ESC HEART FAILURE
ESC Heart Failure 2021; **8**: 26–36
Published online 30 November 2020 in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/ehf2.13122

ORIGINAL RESEARCH ARTICLE

Reporting and interpretation of subgroup analyses in heart failure randomized controlled trials

Muhammad Shahzeb Khan[1], Muhammad Arbaz Arshad Khan[2], Simra Irfan[2], Tariq Jamal Siddiqi[2], Stephen J. Greene[3], Stefan D. Anker[4], Jayakumar Sreenivasan[5], Tim Friede[6], Ayman Samman Tahhan[7], Muthiah Vaduganathan[8], Gregg C. Fonarow[9] and Javed Butler[10]*

[1]*Department of Medicine, Cook County Health Sciences, Chicago, IL, USA;* [2]*Department of Medicine, Dow University of Health Sciences, Karachi, Pakistan;* [3]*Division of Cardiology, Duke University Medical Center, Durham, NC, USA;* [4]*Department of Cardiology (CVK) and Berlin Institute of Health Center for Regenerative Therapies (BCRT), German Centre for Cardiovascular Research (DZHK) partner site Berlin, Charité Universitätsmedizin Berlin, Berlin, Germany;* [5]*Department of Cardiology, Westchester Medical Center and New York Medical College, Valhalla, NY, USA;* [6]*Department of Medical Statistics, University Medical Center Goettingen and DZHK, partnerside Goettingen, Goettingen, Germany;* [7]*Division of Cardiology, Department of Medicine, Emory University School of Medicine, Atlanta, GA, USA;* [8]*Brigham and Women's Hospital Heart & Vascular Center, Boston, MA, USA;* [9]*Division of Cardiology, Ronald Reagan-UCLA Medical Center, Los Angeles, CA, USA;* [10]*Department of Medicine, University of Mississippi, Jackson, MS, USA*

Abstract

Aims This study aimed to investigate the reporting of subgroup analyses in heart failure (HF) randomized controlled trials (RCTs) and to determine the strength and credibility of subgroup claims.

Methods and results All primary HF RCTs published in nine high-impact journals from 1 January 2008 to 31 December 2017 were included. Multivariable regression analysis was used to identify factors that may favour the reporting of results in specific subgroups. Strength of the subgroup effect claimed was classified into (i) strong, (ii) likely, or (iii) suggestive. Credibility of subgroup claim was scored using a pre-specified 10 pointer criteria. Of the 261 HF RCTs studied, 107 (41%) reported subgroup analyses. Twenty-five (23%) RCTs claimed a subgroup effect for the primary outcome of which six (24%) made a strong claim, eight (32%) claimed a likely effect, and 11 (44%) suggested a possible subgroup effect. Seven of the 25 RCTs did not employ interaction testing for subgroup claims of the primary outcome. Three out of 10 pre-specified credibility criteria were satisfied by half of the trials. Fourteen trials justified the choice of subgroups, and 10 explicitly stated they were underpowered to detect differences within subgroups. Source of funding did not influence the frequency of reporting subgroup analyses (OR 0.53, 95% CI 0.78–3.62, $P = 0.52$).

Conclusions Appropriate credibility criteria were rarely met even by HF RCTs that held strong subgroup claims. Subgroup analyses should be pre-specified, be adequately powered, present interaction terms, and be replicated in independent data before being integrated into clinical decision making.

Keywords Subgroup claims; Credibility; Strength of claims; Study characteristics; HF RCTs

Received: 18 August 2020; Revised: 18 October 2020; Accepted: 3 November 2020
**Correspondence to: Javed Butler, Department of Medicine, University of Mississippi Medical Center, 2500 N. State Street, Jackson, MS 39216, USA. Tel: 601 984-5600; Fax: 601 984-5608. Email: jbutler4@umc.edu*

Introduction

Subgroup analyses in randomized controlled trials (RCTs) are frequently conducted with almost 50% of RCTs reporting them.[1,2] Subgroup analysis aims to determine whether the overall treatment effect varies across different patient characteristics.[3,4] Subgroup analyses may offer valuable insights; however, they are prone to yield false positives results owing to multiple comparisons and false negatives owing to inadequate power. Moreover, subgroup analyses can be confounded. Thus, credible reporting and interpretation of subgroup analyses are critical. This is especially true in several noted circumstances, for example, statistically borderline positive or neutral results, unexpected subgroup analysis, when an *a priori* clinical belief system exists, or sometimes payers using data to streamline payment decisions. A noted example may include the claim that amlodipine reduces the risk of death in patients with heart failure (HF) owing to

non-ischaemic cardiomyopathy.[5] However, this claim was later disproven by an RCT.[6]

Subgroup analyses are especially important for RCTs in diseases like HF owing to the heterogeneity of the patient population. Apart from age, sex, and ethnicity, patients with HF may differ in their aetiology (HF with reduced ejection fraction or HF with preserved ejection fraction), severity, co-morbidities, and medications. These differences are multiplied in global trials. It is difficult for investigators to include a homogenous set of patients as stringent criteria would result in insurmountable recruitment challenges and related exorbitant costs and trial duration and would affect the generalizability of results.[7] However, misguided claims from subgroup analyses can result in potentially beneficial therapies being overlooked, potentially harmful treatments being administered, and most importantly ineffective treatments being based on spurious subgroup findings. Indeed, many subgroup results have proven to be subsequently false in the past.[8] Therefore, it is critical that clinicians and health policymakers make appropriate inferences from subgroup analyses.

There has been no study that has systematically assessed the credibility claims of subgroup analyses and their credibility claims comprehensively in patients with HF. In this study, we sought to study the characteristics that influence the reporting of subgroup analyses in HF clinical trials and to assess the credibility and strength of the claims of subgroup effect.

Methods

This systematic review was conducted in accordance with the American Heart Association guidelines[9] and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[10] Our study follows the detailed protocol outlined by Sun and colleagues.[11] Because this is a review of publicly accessible information, no approval from institutional review board was acquired.

