Interactive CardioVascular and Thoracic Surgery (2020) 1–7
doi:10.1093/icvts/ivaa042

Cite this article as: Milojevic M, Nikolic A, Jüni P, Head SJ. A statistical primer on subgroup analyses. Interact CardioVasc Thorac Surg 2020; doi:10.1093/icvts/ivaa042.

A statistical primer on subgroup analyses

Milan Milojevic[a,b,*], Aleksandar Nikolic[c], Peter Jüni[d] and Stuart J. Head[a]

[a] Department of Cardiothoracic Surgery, Erasmus University Medical Center, Rotterdam, Netherlands
[b] Department of Cardiac Surgery and Cardiovascular Research, Dedinje Cardiovascular Institute, Belgrade, Serbia
[c] Department of Cardiac Surgery, Acibadem Sistina Hospital, Skopje, North Macedonia
[d] Applied Health Research Centre, Li Ka Shing Knowledge Institute of St. Michael's Hospital, Department of Medicine, University of Toronto, Toronto, ON, Canada

[*] Corresponding author. Department of Cardiothoracic Surgery, Erasmus University Medical Center, Dr Molewaterplein 40, 3015 GD, PO Box 2040, 3000 CA Rotterdam, Netherlands. Tel: +31-10-7035784; fax: +31-10-7033993; e-mail: m.milojevic@erasmusmc.nl (M. Milojevic).

Received 15 November 2019; received in revised form 23 January 2020; accepted 29 January 2020

Key question

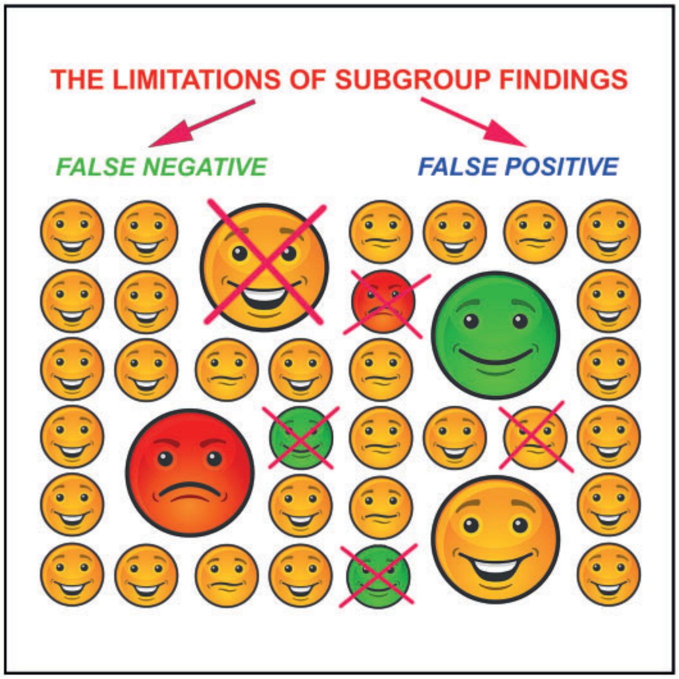
What are essential standards for conducting subgroup analyses in cardiovascular medicine?

Key finding(s)

Subgroup analyses require advanced planning, including adequate statistical power and high reporting standards.

Take-home message

A well-conducted subgroup analysis provides more evidence for personalized medicine and establishes hypotheses for future research.



Summary

Resources for clinical research are limited. With increasing demand for patient-centred care, which is growing into an integral component of modern medicine, studying outcomes of patients with specific clinical characteristics is becoming increasingly important. Given the high cost of clinical trials and the time it takes to complete an investigation, it has become compulsory for investigators to assess not only treatment effects between the main randomized groups but also to try to identify clinically relevant subgroups that may particularly benefit from specific treatments. Publications of subgroup analyses turned out to be prevalent, and more importantly, these findings play a significant role in strategic planning and decision-making processes. Therefore, raising awareness among clinicians about the concepts and values of subgroup analysis is an aspect of improving patient outcomes. In this statistical primer, we give a broad introduction to the topic of subgroup analysis in scientific research. We furthermore discuss the concept of subgroup analysis; the motivation for assessing subgroups; the types of subgroup analyses and the paradigm of hypothesis-generating research; the proper statistical methods for the examination of subgroup effects; and the optimal approach for interpretation of results. Finally, this review establishes the comprehensive users' guide for analysing and reporting subgroup studies on a point-by-point basis, using real-world examples that may help readers to gain experience to pursue their own subgroup analyses or interpret those of others.

