JAMA Internal Medicine | Original Investigation

Evaluation of Evidence of Statistical Support and Corroboration of Subgroup Claims in Randomized Clinical Trials

Joshua D. Wallach, MS, PhD; Patrick G. Sullivan, MD, MS; John F. Trepanowski, PhD; Kristin L. Sainani, MS, PhD; Ewout W. Steyerberg, MSc, PhD; John P. A. Ioannidis, MD, DSc

IMPORTANCE Many published randomized clinical trials (RCTs) make claims for subgroup differences.

OBJECTIVE To evaluate how often subgroup claims reported in the abstracts of RCTs are actually supported by statistical evidence ($P < .05$ from an interaction test) and corroborated by subsequent RCTs and meta-analyses.

DATA SOURCES This meta-epidemiological survey examines data sets of trials with at least 1 subgroup claim, including Subgroup Analysis of Trials Is Rarely Easy (SATIRE) articles and Discontinuation of Randomized Trials (DISCO) articles. We used Scopus (updated July 2016) to search for English-language articles citing each of the eligible index articles with at least 1 subgroup finding in the abstract.

STUDY SELECTION Articles with a subgroup claim in the abstract with or without evidence of statistical heterogeneity ($P < .05$ from an interaction test) in the text and articles attempting to corroborate the subgroup findings.

DATA EXTRACTION AND SYNTHESIS Study characteristics of trials with at least 1 subgroup claim in the abstract were recorded. Two reviewers extracted the data necessary to calculate subgroup-level effect sizes, standard errors, and the *P* values for interaction. For individual RCTs and meta-analyses that attempted to corroborate the subgroup findings from the index articles, trial characteristics were extracted. Cochran *Q* test was used to reevaluate heterogeneity with the data from all available trials.

MAIN OUTCOMES AND MEASURES The number of subgroup claims in the abstracts of RCTs, the number of subgroup claims in the abstracts of RCTs with statistical support (subgroup findings), and the number of subgroup findings corroborated by subsequent RCTs and meta-analyses.

RESULTS Sixty-four eligible RCTs made a total of 117 subgroup claims in their abstracts. Of these 117 claims, only 46 (39.3%) in 33 articles had evidence of statistically significant heterogeneity from a test for interaction. In addition, out of these 46 subgroup findings, only 16 (34.8%) ensured balance between randomization groups within the subgroups (eg, through stratified randomization), 13 (28.3%) entailed a prespecified subgroup analysis, and 1 (2.2%) was adjusted for multiple testing. Only 5 (10.9%) of the 46 subgroup findings had at least 1 subsequent pure corroboration attempt by a meta-analysis or an RCT. In all 5 cases, the corroboration attempts found no evidence of a statistically significant subgroup effect. In addition, all effect sizes from meta-analyses were attenuated toward the null.

CONCLUSIONS AND RELEVANCE A minority of subgroup claims made in the abstracts of RCTs are supported by their own data (ie, a significant interaction effect). For those that have statistical support ($P < .05$ from an interaction test), most fail to meet other best practices for subgroup tests, including prespecification, stratified randomization, and adjustment for multiple testing. Attempts to corroborate statistically significant subgroup differences are rare; when done, the initially observed subgroup differences are not reproduced.

JAMA Intern Med. doi:10.1001/jamainternmed.2016.9125
Published online February 13, 2017.

Invited Commentary
Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: John P. A. Ioannidis, MD, DSc, Stanford University, 1265 Welch Rd, Medical School Office Bldg, Room X306, Stanford, CA 94305 (jioannid@stanford.edu).

In medicine, there is a growing interest in developing treatment and prevention strategies that are tailored to unique patient characteristics (ie, "stratified medicine" or "precision medicine").[1,2] Evidence for these strategies often comes from subgroup analyses reported in randomized clinical trials (RCTs).[3-6] Considering that the results from individual subgroup tests are often misleading and can lead to withholding of treatment or provision of incorrect, ineffective, or harmful treatments, it is important to understand the credibility of subgroup effects reported in RCTs.

Previous research suggests that subgroup analyses are often poorly conducted and reported.[5-9] For example, Wang et al[5] pointed out key problems in subgroup claims in published RCTs. First, most subgroup analyses in RCTs fail to provide basic statistical support for their claims.[5,8-11] The presence of a statistical effect in one subgroup but not the other does not constitute evidence of a subgroup effect, as many authors mistakenly believe; rather, the appropriate statistical approach for establishing a subgroup test is a formal test of interaction.[12] Second, trials often perform numerous subgroup analyses that are not prespecified or adjusted for multiple testing, which increases the probability of false-positive findings.[5,6,9] Third, most trials fail to randomize participants within subgroups (eg, stratified randomization),[5,8,9] which leaves more room for imbalanced confounders between treatment and control arms within subgroups. Collectively, these problems may affect the credibility of subgroup findings from RCTs.