Sources, search, and eligibility of studies

We searched PubMed using the Cochrane highly sensitive search strategy for HF RCTs published from 1 January 2008 to 31 December 2017 in nine high-impact journals [*New England Journal of Medicine (NEJM)*, *The Lancet*, *Journal of the American Medical Association (JAMA)*, *Journal of the American College of Cardiology (JACC)*, *JACC: Heart Failure*, *Circulation*, *Circulation: Heart Failure*, *European Heart Journal*, and *European Journal of Heart Failure*]. A detailed search strategy is provided in *Table S1*. Studies that published the primary results of an HF RCT were included in our analysis.

We excluded any reports published after the primary report of an RCT (i.e. secondary publications), studies comprising non-human subjects, research letters, pharmacokinetic studies, and Phase 1 RCTs. RCTs investigating diagnostic test accuracy and economic evaluations were also not included. The articles retrieved from the systematic search were exported to EndNote reference library, Version X8.1 (Clarivate Analytics), where two independent reviewers (M. A. A. K. and S. I.) assessed the remaining articles, and only those that met the predefined criteria were selected. In case of any inconsistencies in agreement, a third investigator (M. S. K.) was consulted.

Data extraction

For each of the selected RCTs, the following study characteristics were extracted: total number of patients enrolled; type of primary outcome reported as 'time to event', 'continuous', 'binary', or 'others'; main effect of primary outcome, which was classified as 'significant' (when the primary outcome reported a value of $P < 0.05$) or 'non-significant' ($P > 0.05$); study design ('parallel', 'factorial', or 'crossover'); unit of randomization ('individual' or 'cluster'); and funding source, which categorized as 'industry' (where RCTs received total or partial funding from industry) or 'other' (RCTs funded by non-profit organizations, hospitals, or those that received no external funding). The sources of funding were determined according to their specification in the methods, acknowledgements, and the funding section of the studies. Area of study was classified as 'surgical' (studies that focused on surgical interventions) and 'non-surgical' (drug and behavioural interventions).

Subgroup definition

The RCTs reporting subgroup analyses were further surveyed to extract more information. In cases where a claim of subgroup effect was made, the strength and credibility of the claim were evaluated. A subgroup within an RCT was defined as a subset of the enrolled population that shared a common trait among patients or an intervention measured at either baseline or after randomization. Subgroup analysis was defined as a statistical analysis conducted to investigate whether the effects of the intervention (experimental vs. control) varied according to the status of a subgroup variable. A study was considered to have conducted a subgroup analysis if the study had reported: a *P*-value for one or more subgroups, a magnitude of difference in the effect between patient subgroups, the result of an interaction test, or mentioned an explicit statement that a subgroup analysis had been done. Pre-specification of subgroup analyses was determined by searching original trial reports and their

supplementary materials. Study protocols were also searched if they were referred to or provided in the supplementary materials.

Subgroup effects claim and credibility

Strength of a subgroup claim was determined by using the criterion utilized by previous studies.[12,13] According to that criterion based on 7 points, claims were categorized into strong effect, likely effect, and suggestion of an effect. The details are provided in *Table S2*. They were categorized according to strong claim (authors are confident about the claim of the subgroup effect), claim of likely effect (authors believe that a subgroup effect might exist), and claim of a suggestion of possible effect (authors propose that there is a subgroup effect but are dubious about the claim). Credibility of subgroup effect claimed for a primary outcome by an individual study was assessed using the 10-point criteria developed by Sun *et al*.[14] These criteria are an updated version of the original 7-point criteria authored by Sun *et al*.[13] It was devised to help clinicians assess the credibility of a subgroup claim and was later improved upon by Sun *et al*. Four of these criteria were pertinent to design, two looked at analyses, and four explored the context (*Table S3*). If one RCT claimed more than one subgroup effect for the primary outcome, the subgroup effect with the strongest claim was selected. In cases where the strength of two or more claims was the same, the claim mentioned first was preferred over the others. This was done to avoid a possible clustering effect.[11]

Statistical analysis

To determine the association of study characteristics on the reporting of subgroup analyses in frequency tables, we used χ^2 tests. Furthermore, multivariable regression analyses were conducted to examine the predictors of subgroup reporting. Study area (non-surgical vs. surgical), statistical significance of the primary outcome (significant vs. non-significant), mean sample size per arm (calculated by dividing total number of patients enrolled by number of arms), funding source (industry vs. non-industry), and the number of primary outcomes were included in a logistic regression model for subgroup reporting (yes vs. no). To evaluate the joint effect of funding source and statistical significance of the primary outcome on the reporting of subgroup analyses, the interaction term (funding source × statistical significance) was also included in our regression model. If this interaction term was statistically significant, the role of funding source in reporting of subgroup analysis was further investigated in the following two groups: significance and non-significance of the primary outcome. Assessment of subgroup credibility was done by calculating the proportions of claims that met each criterion and the number of trials met by each criteria. A test of trend using one-way ANOVA was done to investigate if stronger claims met more criteria. The κ coefficient measures the degree of agreement between the two independent investigators. Landis and Koch's novel methodological criteria[15] were used to classify the coefficient. Kappa calculator by Statistics Solutions was utilized to calculate the κ statistics with the respective 95% confidence intervals (CIs) for each outcome by entering the frequency of agreements and disagreements between the reviewers. All analyses were performed using Statistical Package for the Social Sciences (SPSS) software Version 23 (International Business Machines Corporation, New York, USA). All comparisons were two-tailed, and $P < 0.05$ was considered statistically significant.

Results

Literature search

After initial screening, 1271 potentially relevant articles were identified, of which 261 met the criteria to be included (*Table S4*). The PRISMA flow chart (*Figure 1*) summarizes the literature search. There was substantial agreement in reproducibility for subgroup reporting (κ = 0.72; 95% CI 0.49–0.85) and *a priori* subgroup hypothesis (κ = 0.77; 95% CI 0.52–0.87).