Keywords: Statistics • Subgroup analysis • Clinical trials • Stratification • Multiple testing • Tests of interaction

ABBREVIATIONS

CABG	Coronary artery bypass grafting
DES	Drug-eluting stent
LMD	Left main disease
PCI	Percutaneous coronary intervention
RCT	Randomized controlled trial

INTRODUCTION

Well-designed and rigorously conducted randomized controlled trials (RCTs) are viewed as the ultimate scientific tool to assess the safety and efficacy of treatments [1]. Among researchers, clinicians and regulatory agencies, results from dedicated RCTs are generally considered necessary to enhance treatment strategies, patient safety and quality of care. The process of assigning study participants to intervention by chance should guarantee that treatment groups are well-balanced in terms of known and unknown confounding factors, thus ensuring that any difference in outcome can be explained only by the treatment effect. However, their conduct is significantly limited by ethical and practical concerns including the long period of recruitment and/or follow-up and the typically high costs. These concerns play a particularly relevant role in phase IV trials, which involve a more substantial number of patients because the goal of the investigators is to establish the gold standard of care for daily practice. Because of the time and resources invested in clinical trials and because a large amount of data on each recruited patient is often collected [2], the investigators attempt to derive as much information as possible from RCTs. Performing subgroup analyses and publishing specific subgroup papers have become popular processes done to increase the impact of an RCT. It has, therefore, become essential to discuss the benefits and challenges of subgroup analyses and to establish the guiding principles for scientific reporting. Hence, in this statistical primer, the key statistical issues that are relevant for high-quality standards to report subgroup analyses of RCTs are discussed further. Of note, many of these aspects can also be applied to observational studies.

(MIS)USE OF SUBGROUP ANALYSES

Subgroup analyses are necessary to evaluate whether the overall treatment effect in a trial is consistent across patients with different risk profiles. Beyond the vital role of identifying significant treatment heterogeneity across different subsets of patients, subgroup analyses may also serve as an outstanding source of information for generating hypotheses for future research and providing insights for personalized medicine [3]. Overall, subgroup analyses carry more weight than retrospective analyses and, if positive, are often used to support the hypotheses for clinical trials. However, subgroup analyses have also been misused to 'save the trial' by identifying a patient subgroup that may benefit from a particular treatment despite the trial reporting a neutral primary end point result. Indeed, only a minority of trials in cardiovascular medicine have sufficient statistical power to detect a treatment effect among subgroups, and the results of many of these analyses have been misleading and overstated in the scientific literature [4, 5]. Almost 2 decades ago, Sleight [6] from Oxford University published numerous examples of wrong interpretations of subgroup analyses that caused severe harm to patients. They found, for example, that restricting the use of thrombolytic and β-blockade therapies only to patients with anterior myocardial infarctions were incredibly harmful, showing that those treatments were beneficial also for myocardial infarctions in other anatomical locations. This editorial should remind physicians why subgroup analyses do not always need to be accurate and highlight potential reasons for low statistical power to detect treatment effects where their true statistical significance exists.

Many systematic reviews and meta-analyses combine patient data from subgroups of patients to estimate the treatment effect for clinically relevant subgroups of patients. Despite the widely acknowledged benefits of meta-analyses of data from individual patients compared to conventional meta-analyses [7, 8] in cardiovascular medicine, data sharing receives insufficient attention from clinical trialists. This attitude may be partially justified by the numerous legal and technical barriers that remain and that are difficult to overcome, but they must be addressed and removed as quickly as possible. A brilliant example is the field of genetics in which the academic community has established an overall agreement to share all available data among investigators for research activities in the same area [9]. In light of current circumstances, the results of subgroup analyses serve as an essential source of information for conventional meta-analyses; hence well-conducted studies can minimize future publication bias.

DEFINING SUBGROUPS

A subgroup can be described as any subset of the overall studied population that is determined by the absence or presence of 1 or more descriptive factors [10]. Numerous intrinsic and extrinsic factors may be used to categorize patients into subgroups, including demographic characteristics, comorbidities, severity or type of disease, clinical presentation or procedural aspects. These factors can be further classified as dichotomous (e.g. men and women), categorical (e.g. different geographical locations), ranked categorical [e.g. low (0–22), intermediate (23–32) and high (≥33) SYNTAX score tertiles] or continuous (e.g. age).

Demographic characteristics, such as sex, are most often naturally defined. Other categorizations depend on additional measures that may potentially lead to a risk of misclassification. There should be a clear justification for the choice of combining patients into subgroups and for the use of cut-off points for continuous variables. Ideally, subgroups should be defined in the trial protocol, with definitions based on clinical relevance or expert consensus definitions. However, in the case of continuous variables, non-linear associations can be explored with splines [11].

The appropriate distinction between the subgroups relative to the overall studied population is of ultimate importance during the study design because any mistake in the definition of a subset can lead to the study not being accepted or, even worse, to the study results being cited and used in daily practice and subsequently retracted by a journal [12].