Previous studies[3,8,9] have assessed the credibility of subgroup differences reported anywhere in the text of RCTs but have not focused on those most likely to be credible (ie, those reported in the articles' abstracts). Presumably, authors are more careful and selective about reporting subgroup differences in abstracts because these claims are most visible to the research community. Furthermore, to our knowledge, no previous studies have attempted to evaluate the credibility of subgroup findings by checking to see if they are corroborated (eg, new studies producing the same results with the same experimental methods). Specifically, we were interested in examining how often subgroup findings with statistical support (a significant formal test result of interaction) from RCTs are corroborated by subsequent RCTs or meta-analyses. The widespread inability to replicate published research and the lack of replication in the biomedical literature highlight the importance of corroborating previous subgroup findings.[13-15]

Herein, we used 2 samples of RCTs with subgroup claims anywhere in the text to answer 4 questions: (1) how often are subgroup claims (with or without statistical support) reported in the abstracts of RCTs? (2) how often do these subgroup claims have formal statistical support? (3) how often are the abstract subgroup claims with formal statistical support based on a subgroup stratification factor at randomization, preplanned, and based on analyses adjusted for multiple comparisons? and (4) how often are the abstract subgroup claims with statistical support corroborated by subsequent RCTs and meta-analyses?

Key Points

Question How often are subgroup claims reported in the abstracts of randomized clinical trials supported by a statistically significant interaction test result and corroborated by subsequent randomized clinical trials and meta-analyses?

Findings In this meta-epidemiological survey, a minority of subgroup claims (46 of 117) in the abstract of randomized clinical trials were supported by their own data. Only 5 of these 46 subgroup findings had at least 1 subsequent corroboration attempt, and none of the corroboration attempts had a statistically significant *P* value from an interaction test.

Meaning Claims of subgroup differences in randomized clinical trials are typically spurious or chance findings.

Methods

Details about the study are available in the eAppendix in the Supplement. This study included no human participants (it is a meta-epidemiological survey based on summary data available to the public). For this reason, there was no need for institutional review board approval.

Identification of RCTs With Subgroup Claims

To identify a sample of RCTs with subgroup claims in the abstract, we analyzed RCTs with at least 1 subgroup claim from the Subgroup Analysis of Trials Is Rarely Easy (SATIRE)[3,9,16] and Discontinuation of Randomized Trials (DISCO)[8] study groups. The SATIRE and DISCO study groups previously investigated characteristics related to the reporting and validity of subgroup claims. Because both study groups had already compiled separate samples of RCTs, we were able to request data from the study authors and construct a new database consisting of only RCTs with at least 1 subgroup analysis. Study descriptions, additional definitions, and inclusion and exclusion criteria appear in the original SATIRE and DISCO publications.

We used the same definition of a *subgroup* as both the SATIRE and DISCO study groups. We focused on subgroup differences that appeared in the abstract of an article, which are the most visible ones. We defined a subgroup effect as *claimed* if there was either a clear or an implied statement that the effects of an intervention (ie, experiment vs control) differed according to the presence of a subgroup variable. Detailed definitions can be found in the eAppendix in the Supplement. The RCTs from the SATIRE and DISCO studies classified as including a subgroup claim in the abstract are hereafter referred to as "index articles." We defined a *subgroup finding* as a subgroup claim with evidence of statistically significant heterogeneity across subgroup levels from an interaction test or where the authors qualitatively implied that there was evidence of such statistically significant heterogeneity. We defined a *pure corroboration attempt* as a subsequent RCT or meta-analysis with an analysis for the exact same subgroup findings as reported in the index article (ie, for same subgroup levels, interventions, outcomes, and study population). A subgroup finding was considered corroborated if a subsequent RCT or meta-analysis provided subgroup-level effect sizes that were in the same direction

as those reported in the index article and had evidence of statistically significant heterogeneity across subgroup levels from an interaction test ($P < .05$).

Two reviewers (J.D.W. and P.G.S.) independently screened all index articles (n = 169) to determine the subset of the articles that made subgroup claims in the abstract. Three additional reviewers arbitrated all potential discrepancies (J.F.T., K.L.S., and J.P.A.I.).

