperscript{16-19} Opponents describe subgroup analyses as "fishing expeditions" and "data dredging exercises."\textsuperscript{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.\textsuperscript{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.\textsuperscript{9} However, another publication in 2008\textsuperscript{8} concerned a post hoc (non-prespecified) 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.\textsuperscript{29} 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.\textsuperscript{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.\textsuperscript{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% ($P < .05$) is for each test to use a criterion of $P < .05/n$, to assess statistical significance (the Bonferroni correction).\textsuperscript{24} For example, if 20 tests are conducted, each should use $P = .0025$ as the threshold for significance.

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\textsuperscript{9} or as $>65$ years or $<65$ years),\textsuperscript{2} a process that allows the production of various, sometimes diverging, results.\textsuperscript{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"\textsuperscript{26} or "$\beta$ blockers are ineffective in patients with inferior (as opposed to other) myocardial infarctions."\textsuperscript{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.\textsuperscript{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.\textsuperscript{28} Discrepancies between the study protocol and the publication were found in all cases.\textsuperscript{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\textsuperscript{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?

References