CONSIDERATIONS FOR PERFORMING SUBGROUP ANALYSES

Prespecified and *post hoc* subgroup analyses

Clinical trial participants are naturally heterogeneous due to individual patient risk profiles such as age, sex, the severity of a disease or additional comorbidities that are known to be associated with prognosis. Subgroup analyses are performed to assess the consistency of treatment effects measured on a relative scale, such as risk ratios, odds ratios or hazard ratios, or, in the case of binary clinical outcomes, across different subgroups. Studies of treatment effects among specific subsets of patients need to be planned during the design phase of an RCT. Such prespecified subgroup analyses are defined and documented in the study protocol before any examination of data. To give more strength to prespecified subgroup analyses, stratified randomization can be used to ensure a balanced distribution of characteristics among subgroups; for example, randomization between coronary artery bypass grafting (CABG) and percutaneous coronary intervention (PCI) can be stratified according to the status of diabetes mellitus, because this variable is known to impact the comparative effectiveness of treatments. However, the lack of stratified randomization does not a priori render subgroup analyses invalid.

In a recent investigation of published study protocols, a majority of cardiovascular trials reported specific articles on subgroup analyses, but only about 20% of these analyses were performed on a prespecified subgroup [13]. All other analyses were performed on *post hoc* subgroups, i.e. they were not defined in the protocol before the investigators saw the data. This type of analysis is particularly challenging because there is a high plausibility that a hypothesis was established after the statistical analyses were performed. *Post hoc* subgroup analyses are sometimes referred to as a ‘fishing expedition’, where investigators are looking for significant results.

Hypothesis-generating subgroup analyses. Although prespecified subgroup analyses are more trustworthy than *post hoc* analyses, clinical investigators should avoid performing a large number of subgroup analyses, whether prespecified or *post hoc*. Both prespecified and *post hoc* subgroup analyses may have a limited ability to inform individual treatment decision-making because of false positive (type I error) results due to multiple testing or false negative (type II error) results because of inadequate sample size and statistical power. Results from a subgroup analysis should not be used to justify a change in clinical practice if the results are not prespecified and do not have enough statistical power to detect a clinically meaningful difference. Accordingly, the results of subgroup analyses should be viewed with caution and generally used for generating hypotheses for future clinical trials.

End points. Clinical trial end points can be classified as primary, secondary or exploratory. The primary end point is the main study outcome and represents the parameter for which the study sample size is determined. Secondary or explanatory end points are those that are used to provide additional clinical characterization of treatment effects. Importantly, the majority of clinical studies lack sufficient power for these end points. Therefore, emphasizing secondary end points carries an extremely high risk of reporting false-positive conclusions. Generally, the judgement of subgroup effects should be based on the primary end point results, whereas secondary end points may be used in rare situations, for example when a strong signal of harm in a subgroup, such as a markedly higher risk of death, outweighs the risk of false positive findings.

Example of a subgroup analysis. An example of where a subgroup analysis was used for the hypothesis of a new RCT comes from the Synergy between Percutaneous Coronary Intervention with TAXUS and Cardiac Surgery (SYNTAX) trial. The trial was designed to examine if PCI with first-generation drug-eluting stents (DESs) was non-inferior to CABG for the 1-year incidence of major adverse cardiac and cerebrovascular events among 1800 patients with *de novo* 3-vessel disease or left main disease (LMD) [14]. The overall 1-year results demonstrated that PCI was not non-inferior to CABG [14]. Importantly, however, subgroup analyses according to (i) the presence of LMD ($n = 705$) and 3-vessel disease without LMD ($n = 1095$) were prespecified to reach the desirable statistical power of 80% for major adverse cardiac and cerebrovascular events; and (ii) the SYNTAX score was determined *post hoc* to determine the impact of coronary lesion complexity on the comparison between PCI and CABG. Whereas CABG was clearly superior to PCI in patients with 3-vessel disease, patients with LMD did well with PCI versus CABG, particularly those with a low or intermediate SYNTAX score (0–32). This result generated the hypothesis that PCI was non-inferior to CABG in patients with LMD and a SYNTAX score $\le$32. The Evaluation of XIENCE versus Coronary Artery Bypass Surgery for Effectiveness of Left Main Revascularization (EXCEL) trial subsequently tested the hypothesis and compared PCI with everolimus-eluting stents (XIENCE Family Stent System, Abbott Vascular, Santa Clara, CA, USA) with CABG among 1905 such patients [15].