Identification of Corroboration Attempts

We used Scopus, a large abstract and citation database of peer-reviewed literature, to search for English-language publications citing each of the eligible index articles with at least 1 subgroup finding (searches updated July 2016). Within Scopus, one can search for the title of a study and obtain a list of all of the articles citing the study of interest. One reviewer (J.D.W.) screened the title and abstract of all citing articles to determine the citing RCTs and meta-analyses. The RCTs and meta-analyses were downloaded and screened by 2 reviewers (J.D.W. and P.G.S.) for evidence of subgroup corroboration attempts. Three additional investigators (J.F.T., K.L.S., and J.P.A.I.) arbitrated any uncertainties.

Data Extraction

For each index article with at least 1 subgroup claim in the abstract, we recorded the first author, year of publication, journal, and sample size randomized. We also extracted the compared interventions, population assessed, and outcomes for each individual subgroup claim. We noted the total number of subgroup claims, the number of claims where a *P* value was provided from a test for interaction, the number of claims where a statistically significant *P* value from a test for interaction was reported, the number of claims where there was not enough information provided in the full text to formally test for subgroup heterogeneity, and the number of claims where there was a statement in the full text indicating a subgroup finding (eg, "the interaction term was statistically significant").

For claims without clear evidence of statistical heterogeneity, 2 reviewers (J.D.W. and P.G.S.) extracted the relative or absolute effect sizes, CIs, standard errors, or any other available data to calculate subgroup-level effect sizes and standard errors. When the index articles did not provide effect measures for the subgroups of interest, we used our best judgment to determine whether to calculate a relative or absolute effect measure, depending on the other effect measures reported in the index article. When the choice was unclear, we calculated relative effect measures because multiplicative scale interactions are more often assessed and reported based on logistic or Cox proportional hazards regression models in RCTs.[17-19] An online digitizer (WebPlotDigitizer; http://arohatgi.info/WebPlotDigitizer) was used to extract approximate values from figures. When exact calculations were not possible, 2 reviewers (J.D.W. and K.L.S.) discussed the information and determined if it was possible to approximate the *P* value for interaction with enough precision to confidently classify it as significant or not significant.

For individual RCTs and meta-analyses that attempted to corroborate the subgroup finding from the index article, we extracted the first author, journal, year of publication, and whether there was any overlap in authorship with the index article. When there were several meta-analyses attempting the same corroboration, we focused on the most inclusive one. For any meta-analysis citing an index article with a subgroup finding, we extracted the number of studies and number of participants included in the subgroup effect calculation, the number of studies included in the calculation of the average effect size at the subgroup level that were published after the index study, the overall summary effect size and 95% CI, and the summary effect size and 95% CI in each pertinent subgroup level. This information was used to reevaluate heterogeneity using data from all available individual trials.

After we implemented suggestions raised by peer reviewers, we also screened all index articles to determine how often the abstract subgroup claims with formal statistical support were based on a stratification factor at randomization; were prespecified in the abstract, methods, or results of the trials; and were based on analyses adjusted for multiple comparisons. Finally, we considered the possibility that the index articles themselves might be corroboration attempts of previously published subgroup findings. To evaluate this possibility, we determined whether index articles cited previous RCTs with similar subgroup findings (ie, for same comparison, outcomes, and subgroup levels and with a significant *P* value for interaction).

Statistical Analysis

Subgroup-level effect estimates and standard errors were entered into a software program (R, version 3.2.3; The R Project for Statistical Computing), and the metafor package was used to test for heterogeneity using Cochran *Q* test.[20] When index articles reported hazard ratios, another software program (RevMan, version 5.4; Cochrane Collaboration) was used to test for heterogeneity (J.D.W. and P.G.S.). A third investigator (K.L.S.) reviewed all subgroup claim classifications and reevaluated the test for interactions applying the method by Altman and Bland.[21]

For any meta-analysis attempting to corroborate a subgroup finding, we extracted the available data and tested for interaction using Cochran *Q* test. Trial data were combined within each subgroup level based on the DerSimonian and Laird procedure for random effects. *P* values were 2-tailed.

