European Journal of Endocrinology, 2023, **189**, E1–E3
https://doi.org/10.1093/ejendo/lvad089
Advance access publication 1 August 2023
Methodological Editorial

Subgroup analyses in clinical research: too tempting?

Rolf H.H. Groenwold[1,2,*] and Olaf M. Dekkers[1,3]

[1]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, 2333 ZA, the Netherlands
[2]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, 2333 ZA, the Netherlands
[3]Department of Endocrinology, Leiden University Medical Center, Leiden, 2333 ZA, the Netherlands

***Corresponding author:** Department of Clinical Epidemiology, Leiden University Medical Center, Albinusdreef 2, Leiden, ZA 2333, the Netherlands.
Email: R.H.H.Groenwold@lumc.nl

Abstract

In many biomedical studies, subgroup analyses are performed to identify subgroups of patients in whom a treatment is most effective, or a risk factor has the largest effect. While both are referred to as subgroup analysis, it is important to distinguish between the estimation of effects within subgroups and the comparison of effects across subgroups. Both are discussed, and we outline the implications regarding sample size and statistical methods for estimation of effects. Also, the risk of false-positive findings—which potentially increases with subgroup analysis—is discussed, as well as the distinction between effect modification and interaction.

Keywords: effect modification, interaction, subgroup analysis, clinical research

Introduction

Humans differ in many respects, eg, regarding genetic make-up, body weight, co-morbidities, and lifestyle. Each of these factors could influence for example the effects of medical treatments or the extent to which a risk factor has an effect, which then can be investigated through subgroup analysis. Subgroup analyses are appealing because they could inform more personalized medical decisions: instead of knowing the average effect of a medical treatment (the so-called population average effect), subgroup analyses provide the opportunity to identify those subjects in whom the treatment is most effective or in whom the treatment adverse effects are more likely.

Here, we discuss several methodological aspects of subgroup analysis to inform researchers, readers, and reviewers about some key aspects of those analyses. While subgroup analyses may seem appealing, they come with the drawbacks of incorrect conclusions. We discuss subgroup effects in study of medical treatments, but the topics discussed also apply to studies of risk factors (etiology).

Estimation of subgroup effects

There is a difference between the estimation of effects within subgroups and the comparison of effects across subgroups (see Figure 1). The former requires stratification by the subgrouping variable and subsequent estimation of the treatment effect within each of the subgroups, eg, for women and men separately. For inference, in each of the subgroups, the estimated treatment effect and its corresponding 95% confidence interval (or *P*-value) is compared against a neutral effect (eg, absence of a treatment effect). The latter—a comparison across subgroups—also requires stratification by the subgrouping variable as first step, yet subsequently the treatment effects estimated in the subgroups are compared against each other. Preferably, this comparison is based on a statistical test comparing these treatment effects, for example a test of interaction (see Figure 1).

There are two poor men's solutions. Some researchers compare confidence intervals across subgroups, where non-overlapping confidence intervals are indicative of difference in treatment effects across subgroups. Note that this approach is too conservative (ie, has relatively low statistical power to detect differences in effect size across subgroups).[1] Secondly, some researchers may compare the significance of two (or more) treatment effects and conclude that there is a difference in effects if one is significant and the other not. This approach is simply incorrect.

In studies of medical treatments, subgroup analysis is almost always preceded by estimation of the treatment effect in the entire study group, yielding a so-called average treatment effect. When the average treatment effect is neutral (ie, there appears to be no relevant treatment effect), performing subgroup analyses implies that one expects a possible positive treatment effect in one subgroup and thus a negative treatment effect in another subgroup (otherwise the average treatment effect could not be neutral). This is one reason why—in case of a neutral average treatment effect—subgroup analysis is generally not likely to provide meaningful results.