Multiple testing

An alpha of 0.05 implies that there is a 5% probability of claiming statistical significance when there is no actual difference between the analysed groups. However, this interpretation is overoptimistic in most cases [16]. Even if the interpretation was always appropriate, if 20 independent statistical tests are performed with an alpha of 0.05, there is a 64% $[1 - (1 - 0.05)^{20}]$ chance of observing at least 1 significant result (Table 1). In the cardiovascular literature, the total number of analyses reported is almost always higher than 20; therefore, the probability that published findings are likely to be false positive is significant (Fig. 1). Crucially, false positive results may cause serious medical, economic and legal issues, especially if the study results are published in journals that have substantial credibility among health care professionals. This specific issue has received increasing attention over the last decades, but there is still no consensus on how to deal with multiple comparisons [18]. The Bonferroni correction and Benjamini–Hochberg approaches are most commonly used to reduce the number of false-positive results. They do not address the fundamental problems of significance testing at an alpha level of 0.05 in the presence of underpowered tests and the potentially low prevalence of real between-group differences [16]. They may increase the number of false negative results at the same time but are frequently performed to address the overall false positive rate of multiple tests at least partially.

Test for interaction

Moving beyond the question of whether a statistical analysis should include an adjustment for multiple testings, a proper analytical approach for the examination of subgroup effects must include a test for treatment by subgroup interaction [19]. This statistical approach is the most effective tool to overcome the concerns of false positive findings. It determines whether there is a significant difference in the treatment effect measured on a relative scale according to a specific characteristic. This approach is opposed to the 'main effect', which is the effect of a single independent variable, in trials of the treatment allocation, on the dependent variable. The types of interactions are divided into quantitative and qualitative for dichotomous subgroups. As shown in Fig. 2, a quantitative interaction implicates the variation in the magnitude of treatment effects whereas a qualitative interaction implies a change in the direction of the treatment effects among different subgroups. Qualitative interactions also include the only situation in which an interaction test is significant but in which there is no difference in the main results of the individual subsets. In general, quantitative interactions are more likely to be genuine than interactions that are qualitative.

Table 1: Testing multiple hypotheses

Number of analyses (*N*)	The probability of finding one or more false positive results by chance 100 - (1 - the significance level)^{N} (suppose $\alpha = 0.05$) (%)	Expected number of false positive findings
1	5	0.05
2	10	0.1
4	23	0.2
10	40	0.5
20	64	1
40	92	2
100	99	5

Expected probability of finding one or more false positive and expected number of false positives under the optimistic assumption that 50% of tested differences are real and the power of each performed analysis is 95%. False positive rates will increase dramatically with decreasing power and decreasing probability that a tested difference is real [16].

Due to the nature of interactions, these analyses can only be explored if both groups of a subgroup analysis are reported. However, this is not always the case. In a comparison of bare-metal stents (BMSs) versus DESs, Stefanini *et al*. [20] reported that outcomes favoured DES over BMS concerning safety and efficacy among women, on the basis of the individual *P*-values of the analyses. However, because no men were included in the study, it is unclear whether this treatment effect associated with DES is similar in men and whether there is any difference in the direction and magnitude of the treatment effect according to sex.

The presence of a positive test result for a treatment by subgroup interaction, as defined by a significant *P*-value for interaction at a 2-sided alpha of 0.05 or smaller, is particularly relevant for the interpretation of subgroup results. If the test results are negative, one must assume that the difference in the treatment