Results

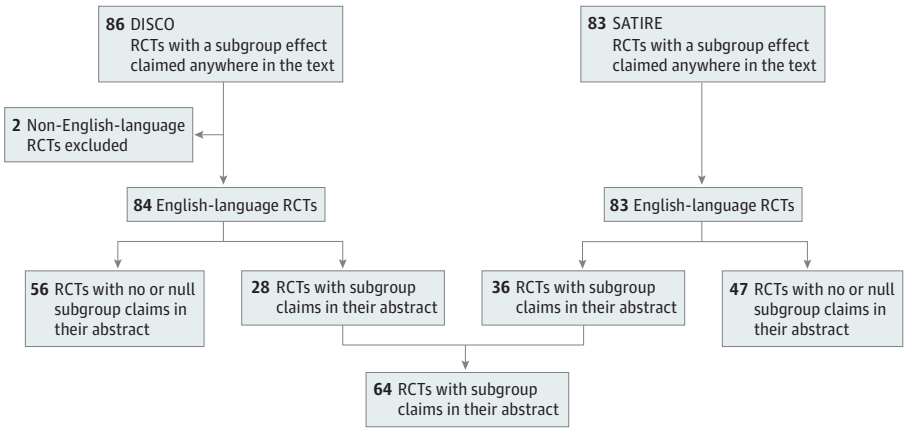
Search Findings

Among the 169 articles with a subgroup effect claimed anywhere in the text, there were 64 articles (37.9%) with at least 1 subgroup claim made in the abstract. In these 64 articles, a total of 117 individual abstract subgroup claims were made (**Figure**).

Frequency and Characteristics of Subgroup Claims

Table 1 summarizes the characteristics of the subgroup claims evaluated. Among the 117 subgroup claims, there were 33 (28.2%) with a corresponding *P* value from a statistical test for interaction anywhere in the text and 47 (40.2%) with data that could be extracted to assess whether there was statistical interaction. We found that more than half of the subgroup claims made (83 [70.9%]) pertained to primary outcomes.

Figure. Flow Diagram of DISCO and SATIRE Articles Review Process



DISCO indicates Discontinuation of Randomized Trials; RCTs, randomized clinical trials; and SATIRE, Subgroup Analysis of Trials Is Rarely Easy.

Table 1. Characteristics of Subgroup Claims

Characteristic	Value
Sample size among index articles with subgroup claims, median (interquartile range)	450 (202-876)
Subgroup claims made in the abstract, No. (%)[a]	117
P value for interaction provided in the abstract	10 (8.5)
P value for interaction provided in the text only	23 (19.7)
Information extracted to calculate Cochran *Q*	47 (40.2)
Not enough information for exact heterogeneity calculations	26 (22.2)
Qualitative statement implying no subgroup effect provided in the text	9 (7.7)
Qualitative statement claiming subgroup effect provided in the text	2 (1.7)

[a] Subgroup claims were clear or implied statements in the abstracts of randomized clinical trials that the effects of an intervention differed according to the presence of a subgroup variable.

Table 2. Characteristics of Subgroup Findings (Subgroup Claims With Statistical Support)[a]

Characteristic	Value
Sample size among index articles with subgroup findings, median (interquartile range)	445 (210-821)
Statistical support provided, No. (%)	46
P value for interaction provided in the abstract	10 (21.7)
P value for interaction provided in the text only	14 (30.4)
Information available to calculate *P* value for interaction	18 (39.1)
Not enough information for exact *P* value calculation but classified as a subgroup finding based on approximations	2 (4.3)
Qualitative statement indicating statistically significant interaction included in the text	2 (4.3)
Subgroup analyses prespecified in the abstract, methods, or results, No. (%)	13 (28.3)
Subgroup variable a stratification factor at randomization, No. (%)[b]	16 (34.8)
P value for interaction adjusted for multiple testing, No. (%)	1 (2.2)

[a] Subgroup findings were subgroup claims with evidence of statistical significant heterogeneity ($P < .05$) across subgroup levels from an interaction test or where the authors qualitatively implied that there was evidence of such statistically significant heterogeneity.

[b] Some index articles had multiple subgroup analyses for different outcomes based on the same subgroup variable used as a stratification factor at randomization. There were 13 individual subgroup variables used as a stratification factor at randomization, ignoring multiple outcomes per study.

We found that most (24 of 33 [72.7%]) of the claims with a reported interaction *P* value were statistically significant and that less than half (18 of 47 [38.3%]) of the claims for which data were extracted to calculate a *P* value for interaction were statistically significant. Overall, of the 117 subgroup claims evaluated, only 46 (39.3%) had statistical support (ie, a significant *P* value for interaction).