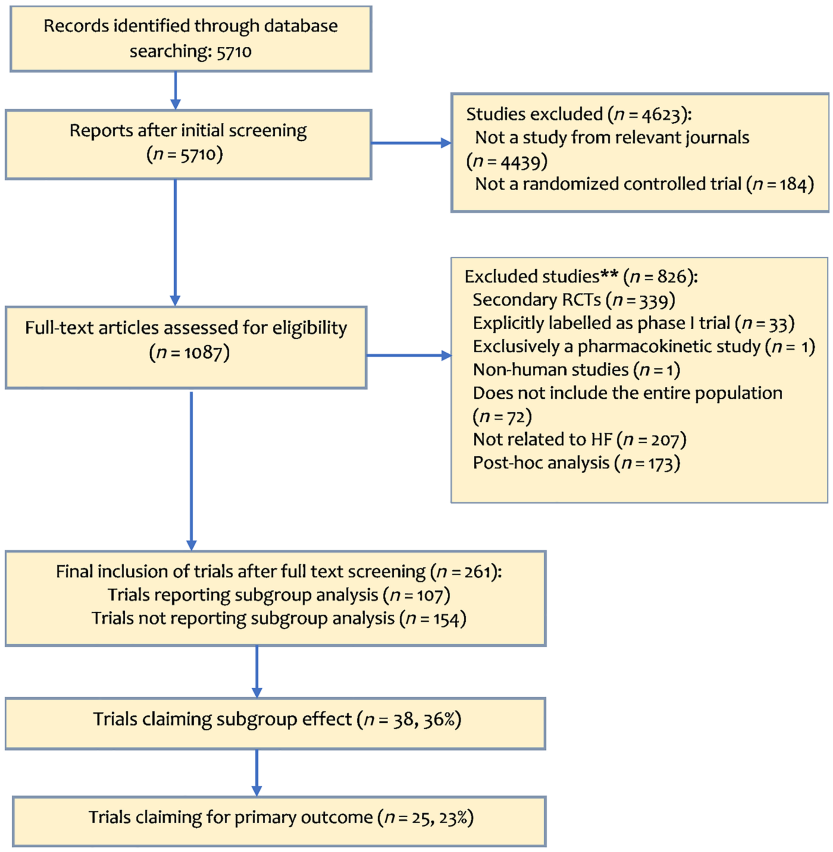
Trial characteristics

A total of 107 (41%) RCTs reported subgroup analyses. *Table 1* outlines the study characteristics of the included RCTs that did and did not report subgroup analyses. Of the 107 RCTs reporting subgroup analyses, 38 (36%) had claimed subgroup effects, of which 25 (23%) had claimed effects for the primary outcomes. Of the 107 RCTs that conducted subgroup analyses, 72 (67%) RCTs had pre-specified them. The most frequently reported subgroup analysis was sex (56%), age (55%), race (23%), severity of disease (64%), aetiology of HF (50%), and co-morbidities (53%). Fourteen trials justified the choice of the subgroups, and 10 trials explicitly stated that they were underpowered to detect the differences within subgroups. Seven trials reemphasized their claim at the end of the study.

Effect of study characteristics on subgroup reporting

Table S5 shows the results of the logistic regression examining the relation between-study characteristics and subgroup analysis. Multivariable regression showed that only sample size (OR 14.1, 95% CI 5.2–38.3 P = 0.003) was statistically significant in the association of study characteristics and

Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow chart summarizes the literature search process. A total of 1271 potentially relevant articles were identified, of which 261 met the criteria to be included.



*New England Journal of Medicine (NEJM), Lancet, Journal of the American Medical Association (JAMA), Journal of the American College of Cardiology (JACC), JACC: Heart Failure, Circulation, Circulation: Heart Failure, European Heart Journal, and European Journal of Heart Failure.

** Studies may be excluded for multiple reasons.

subgroup reporting with larger mean sample size per arm trials being more likely to report subgroup analyses. The interaction term (funding source statistical significance) was not statistically significant (OR 1.84, 95% CI 0.57–5.92, P = 0.30).

Characteristics of subgroup claims

Of the 38 trials that claimed subgroup effects, 28 (74%) trials made one claim, eight (21%) made two claims, and two (5%) made three or more claims. Eleven (29%) made strong claims, nine (24%) claimed a likely effect, and 18 (47%) suggested a possible effect. Thirteen trials made 20 claims for secondary endpoints. Twenty-five (23%) trials claimed a subgroup effect for the primary outcome of which six (24%) made a strong claim (1/6 hypothesized an effect *a priori*), eight (32%) claimed a likely effect (1/8 hypothesized an effect *a priori*), and 11 (44%) suggested a possible subgroup effect (1/11 hypothesized an effect *a priori*). *Table 2* outlines the strengths of subgroup claims for primary endpoints.

Trends in credibility

In the 25 trials that reported subgroup effects for primary outcomes, the criteria satisfied most often (23; 92%) was measurement of the subgroup variable at baseline (*Table 2*). Only three criteria among the 10 were met by more than half (13; 50%) of the studies, including enrolment stratified by subgroup variable, significant test of interaction, and subgroup variable being a baseline characteristic, regardless of the strength of subgroup claim (*Figure 2*). There was a total of 34 subgroup effects claimed for the primary outcome by 25 RCTs (*Table 4*), and only four trials satisfied five or more

Table 1 Study characteristics of randomized controlled trials reporting subgroup analyses