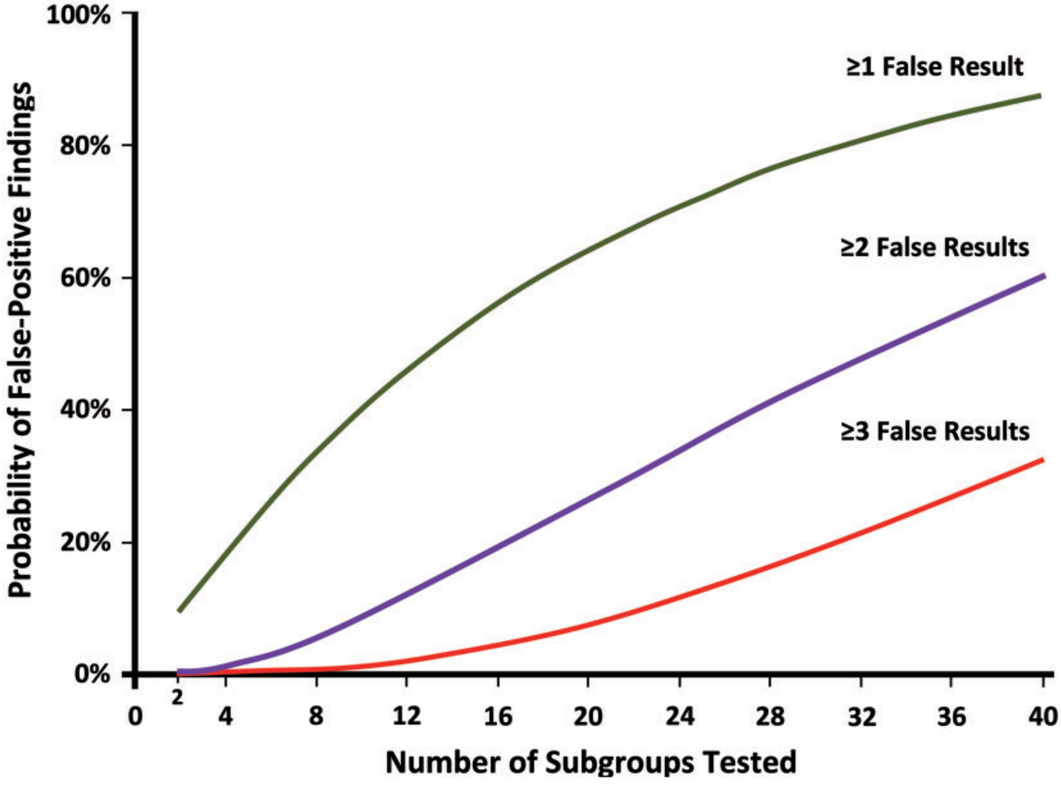


Figure 1: The risk of false-positive findings after performing multiple subgroup analyses (adapted from an illustration by Lagakos [17]).

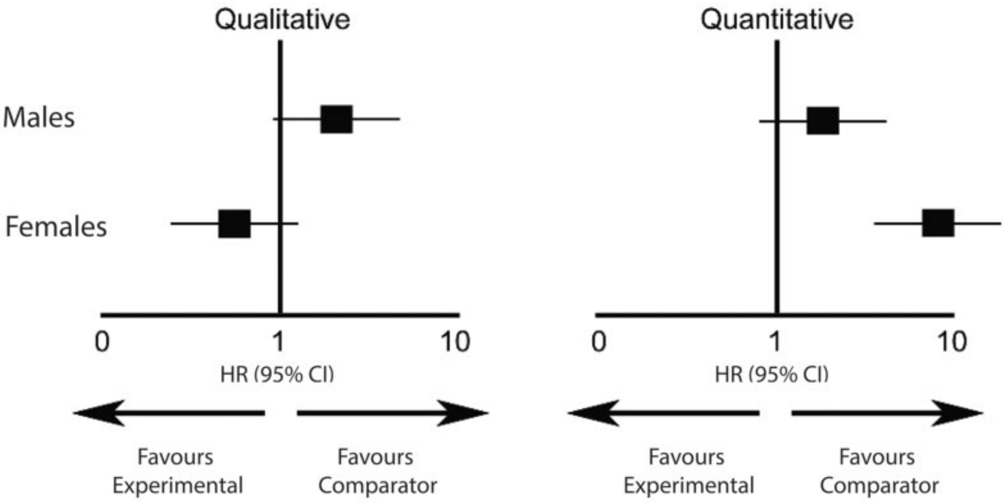
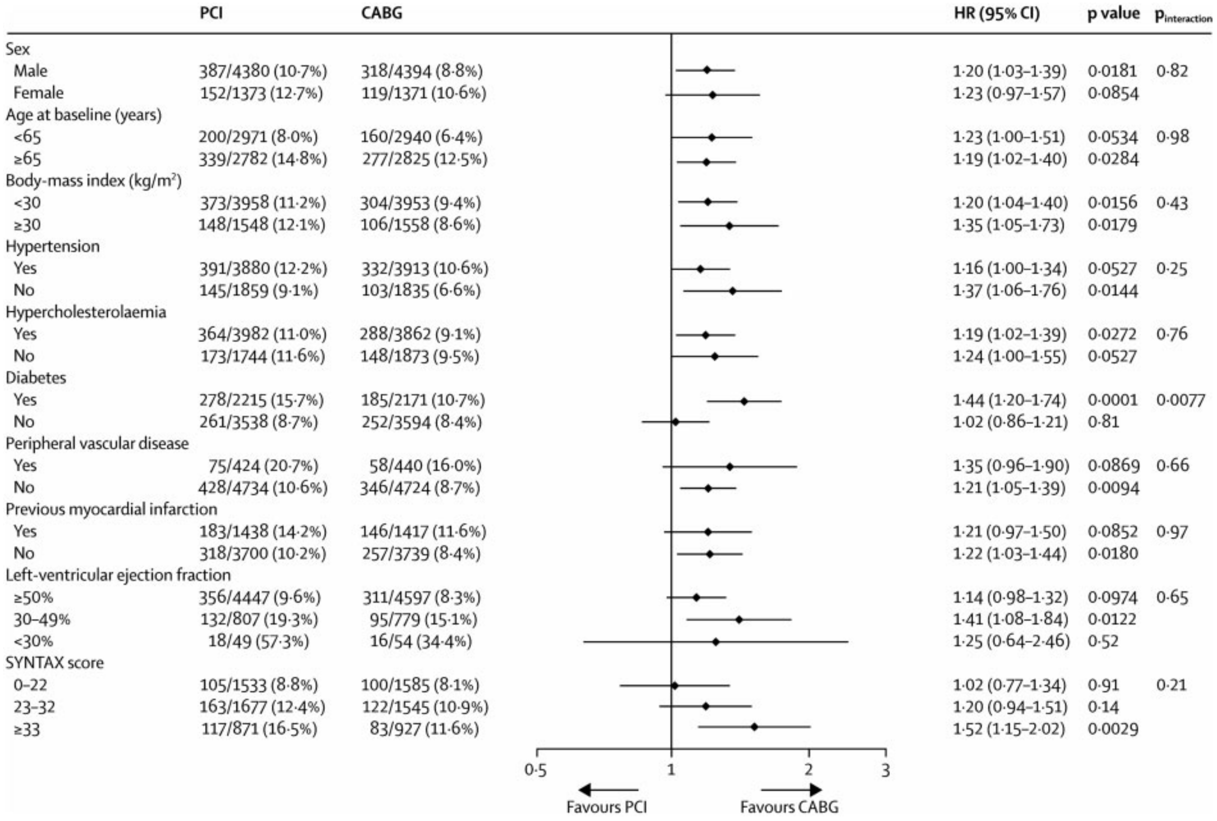