When interest lies in treatment effects within subgroups, then—ideally—the sample size of each subgroup should be such that the power to detect a subgroup-treatment effect (should it exist) is sufficiently large. Practically, this could mean that a regular sample size calculation is performed, the results of which indicates the minimum number of participants in the smallest subgroup. The other subgroup(s) are larger and therefore are expected to have higher power (if the expected effect size is approximately similar). For example, in an RCT assessing the treatment effect of antidiabetic medication, researchers may be interested whether the effect exists in both men and women. In this case, the power should be sufficient in both groups. When interest lies in the difference in

Received: May 31, 2023. **Editorial Decision:** June 6, 2023. **Accepted:** June 6, 2023

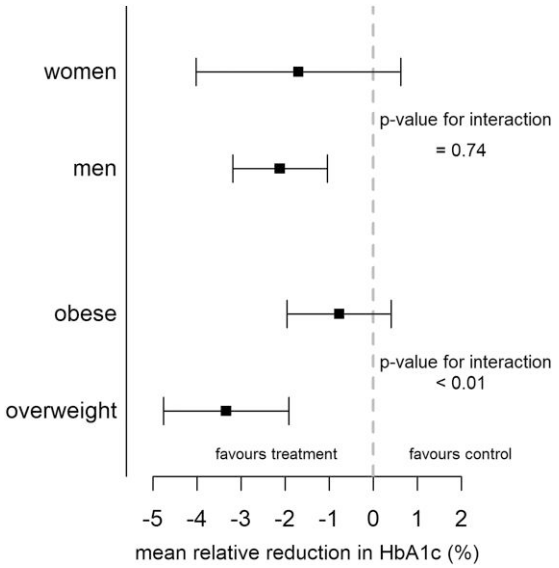

Figure 1. Forest plot of relative reduction of HbA1c of within subgroups of a hypothetical study of a glucose-lowering treatment vs placebo. Figure presents point estimate (box) and corresponding 95% confidence interval (bars) of the effect of glucose-lowering treatment on HbA1c levels. The dashed vertical line indicates a neutral effect (no difference in glucose-lowering effect). In the analysis that is stratified by sex, the treatment effect is statistically significant in men, but not in women. However, a test of interaction indicates that the treatment effect does not differ between subgroups ($P = .74$). In the analysis stratified by BMI, the treatment effect is statistically significant in overweight patients but not in obese patients. A test of interaction indicates that the treatment effect differs between subgroups ($P < .01$), despite the overlapping 95% confidence intervals.

treatment effect across subgroups, it is that difference that determines samples size and hence the expected difference in effect across subgroups needs to be clearly articulated in study protocol and report. As a rule of thumb, one could say that at least four times as many participants need to be included as compared to a study investigating an average treatment effect.[2]

False-positive results

With an increasing number of statistical tests, the probability of at least one false-positive result increases, a problem also known as multiple testing.[3] Subgroup analyses are no exception to this rule: with an increasing number of subgroup analyses, the probability of at least one false-positive subgroup effect increases, even more so when subgroup classification is data driven (eg, subgroup analysis after identifying most distinctive subgroups). Methods exist to correct for multiple testing, such that the overall probability of false-positive results is controlled. For such methods to behave properly, it is important to know the total number of subgroup analyses that has been performed. Pre-specification of subgroup analyses does not protect against false-positive results, yet it provides insight into the number of subgroup analyses to be performed, and thus the extent to which correction for multiple testing is warranted. Also, pre-specification of subgroup analysis, including a rationale and anticipated direction may benefit the credibility of found subgroup effects.[4]

Obviously, subgroup analyses may also lead to false-negative results. One example of this is a subgroup analysis of the ISIS-2 trial, which investigated the effects of aspirin and streptokinase in patients suspected of myocardial infarction.[5] While aspirin was found to be highly effective in reducing mortality risk, it was found ineffective in patients born under the astrological signs of Libra or Gemini. The only reasonable explanation for this observation seems to be chance, ie, a false-negative finding.