Study characteristics	RCTs reporting subgroup analyses (n = 107) (%)		RCTs not reporting subgroup analyses (n = 154) (%)
Median (inter-quartile range) sample size per study arm	188 (76–687)		41.3 (18.5–123)
Source of funding:		P = 0.001	
Industry	68 (63.8)		68 (55.8)
Other	39 (36.2)		86 (44.2)
Study area:		P = 0.003	
Non-surgical	80 (74.3)		130 (84.6)
Surgical	27 (25.7)		24 (15.4)
Main effect for primary outcome:		P = 0.018	
Statistically significant	40 (37.4)		89 (57.7)
Statistically non-significant	67 (62.6)		65 (42.3)
Study design:		$P < 0.001$	
Parallel	103 (99.4)		139 (90.8)
Factorial	1 (0.1)		1 (0.6)
Crossover	2 (0.4)		12 (7.5)
Single group	1 (0.1)		2 (1.1)
Unit of randomization:		$P < 0.001$	
Individual participant	106 (98.9)		154 (100)
Cluster of participants	1 (1.1)		0 (0)
Type of selected primary outcome:		P = 0.05	
Time to event	13 (11.4)		7 (5.2)
Binary	20 (19)		8 (5.1)
Continuous	43 (40)		107 (69.2)
Others	31 (29.5)		32 (20.5)

than five of the 10-point criteria. One trial correctly pre-specified the direction of the subgroup effect; however, three trials pre-specified subgroup hypotheses in total. The median number of criteria met by trials was as follows: strong claim, 2.5 (1.75–3.25); claim of likely effect, 4 (3–5); and suggestion of a possible effect, 3 (2–4). Test of trend revealed no significant differences between any of the three groups.

Discussion

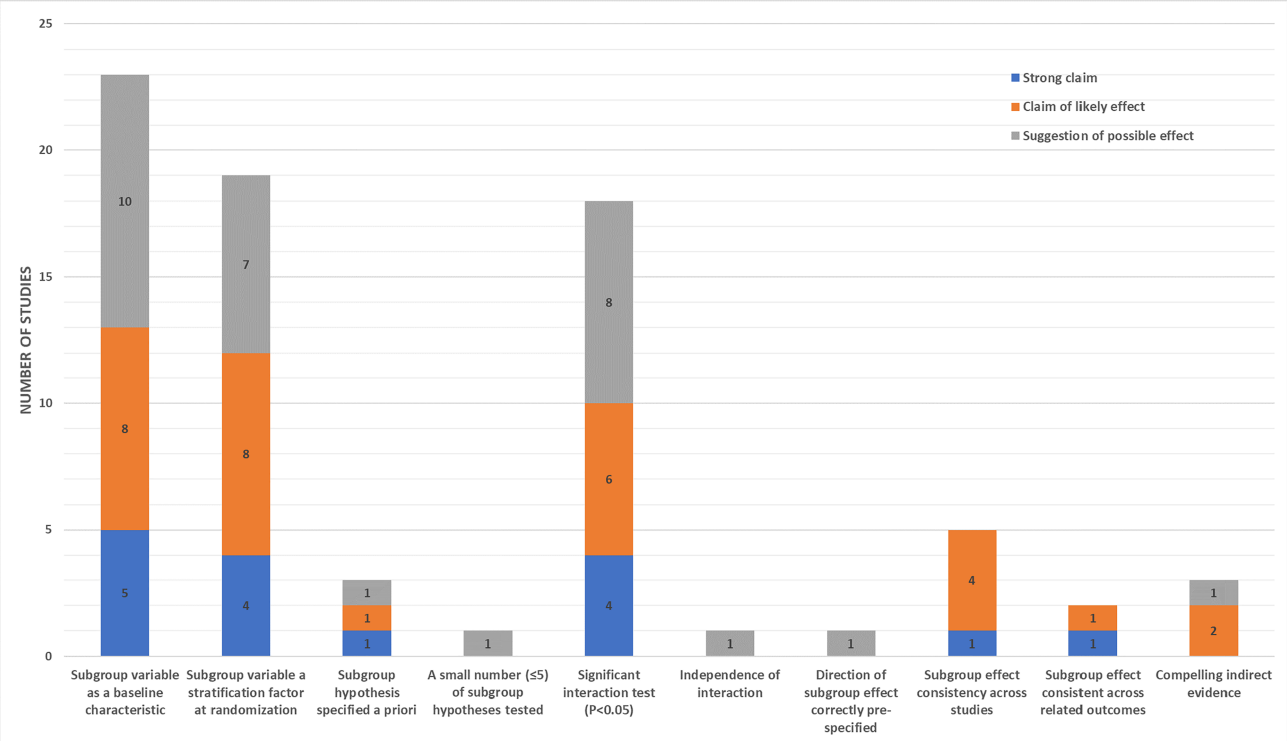
In this analysis of contemporary HF RCTs, we found that the credibility of subgroup claims was generally low across all strengths of claims. One-third of all RCTs claiming subgroup effects failed to report a test of interaction. Only one trial pre-specified the direction of the subgroup hypothesis. We also found no association between industry funding and

Table 2 Proportion of claims meeting subgroup criteria for primary outcomes

Criteria	Strong claim (n = 6)	Claim of likely effect (n = 8)	Suggestion of possible effect (n = 11)	Total (n = 25)
Subgroup variable as a baseline characteristic	5 (83)	8 (100)	10 (91)	23 (92)
Subgroup variable a stratification factor at randomization	4 (67)	8 (100)	7 (64)	19 (76)
Subgroup hypothesis specified *a priori*	1 (17)	1 (13)	1 (9)	3 (12)
A small number (≤5) of subgroup hypotheses tested	0 (0)	0 (0)	1 (9)	1 (4)
Significant interaction test ($P < 0.05$)	4 (67)	6 (75)	8 (73)	18 (72)
Independence of interaction	0 (0) (n = 3[a])	0 (0) (n = 4[a])	1 (9) (n = 1[a])	1 (4) (n = 8[a])
Direction of subgroup effect correctly pre-specified	0 (0)	0 (0)	1 (9)	1 (4)
Subgroup effect consistency across studies	1 (17)	4 (50)	0 (0)	5 (20)
Subgroup effect consistent across related outcomes	1 (17)	1 (13)	0 (0)	2 (8)
Compelling indirect evidence	0 (0)	2 (25)	1 (10)	3 (12)

[a]Number of trials having two or more subgroup claims

Figure 2 The bar chart summarizes the proportion of trials meeting the 10-point criteria across all three strengths of claims. Blue colour represents strong claim, orange colour represents claim of a likely effect, and grey colour represents suggestion of a possible effect.



reporting of subgroup analyses. These results have important clinical implications as misguided claims from subgroup analyses can result in potentially beneficial therapies being overlooked and potentially harmful treatments or less efficacious treatments being administered to the patient.