Figure 2: Different types of treatment-variable interactions. CI: confidence interval; HR: hazard ratio.



	PCI	CABG	HR (95% CI)	p value	$P_{interaction}$
Sex					
Male	387/4380 (10·7%)	318/4394 (8·8%)	1·20 (1·03–1·39)	0·0181	0·82
Female	152/1373 (12·7%)	119/1371 (10·6%)	1·23 (0·97–1·57)	0·0854	
Age at baseline (years)					
<65	200/2971 (8·0%)	160/2940 (6·4%)	1·23 (1·00–1·51)	0·0534	0·98
≥65	339/2782 (14·8%)	277/2825 (12·5%)	1·19 (1·02–1·40)	0·0284	
Body-mass index (kg/m²)					
<30	373/3958 (11·2%)	304/3953 (9·4%)	1·20 (1·04–1·40)	0·0156	0·43
≥30	148/1548 (12·1%)	106/1558 (8·6%)	1·35 (1·05–1·73)	0·0179	
Hypertension					
Yes	391/3880 (12·2%)	332/3913 (10·6%)	1·16 (1·00–1·34)	0·0527	0·25
No	145/1859 (9·1%)	103/1835 (6·6%)	1·37 (1·06–1·76)	0·0144	
Hypercholesterolaemia					
Yes	364/3982 (11·0%)	288/3862 (9·1%)	1·19 (1·02–1·39)	0·0272	0·76
No	173/1744 (11·6%)	148/1873 (9·5%)	1·24 (1·00–1·55)	0·0527	
Diabetes					
Yes	278/2215 (15·7%)	185/2171 (10·7%)	1·44 (1·20–1·74)	0·0001	0·0077
No	261/3538 (8·7%)	252/3594 (8·4%)	1·02 (0·86–1·21)	0·81	
Peripheral vascular disease					
Yes	75/424 (20·7%)	58/440 (16·0%)	1·35 (0·96–1·90)	0·0869	0·66
No	428/4734 (10·6%)	346/4724 (8·7%)	1·21 (1·05–1·39)	0·0094	
Previous myocardial infarction					
Yes	183/1438 (14·2%)	146/1417 (11·6%)	1·21 (0·97–1·50)	0·0852	0·97
No	318/3700 (10·2%)	257/3739 (8·4%)	1·22 (1·03–1·44)	0·0180	
Left-ventricular ejection fraction					
≥50%	356/4447 (9·6%)	311/4597 (8·3%)	1·14 (0·98–1·32)	0·0974	0·65
30–49%	132/807 (19·3%)	95/779 (15·1%)	1·41 (1·08–1·84)	0·0122	
<30%	18/49 (57·3%)	16/54 (34·4%)	1·25 (0·64–2·46)	0·52	
SYNTAX score					
0–22	105/1533 (8·8%)	100/1585 (8·1%)	1·02 (0·77–1·34)	0·91	0·21
23–32	163/1677 (12·4%)	122/1545 (10·9%)	1·20 (0·94–1·51)	0·14	
≥33	117/871 (16·5%)	83/927 (11·6%)	1·52 (1·15–2·02)	0·0029	

Figure 3: Caterpillar plot of subgroup analyses of 5-year mortality rates after CABG and PCI obtained for the pooled analysis of individual patient data from 11 clinical trials. Reprinted with permission from Elsevier [23]. CABG: coronary artery bypass grafting; CI: confidence interval; HR: hazard ratio; PCI: percutaneous coronary intervention.