Often, studies of medical treatments are underpowered for subgroup analyses. (This is understandable given that trials are very expensive and adequately powered subgroup analyses will increase the sample size, which substantially increases the costs.) In that case, subgroup analyses may still be performed, but these are indicated to be "exploratory". Nevertheless, when striking results are found, it might be tempting to provide post hoc explanations for the observed subgroup effect. Given the general risk of false-positive subgroup effects, researchers should be reluctant to claim a subgroup effect. What is more, given the subgroup analysis is likely underpowered, finding a significant subgroup effects, suggests that the magnitude of the effect is likely overestimated.

Subgroups, effect modification, and interaction

Subgroup analysis is a general term that encompasses effects within subgroups as well as differences in effects across subgroups. Regarding differences in effects across subgroups, a further distinction can be made between effect modification and interaction.[6] While an analysis of effect modification describes solely the difference in treatment effects across subgroups, interaction analysis aims to understand and explain those differences causally (see below). In many books about statistics, the distinction between effect modification and interaction is not made and both are referred to as (statistical) interaction.

Consider a randomised controlled trial (RCT) in diabetes, and suppose the blood glucose-lowering effect of the investigated treatment is found in overweight patients, but not in obese. It is then a fair conclusion that there is *effect modification* by body mass index (BMI), since effect modification pertains to the situation where the effect (of a treatment) differs across subgroups. The subgrouping variable characterizes the subgroup, but no claims are made about the effect of the subgrouping variable itself. But mind that even in an RCT, comparing effects of subgrouping variables is "observational" in nature.[7] This means that, whereas the overall (or subgroup) treatment effect can be estimated unconfoundedly, the comparison between subgroups can be confounded. If there is, based on RCT data, evidence that the treatment is effective in overweight patients only, this need not necessarily indicate a formal *interaction* by BMI. It may be due to underlying differences between overweight and obese patient, for example in genetic profile, comorbidities, or co-medication use. Investigation of interaction requires either a more rigorous study design (eg, factorial trial design) or more assumption to be made (eg, about confounders of the relation between the subgrouping variable and the outcome being measured).

Conclusion

We discussed the topic of subgroup analysis, which refers to estimation of effects within subgroups as well as comparison of effects across subgroups. Both are prone to false-positive findings, particularly when decisions about subgroup analyses are post hoc and/or data driven. Many subgroup analyses

investigate effect modification, whereas interaction requires more from either the design of the study or the underlying assumptions. Researchers should report clearly and transparently about subgroup analyses, including their rationale, the number of analyses, whether pre-specified, and their interpretation, in order to allow readers a fair assessment of its results.

Funding

None declared.

Conflicts of interest: R.G. reports no conflicts of interest. O.M.D. is a Deputy Editor for European Journal of Endocrinology. He was not involved in the review or editorial process for this article, on which he is listed as an author.

References

1. Knol MJ, Pestman WR, Grobbee DE. The (mis)use of overlap of confidence intervals to assess effect modification. *Eur J Epidemiol*. 2011;26(4):253-254. https://doi.org/10.1007/s10654-011-9563-8
2. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol*. 2004;57(3):229-236. https://doi.org/10.1016/j.jclinepi.2003.08.009
3. Groenwold RHH, Goeman JJ, Cessie SL, Dekkers OM. Multiple testing: when is many too much? *Eur J Endocrinol*. 2021;184(3):E11-E14. https://doi.org/10.1530/EJE-20-1375
4. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ*. 2010;340:c117. [Published online on March 30, 2010]. https://doi.org/10.1136/bmj.c117
5. 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):349-360. https://doi.org/10.1016/S0140-6736(88)92833-4
6. VanderWeele TJ. On the distinction between interaction and effect modification. *Epidemiology*. 2009;20(6):863-871. https://doi.org/10.1097/EDE.0b013e3181ba333c
7. Groenwold RH, Donders AR, van der Heijden GJ, Hoes AW, Rovers MM. Confounding of subgroup analyses in randomized data. *Arch Intern Med*. 2009;169(16):1532-1534. https://doi.org/10.1001/archinternmed.2009.250