Our results show that widely accepted and preferred methods of subgroup reporting are infrequently followed.[16–18] These findings are consistent with studies published in other fields as well.[19] Pre-specifying subgroup analyses is an integral part of analyses, as it reflects that the investigators expected a certain subgroup effect beforehand. Not pre-specifying subgroup analyses is analogous to betting on a sports team after the match has concluded, lessening the credibility of subgroup claims. Such findings should be interpreted with caution until they can be replicated in subsequent studies. Moreover, authors should consider providing a subgroup analyses credibility criteria list and referencing relevant sections of the protocol and/or statistical analysis plan.

Around one-third of the RCTs did not report a test of interaction value. Test of interaction represents an important criterion in establishing the eligibility of claimed subgroup effect. The *P*-value of interaction test aims to assess the probability that if true diverging effects exist between two groups, they can be explained by chance or an accidental finding rather than an actual effect. A previous study showed that about 30% of cardiovascular RCTs from 2002 to 2004 reported interaction testing in subgroup analyses[1]; however, our results showed that approximately three-fourths (72%) of HF RCTs had reported interaction testing, indicating that trends are getting more favourable.

Direction of subgroup hypothesis was reported seldom, which is also consistent with previous reviews.[11] It is imperative to specify the correct direction of subgroup hypotheses in addition to *a priori* specification of subgroup hypothesis as illustrated by the following example: in a trial investigating vasopressin and norepinephrine in septic shock patients, investigators pre-specified the hypothesis and direction; in patients with more severe septic shock, reduced mortality will be observed to be attributable to vasopressin over norepinephrine.[20] However, vasopressin was only found to reduce mortality in patients with less severe septic shock. This failure to correctly identify the hypothesis direction considerably lessened the inference that vasopressin was superior to norepinephrine in less severely ill patients.

We did not find any association between industry funding and subgroup reporting, contrary to previous studies.[11,15] Industry-funded trials usually involve multiple centres, numerous researchers, and analysts over an expansive setting, reflecting the substantial energy and resources dispensed to successfully execute a trial. In cases where the outcome is

DOI: 10.1002/ehf2.13122

not statistically significant, researchers may be inclined to look for a statistically significant result in a subset of the population, which may lead to spurious outcomes. However, our results showed that while two-thirds of all RCTs that reported subgroup analyses were funded by industry and 60% of these had non-significant primary outcomes, the difference in reporting of subgroup analyses between industry and non-industry-funded trials was not significant.

It is important to note that subgroup analyses might be conducted for a myriad of reasons. Subgroup analyses may assist in generating important hypotheses, and if conducted appropriately, such exploration can help with better informed clinical decision making and improved patient care.[21] This can lead to the identification of baseline factors that show an increased efficacy among certain patient subgroups. Similarly, subgroup analysis can shed light on specific subgroups that show more deleterious side effects. There have been quite a few studies outlining steps to help clinicians differentiate between true and false subgroup effects.[20–24] Suggestions included following a stringent criteria to assess credibility of the claims outlined by Sun *et al*.[22] and Burke *et al*.[23] The criteria include the following: whether chance explains the subgroup effect, if the subgroup effect is consistent across studies, whether the subgroup hypothesis developed *a priori* with direction specified, whether there is strong biological support, whether evidence supporting the effect based on within-study or between-study comparisons is present, whether the subgroup effect being investigated had a prior probability of at least 20% (preferably >50%), and whether more than one subgroup analysis is performed. Others aim at documentation of all planned subgroup analyses in RCT registries such as clinicaltrials.gov and strongly adhering to protocols of RCTs such as those described in the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) statement.[24,25]

Subgroup results may be interpreted in two ways: researchers may only wish to demonstrate the presence of a heterogenic effect in a particular subgroup (known simply as effect heterogeneity), or they may use it to entirely attribute that effect to the secondary factor that defines that subgroup, that is, age, sex, and race. This latter interpretation would be termed as establishing a causal relationship between the secondary subgroup factor and subgroup effect, and it would only be appropriate to establish this relationship if other confounding variables within that subgroup have been controlled for during their analysis. For example, consider a treatment effect that was more prominent in the female subgroup. If the men were generally older than women and the treatment effect was more prominent in younger individuals, therefore, age might be the confounding factor responsible for the heterogeneity. Furthermore, the precise interpretation of the results of a positive subgroup effect can be masked by the heterogeneity of HF population, particularly when employing multiple variables.

In summary, researchers should be careful when undertaking subgroup analyses. All-subgroup analyses should be pre-specified to a few variables and confined only to the primary outcome. Moreover, while interpreting subgroup analyses, results should be considered exploratory and should rarely be mentioned in conclusion. Interaction test should be used instead of the *P*-values in each subgroup, and researchers should clearly mention whether the RCT was powered enough to detect these subgroup differences to avoid false-negative results. *Table 3* provides a checklist for best practices of reporting and interpretation of subgroup analyses in HF RCTs. The current Consolidated Standards of

Table 3 Checklist for reporting and interpreting subgroup analyses in heart failure trials