effect between subgroups occurred by chance alone. In this situation, the main effect observed in the trial overall needs to be considered as the most representative for all subgroups and no subgroup-specific treatment effects can be claimed. Positive test results for an interaction provide some evidence for a difference in treatment effects between subgroups, but the caveats discussed above regarding false positives need to be taken into account. *P*-values for the interaction in the study report typically use the same threshold of significance as conventional analyses. The sole reporting of a *P*-value for an interaction is not adequate and needs to be complemented by treatment effects with the corresponding confidence intervals observed in each of the

Table 2: Critical points for reporting and interpreting subgroup analytic data

The study design:

- Is the subgroup analysis prespecified in the study protocol or designed *post hoc*?
- Are both the subgroup of interest and the complementary subgroup included?
- Is the subgroup analysis powerful enough for the proposed end points?
- Is the subgroup variable a stratification factor during the randomization?
- How is the subgroup defined?
- Does the subgroup analysis have a prior statistical power >50% to detect subgroup effect?
- Is the subgroup analysis restricted to groups >35% of the original cohort?

Statistical analysis:

- Is the intention-to-treat principle being used in the subgroup analysis?
- Is the test of interaction applied?
- Is there any adjustment for multiple testing?
- Is the analysis adjusted for any variety of critical prognostic factors among subgroups?

Reporting results:

- Are the differences in baseline characteristics presented (detailed description may include supplemental material)?
- Does the results section focus mainly on the primary end point(s)?
- Are event numbers and denominators presented together with a test of statistical significance and interaction terms?
- Is the graphical presentation of the subgroup effects included?

Discussion:

- Is the main emphasis placed on the primary outcome of the study?
- Is the subgroup effect consistent or conflicting with evidence from previous related studies?
- Is there any indirect evidence to support the differential effects?
- Are the limitations of the study emphasized correctly?
- Is the hypothesis-generating study design emphasized appropriately?
- Is the study conclusion based on the primary end point results?

subgroups. A graphical presentation using a caterpillar plot is desirable.

REPORTING SUBGROUP ANALYSES

The pattern of inadequate reporting of subgroup analyses was recognized almost 3 decades ago [21]. Despite numerous published editorials [10, 19, 22], there has been no relevant improvement in the reporting of subgroup effects among publications between 2007 and 2014 in the top 5 medical journals: *The New England Journal of Medicine; The Lancet; JAMA: The Journal of American Medical Association; Annals of Internal Medicine;* and *BMJ: The British Medical Journal* [5]. Of the 270 investigated subgroup analyses, roughly two-thirds did not apply any formal test for heterogeneity or interaction across subgroups. Furthermore, the authors found that the numbers of subgroup analyses using appropriate methods decreased from 77% in 2007 to 63% in 2014. Findings from large numbers of subgroup analyses may, therefore, be misleading.

Done correctly, the results of subgroup analyses should be reported in detail as follows: (i) the total number of investigated patients; (ii) the total number of events also expressed by percentages; (ii) the hazard ratio (or another relative measure of treatment effect) and the corresponding 95% confidence intervals; (iv) the *P*-value for treatment by subgroup interaction; and (v) a caterpillar plot as a graphical illustration of the treatment effect (Fig. 3).

However, a study report must also include specific information on the assessment of the validity and reliability of a subgroup finding. Thus, critical considerations (checklist) for conducting and interpreting a subgroup analysis are given in Table 2. Although the randomization process is expected to ensure the balance of risk factors between the overall treatment groups, it is less likely that subgroups will remain balanced if subgroups are small, unless randomization has been stratified according to the particular subgroup characteristic. A significant difference in a potent prognostic factor could cause confounding, which can be addressed by tests for interaction adjusted or stratified by this confounding factor.

The acknowledgment of a study's limitations provides the authors an opportunity to demonstrate critical thinking concerning the methods selected for investigation of the research problem and helps to convince the editorial board, the reviewers and potential readers that the study was conducted according to the highest scientific standards. Therefore, it is essential to recognize the 'hypothesis-generating' nature of subgroup analyses and acknowledge the limitations of the analyses.

CONCLUSIONS

Subgroup analyses are performed in the majority of clinical trials. Unfortunately, reporting of subgroup analyses is often inadequate. Misinterpretation of subgroup results may lead to suboptimal treatment and could mislead subsequent research. However, if subgroup analyses are well-designed and adequately analysed, the results may be crucial for informing physicians about treatment by subgroup interactions. Findings of *post hoc* subgroup analyses can be hypothesis-generating and assist in performing a dedicated, powered, confirmatory clinical trial. There remains a clear opportunity for significant improvements in the design and reporting of subgroup analyses. Strategies for improving the current practice of subgroup analyses require better education and more precise reporting standards. This article provides the checklist needed for the study design, statistical analysis and reporting of a subgroup study, and in addition, may serve as a vital tool in critical appraisal of scientific papers.

Conflict of interest: Stuart J. Head is an employee of Medtronic, outside the submitted work. All other authors declared no conflict of interest.

Author contributions

Milan Milojevic: Conceptualization; Writing—original draft. **Aleksandar Nikolic:** Writing—review & editing. **Peter Jüni:** Writing—review & editing. **Stuart J. Head:** Methodology; Writing—original draft.

REFERENCES

[1] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J *et al.* GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011; 64:401–6.

[2] Calvo G, McMurray JJV, Granger CB, Alonso-García Á, Armstrong P, Flather M *et al.* Large streamlined trials in cardiovascular disease. Eur Heart J 2014;35:544–8.

[3] Sun X, Ioannidis JA, Agoritsas T, Alba AC, Guyatt G. How to use a subgroup analysis: users' guide to the medical literature. JAMA 2014;311: 405–11.

[4] Burke JF, Sussman JB, Kent DM, Hayward RA. Three simple rules to ensure reasonably credible subgroup analyses. BMJ 2015;351:h5651.

[5] Gabler NB, Duan N, Raneses E, Suttner L, Ciarametaro M, Cooney E *et al.* No improvement in the reporting of clinical trial subgroup effects in high-impact general medical journals. Trials 2016;17:320.

[6] Sleight P. Debate: subgroup analyses in clinical trials: fun to look at—but don't believe them! Curr Control Trials Cardiovasc Med 2000;1: 25–7.

[7] Academic Research Organization Consortium for Continuing Evaluation of Scientific Studies, Patel MR, Armstrong PW, Bhatt DL, Braunwald E, Camm AJ *et al.* Sharing data from cardiovascular clinical trials–A proposal. N Engl J Med 2016;375:407–9.

[8] Koopman L, van der Heijden GJ, Hoes AW, Grobbee DE, Rovers MM. Empirical comparison of subgroup effects in conventional and individual patient data meta-analyses. Int J Technol Assess Health Care 2008;24: 358–61.

[9] Mandl KD, Glauser T, Krantz ID, Avillach P, Bartels A, Beggs AH *et al.* The Genomics Research and Innovation Network: creating an interoperable, federated, genomics learning system. Genet Med 2020;22:371–80.

[10] Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. N Engl J Med 2007;357:2189–94.

[11] Gaudino M, Benedetto U, Fremes S, Biondi-Zoccai G, Sedrakyan A, Puskas JD *et al.* Radial-artery or saphenous-vein grafts in coronary-artery bypass surgery. N Engl J Med 2018;378:2069–77.

[12] Cole GD, Nowbar AN, Mielewczik M, Shun-Shin MJ, Francis DP. Frequency of discrepancies in retracted clinical trial reports versus unretracted reports: blinded case-control study. BMJ 2015;351:h4708.

[13] Kasenda B, Schandelmaier S, Sun X, von Elm E, You J, Blümle A *et al.* Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications. BMJ 2014;349:g4539.

[14] Serruys PW, Morice M-C, Kappetein AP, Colombo A, Holmes DR, Mack MJ *et al.* Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. N Engl J Med 2009;360: 961–72.

[15] Kappetein A-P, Serruys PW, Sabik JF, Leon MB, Taggart DP, Morice M-C *et al.* Design and rationale for a randomised comparison of everolimus-eluting stents and coronary artery bypass graft surgery in selected patients with left main coronary artery disease: the EXCEL trial. EuroIntervention 2016;12:861–72.

[16] Sterne JA, Davey SG. Sifting the evidence-what's wrong with significance tests? BMJ 2001;322:226–31.

[17] Lagakos SW. The challenge of subgroup analyses—reporting without distorting. N Engl J Med 2006;354:1667–69.

[18] Althouse AD. Adjust for multiple comparisons? It's not that simple. Ann Thorac Surg 2016;101:1644–45.

[19] Pocock SJ, Assmann SE, 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.

[20] Stefanini GG, Baber U, Windecker S, Morice MC, Sartori S, Leon MB *et al.* Safety and efficacy of drug-eluting stents in women: a patient-level pooled analysis of randomised trials. Lancet 2013;382:1879–88.

[21] 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.

[22] Head SJ, Kaul S, Tijssen JP, Serruys PW, Kappetein A. Subgroup analyses in trial reports comparing percutaneous coronary intervention with coronary artery bypass surgery. JAMA 2013;310:2097–98.

[23] Head SJ, Milojevic M, Daemen J, Ahn J-M, Boersma E, Christiansen EH *et al.* Mortality after coronary artery bypass grafting versus percutaneous coronary intervention with stenting for coronary artery disease: a pooled analysis of individual patient data. Lancet 2018;391:939–48.