#	Checklist item	Reported
1	Was *P*-interaction used to evaluate differences in subgroup?	
2	Was the subgroup variable tested pre-specified with correct direction?	
3	Was the selection of all pre-specified subgroups justified?	
4	Was the subgroup effect tested only for primary outcome?	
5	Was the subgroup claim mentioned in the conclusion?	
6	If possible and where appropriate, was pre-specified subgroup analyses performed based on A. Sex B. Age C. Left ventricular ejection fraction D. Any other clinically important variable pertaining to the topic	
7	Was the subgroup effect being tested?	
8	Was the RCT powered enough to detect the subgroup differences?	
9	Was there more than one subgroup analysis reported?	
10	Were study characteristics (length of follow-up, type of patients enrolled, clinical setting, and study design) and patient characteristics (type of HF, baseline medication usage, doses, age range, differing interventions, different comparators, different definition of outcomes, and varying baseline risk) of the included studies described in detail and accounted as possible sources of heterogeneity	
11	Was the subgroup effect claimed congruous with other studies?	
12	Was there a compelling biological evidence to support the claim?	
13	Was a satisfactory explanation of the heterogeneity observed and resultant impact on study findings reported in the discussion section? (e.g. exploratory and/or chance findings)	

Table 4 List of subgroup claims for only the primary outcomes and their corresponding level of strength

Trial	PMID	Claim	Claim strength
Ivabradine and outcomes in chronic heart failure (SHIFT): a randomised placebo-controlled study.	20801500	(1) A significant treatment effect observed in subgroup with baseline heart rate higher than the median 77 b.p.m.	Strong claim
		(2) A significant improvement observed in NYHA class subgroup[a]	Strong claim
Impact of oxypurinol in patients with symptomatic heart failure. Results of the OPT-CHF study.	18549913	Elevated SUA responded favourably to oxypurinol	Strong claim
Short-Term Effects of Tolvaptan in Patients With Acute Heart Failure and Volume Overload.	28302292	(1) Patients without elevated jugular venous pressure showed directional favourability of tolvaptan	Strong claim
		(2) Patients without ascites showed directional favourability of tolvaptan	Strong claim
Targeted left ventricular lead placement to guide cardiac resynchronization therapy: the TARGET study: a randomized, controlled trial.	22405632	Greatest clinical response is seen in patients with a concordant LV lead	Strong claim
A randomized study of haemodynamic effects and left ventricular dyssynchrony in right ventricular apical vs. high posterior septal pacing in cardiac resynchronization therapy.	22286156	Concordant LV leads provided superior LV reverse remodelling and LV reverse dys-synchrony	Strong claim
Cardiac-Resynchronization Therapy for the Prevention of Heart-Failure Events	19723701	(1) CRT–ICD therapy was associated with a greater benefit in women	Strong claim
		(2) CRT–ICD therapy was associated with a greater benefit in patients with a QRS duration of 150 ms or more	Strong claim
Results of a non-specific immunomodulation therapy in chronic heart failure (ACCLAIM trial): a placebo-controlled randomised trial.	18207018	(1) In NYHA Class II patients, IMT was associated with greater reduction	Claim of likely effect
		(2) In patients with no history of myocardial infarction, IMT was associated with greater reduction	Claim of likely effect
Chronic kidney disease and cardiac remodelling in patients with mild heart failure: results from the REsynchronization reVErses Remodeling in Systolic Left vEntricular Dysfunction (REVERSE) study.	22956574	In participants assigned to CRT, those without CKD had significantly greater improvements in LV structural parameters	Claim of likely effect
Low-dose dopamine or low-dose nesiritide in acute heart failure with renal dysfunction: the ROSE acute heart failure randomized trial.	24247300	(1) The treatment effect was significant with low-dose dopamine in subgroups of patients with higher ejection fraction	Claim of likely effect
		(2) The treatment effect was significant with low-dose dopamine of patients with higher baseline blood pressure	Claim of likely effect
Spironolactone for heart failure with preserved ejection fraction.	24716680	Significant treatment effect observed among patients enrolled on the basis of an elevated natriuretic peptide level	Claim of likely effect
Tailored telemonitoring in patients with heart failure: results of a multicentre randomized controlled trial.	22588319	(1) Subgroup analysis showed greater effects in patients with a heart failure duration $<$ 18 months	Claim of likely effect
		(2) Subgroup analysis showed greater effects in patients having a pacemaker	Claim of likely effect
		(3) Subgroup analysis showed greater effects in patients not living alone	Claim of likely effect
Assessment of long-term effects of irbesartan on heart failure with preserved ejection fraction as measured by the Minnesota living with heart failure questionnaire in the irbesartan in heart failure with preserved systolic function (I-PRESERVE) trial.	22267751	Significant outcomes observed in subgroups that had only slightly better or worse symptoms of heart failure or a change in one NYHA class	Claim of likely effect
Cardiac-resynchronization therapy for mild-to-moderate heart failure.	21073365		Claim of likely effect

(*Continues*)

Table 4 (continued)

Trial	PMID	Claim	Claim strength
		(1) There was a significant interaction between treatment and QRS duration of 150 ms or more	
		(2) Patients with left bundle branch block appeared to have a greater benefit than patients with non-specific intra-ventricular conduction delay	Claim of likely effect
Cardiovascular Outcomes with Minute Ventilation-Targeted Adaptive Servo-Ventilation Therapy in Heart Failure: The CAT-HF Trial.	28335841	A positive effect of ASV observed in patients with HF with preserved ejection fraction	Claim of likely effect
Effect of Ularitide on Cardiovascular Mortality in Acute Heart Failure.	28402745	A significant treatment effect observed for geographical region	Suggestion of a possible effect
A prospective comparison of alginate-hydrogel with standard medical therapy to determine impact on functional capacity and clinical outcomes in patients with advanced heart failure (AUGMENT-HF trial).	26082085	A significant treatment effect observed in subgroups of patients split by median 6MWT distance	Suggestion of a possible effect
Increased mortality after dronedarone therapy for severe heart failure.	18565860	The risk of death associated with dronedarone was increased among patients who had a lower wall-motion index as compared with those who had a higher wall-motion index	Suggestion of a possible effect
Short- and long-term treatment of dilutional hyponatraemia with satavaptan, a selective arginine vasopressin V2-receptor antagonist: the DILIPO study.	21199833	A significantly higher response rates seen with both 25 and 50 mg/day in the subgroup of CHF patients	Suggestion of a possible effect
The angiotensin receptor neprilysin inhibitor LCZ696 in heart failure with preserved ejection fraction: a phase 2 double-blind randomised controlled trial.	22932717	A significant treatment effect seen in patients with diabetes	Suggestion of a possible effect
Vagus Nerve Stimulation for the Treatment of Heart Failure: The INOVATE-HF Trial.	27058909	The significant treatment effect observed among women	Suggestion of a possible effect
Defibrillator Implantation in Patients with Nonischemic Systolic Heart Failure.	27571011	A significant treatment-by-subgroup interaction observed in subgroup differing by age	Suggestion of a possible effect
Early eplerenone treatment in patients with acute ST elevation myocardial infarction without heart failure: the Randomized Double-Blind Reminder Study.	24780614	(1) A significant subgroup observed in the subgroup split by heart rate	Suggestion of a possible effect
		(2) A significant subgroup observed in the subgroup split by timing of acute reperfusion	Suggestion of a possible effect
Effect of Aldosterone Antagonism on Exercise Tolerance in Heart Failure With Preserved Ejection Fraction.	27765184	A significant improvement with spironolactone in subgroups above and below an RER of 1	Suggestion of a possible effect
Primary results from the SmartDelay determined AV optimization: a comparison to other AV delay methods used in cardiac resynchronization therapy (SMART-AV) trial: a randomized trial comparing empirical, echocardiography-guided, and algorithmic atrioventricular delay programming in cardiac resynchronization therapy.	21098426	Women optimized with SD and echo responded more favourably to the treatment effect	Suggestion of a possible effect
Effects of high-dose versus low-dose losartan on clinical outcomes in patients with heart failure (HEAAL study): a randomised, double-blind trial.	19922995	Patients without a history of hypertension had greater treatment benefit	Suggestion of a possible effect

6MWT, 6-min walk test; ASV, adaptive servo-ventilation; CHF, chronic heart failure; CKD, chronic kidney disease; CRT, cardiac resynchronization therapy; ICD, implantable cardioverter–defibrillator; IMT, immunomodulation therapy; LV, left ventricle; NYHA, New York Heart Association; RER, respiratory exchange ratio; SD, smart delay; SUA, serum uric acid.
[a]Some trials made more than one claim.

Reporting Trials (CONSORT) guidelines offer minimal recommendation and guidance about subgroup analyses.

Several limitations need to be considered. First, we only included RCTs published in high-impact factor journals from 2008 to 2017, and therefore, our study does not adequately represent the RCTs from lower-impact journals and those published outside this time period. However, we anticipate that the quality of subgroup analysis reporting will be even lower for other trials published in lower-impact factor journals. Second, authors most often do not describe the conduct and results of subgroup analyses in adequate detail, especially regarding biological rationale; therefore, our results cannot take into account the data that were not reported in the RCTs. However, we argue that for readers to judge the credibility of subgroup analyses and interpret it appropriately, authors must provide sufficient information at least in the appendix. Third, the 10-point criteria utilized to assess the credibility of subgroup effect may vary in importance, and not all 10 points may be equally important. For example, pre-specification of subgroup analyses and utilizing statistical significance of interaction tests may be more important than other points. Lastly, we did not include secondary publications based on analyses of subgroups. However, we expect that these secondary publications will have even more frequent claiming of subgroup effect claim with lower credibility.

In conclusion, the credibility of subgroup effect claims was low across all strengths of claims including strong claims. Clinicians should be aware that even strong claims from subgroup analyses are often inconsistent and, therefore, must exercise caution when using subgroup effect claims to guide clinical decision making. Additionally, reporting of primary outcome subgroup analyses should follow a strict standardized criterion. Subgroup analyses should be pre-specified with direction hypothesized, tested only for primary outcome, *P*-interaction reported, and their selection justified.

Conflict of Interest

SJG has received a Heart Failure Society of America/Emergency Medicine Foundation Acute Heart Failure Young Investigator Award funded by Novartis, receives research support from the American Heart Association, Amgen, AstraZeneca, Bristol-Myers Squibb, Merck, and Novartis; serves on advisory boards for Amgen and Cytokinetics; and serves as a consultant for Amgen and Merck. SDA reports receiving fees from Bayer, Boehringer Ingelheim, Cardiac Dimension, Impulse Dynamics, Novartis, Servier, St. Jude Medical, and Vifor Pharma and grant support from Abbott Vascular and Vifor Pharma. AST is supported by the Abraham J. & Phyllis Katz Foundation (Atlanta, GA) and NIH/NIA grant AG051633. MV is supported by the KL2/Catalyst Medical Research Investigator Training award from Harvard Catalyst (NIH/NCATS Award UL 1TR002541); receives research grant support from Amgen and Boehringer Ingelheim; serves on advisory boards for Amgen, American Regent, AstraZeneca, Baxter Healthcare, Bayer AG, Boehringer Ingelheim, Cytokinetics, and Relypsa; and participates on clinical endpoint committees for studies sponsored by Galmed, Novartis, and the NIH. GCF reports consulting for Abbott, Amgen, AstraZeneca, Bayer, CHF Solutions, Janssen, Medtronic, Merck, and Novartis. JB declares that he serves as a consultant for Abbott, Adrenomed, Amgen, Array, AstraZeneca, Bayer, Boehringer Ingelheim, Bristol Myers Squib, CVRx, G3 Pharmaceutical, Impulse Dynamics, Innolife, Janssen, LivaNova, Luitpold, Medtronic, Merck, Novartis, NovoNordisk, Relypsa, Roche, V-Wave Limited, and Vifor. TF reports personal fees from Novartis, Bayer, Janssen, SGS, Roche, Boehringer Ingelheim, Daiichi-Sankyo, Galapagos, Penumbra, Parexel, Vifor, BiosenseWebster, CSL Behring, Fresenius Kabi, Coherex Medical, LivaNova; all outside the submitted work.

Funding

None.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Detailed search strategy.
Table S2. Seven-point criteria used to assess the strength of claims.
Table S3. Ten-point criteria used to assess credibility of subgroup claims.
Table S4. List of the included trials.
Table S5. Regression analyses of factors associated with reporting versus not reporting of subgroup analyses.

References

1. Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? *Am Heart J* 2006; **151**: 257–264.
2. 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**: 426–432.
3. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q* 2004; **82**: 661–687.
4. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005; **365**: 176–178.
5. Packer M, O'Connor C, Ghali J, Pressler ML, Carson PE, Belkin RN, Miller AB, Neuberg GW, Frid D, Wertheimer JH, Cropp AB. Effect of amlodipine on morbidity and mortality in severe chronic heart failure. *N Engl J Med* 1996; **335**: 1107–1114.
6. Packer M, Carson P, Elkayam U, Konstam MA, Moe G, O'Connor C, Rouleau JL, Schocken D, Anderson SA, DeMets D, PRAISE-2 Study Group. Effect of amlodipine on the survival of patients with severe chronic heart failure due to a nonischemic cardiomyopathy. *JACC: Heart Failure* 2013; **1**: 308–314.
7. Ford I, Norrie J. Pragmatic trials. *N Engl J Med* 2016; **375**: 454–463.
8. Guyatt G, Wyer PC, Ioannidis J. When to believe a subgroup analysis. In Guyatt G., Rennie D., Meade M. O., Cook D. J., eds. *User's Guide to the Medical Literature: A Manual for Evidence-based Clinical Practice*, 2nd ed. AMA; 2008. p 571–583.
9. Rao G, Lopez-Jimenez F, Boyd J, D'Amico F, Durant NH, Hlatky MA, Howard G, Kirley K, Masi C, Powell-Wiley TM, Solomonides AE, West CP, Wessel J, American Heart Association Council on Lifestyle and Cardiometabolic Health; Council on Cardiovascular and Stroke Nursing; Council on Cardiovascular Surgery and Anesthesia; Council on Clinical Cardiology; Council on Functional Genomics and Translational Biology; and Stroke Council. Methodological standards for meta-analyses and qualitative systematic reviews of cardiac prevention and treatment studies: a scientific statement from the American Heart Association. *Circulation* 2017; **136**: e172–e194.
10. Moher D, Liberati A, Tetzlaff J, Altman D. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. *PLoS Med* 2009; **6**: e1000097.
11. Sun X, Briel M, Busse J, You JJ, Akl EA, Mejza F, Bala MM, Bassler D, Mertz D, Diaz-Granados N, Vandvik PO. The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. *BMJ* 2011; **342**: d1569–d1569.
12. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, Bala MM, Bassler D, Mertz D, Diaz-Granados N, Vandvik PO, Malaga G, Srinathan SK, Dahm P, Johnston BC, Alonso-Coello P, Hassouneh B, Walter SD, Heels-Ansdell D, Bhatnagar N, Altman DG, Guyatt GH. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ* 2012; **344**: e1553.
13. Sun X, Briel M, Busse JW, Akl EA, You JJ, Mejza F, Bala M, Diaz-Granados N, Bassler D, Mertz D, Srinathan SK, Vandvik PO, Malaga G, Alshurafa M, Dahm P, Alonso-Coello P, Heels-Ansdell DM, Bhatnagar N, Johnston BC, Wang L, Walter SD, Altman DG, Guyatt GH. 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.
14. Oxman AD. Subgroup analyses. *BMJ* 2012; **344**: e2022.
15. Landis J, Koch G. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–174.
16. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992; **116**: 78–84.
17. 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**: 2917–2930.
18. 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**: 93–98.
19. Kasenda B, Schandelmaier S, Sun X, von Elm E, You J, Blümle A, Tomonaga Y, Saccilotto R, Amstutz A, Bengough T, Meerpohl JJ, Stegert M, Olu KK, Tikkinen KA, Neumann I, Carrasco-Labra A, Faulhaber M, Mulla SM, Mertz D, Akl EA, Bassler D, Busse JW, Ferreira-González I, Lamontagne F, Nordmann A, Gloy V, Raatz H, Moja L, Rosenthal R, Ebrahim S, Vandvik PO, Johnston BC, Walter MA, Burnand B, Schwenkglenks M, Hemkens LG, Bucher HC, Guyatt GH, Briel M, DISCO Study Group. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications. *BMJ* 2014; **349**: g4539–g4539.
20. Russell JA, Walley KR, Singer J, Gordon AC, Hébert PC, Cooper DJ, Holmes CL, Mehta S, Granton JT, Storms MM, Cook DJ, VASST Investigators. Vasopressin versus norepinephrine infusion in patients with septic shock. *N Engl J Med* 2008; **358**: 877–887.
21. Rothwell PM, Eliasziw M, Gutnikov SA, Fox AJ, Taylor DW, Mayberg MR, Warlow CP, Barnett HJ, Carotid Endarterectomy Trialists' Collaboration. Analysis of pooled data from the randomised controlled trials of endarterectomy for symptomatic carotid stenosis. *Lancet* 2003; **361**: 107–116.
22. Sun X, Ioannidis JPA, Agoritsas T, Alba AC, Guyatt G. How to use a subgroup analysis: users' guide to the medical literature. *JAMA* 2014; **311**: 405–411.
23. Burke J, Sussman J, Kent D, Hayward R. Three simple rules to ensure reasonably credible subgroup analyses. *BMJ* 2015; **351**: h5651.
24. Chan AW, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, Hróbjartsson A, Mann H, Dickersin K, Berlin JA, Doré CJ, Parulekar WR, Summerskill WS, Groves T, Schulz KF, Sox HC, Rockhold FW, Rennie D, Moher D. SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Ann Intern Med* 2013; **158**: 200–207.
25. Chan AW, Tetzlaff JM, Gotzsche PC, Altman DG, Mann H, Berlin JA, Dickersin K, Hrobjartsson A, Schulz KF, Parulekar WR, Krleza-Jeric K, Laupacis A, Moher D. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ* 2013; **346**: e7586.