Frequency and Characteristics of Subgroup Findings

Table 2 lists the characteristics of the 46 subgroup findings (ie, the subgroup claims with statistical support). For 13 (28.3%) of the 46 subgroup findings, the analyses were listed as prespecified in the abstract, methods, or results sections of the corresponding RCTs. Furthermore, it was evident that the language used to discuss prespecification (ie, preplanned, a priori, previously suggested, planned, and prespecified) and nonprespecification (ie, secondary, explanatory, preliminary, and post hoc) lacked consistency across studies. For 16 of the 46 subgroup findings (34.8%), the subgroup variable was used as a stratification variable during randomization. Only 1 subgroup finding was adjusted for multiple comparisons (the Bonferroni-Holm step-down procedure). Overall, the most common medical fields represented were cardiovascular (n = 7) and infectious disease (n = 5).

Corroboration of Subgroup Findings

Among the 46 subgroup findings, only 5 (10.9%) had at least 1 subsequent pure corroboration attempt by a meta-analysis or an RCT (**Table 3**). None of the corroboration attempts had $P < .05$ from an interaction test, and the subgroup-level effect estimates based on meta-analyzed data were generally attenuated toward the null (relative risk, odds ratio, or hazard ratio of 1.0) compared with the index article. One of the full corroboration attempts is described in the **Box** for illustration, and the remaining descriptions can be found in the

Table 3. Five Subgroup Findings With Full Corroboration Attempts

Characteristics of the Subgroup Findings (Index Articles)					Results of the Corroboration Attempts[a]
Comparison (Year)	Subgroups	Population Characteristics	Outcome (Primary)	Index Article *P* Value for Interaction	*P* Value for Interaction Corroboration Attempt
Supportive expressive group therapy vs control (2007)[22]	Estrogen receptor status negative vs positive	Women with confirmed metastatic or locally recurrent breast cancer	Survival (yes)	.002	.71
Standard care vs standard care without intravenous cooling (2007)[23]	Patients with initial ventricular fibrillation vs no ventricular fibrillation	Patients aged ≥18 y, resuscitated by paramedics from nontraumatic, out-of-hospital cardiac arrest	Awakening (no)	.046[b]	No *P* value provided, no evidence of subgroup difference
			Discharge alive from hospital (no)	.048[b]	.52[b]
Dexamethasone sodium phosphate vs placebo (2007)[24]	Patients with confirmed bacterial meningitis vs probable meningitis	Patients aged ≥14 y with suspected bacterial meningitis	Risk of death at 1 mo (yes)	.01[c]	.23
N-terminal brain natriuretic peptide-guided treatment vs symptom-guided treatment (2009)[25]	Patients aged 60-74 vs ≥75 y	Patients aged ≥60 y with systolic heart failure, New York Heart Association class II or greater, prior hospitalization for heart failure within 1 y, and N-terminal brain natriuretic peptide level ≥2 times the upper limit of normal	Mortality (no)	.01	.22

[a] From the most inclusive corroboration meta-analysis or individual randomized clinical trial.

[b] Not provided by the authors; calculated by us based on risk ratios.

[c] Provided by the authors based on relative risk. Because the corroborating meta-analysis provided only information based on odds ratios, we also reevaluated the interaction from the index article based on odds ratios. The interaction *P* value was no longer statistically significant (*P* = .12).

Box. Pure Corroboration Example

Index Study Description

A 2007 article[23] compared standard care with or without intravenous cooling for patients with nontraumatic cardiac arrest resuscitated by paramedics. The authors reported a subgroup claim for the secondary outcome of being discharged alive from the hospital: They reported that infield cooling improved hospital survival for patients with ventricular fibrillation but reduced survival for patients without ventricular fibrillation.

Calculation of *P* Value for Interaction

The authors of the 2007 article[23] did not report a *P* value for interaction, but we were able to calculate this statistic based on data available in the article. When we recalculated the effect sizes on the risk ratio scale, we found risk ratios for survival of 1.44 (95% CI, 0.84-2.44) for patients with ventricular fibrillation and 0.29 (95% CI, 0.07-1.29) for patients without ventricular fibrillation. We found that the *P* value for interaction achieved statistical significance (*P* = .048). The overall treatment effect for cooling was null (risk ratio, 1.20; 95% CI, 0.73-1.98).

Subgroup Corroboration Attempts

We identified 2 meta-analyses[26,27] and 2 randomized clinical trials[28,29] that attempted to corroborate the subgroup finding for the same outcome. We used information from both meta-analyses to identify individual studies. Three trials provided data for both subgroup levels. One trial provided data for only the ventricular fibrillation subgroup, and 1 trial provided data for only the non–ventricular fibrillation subgroup. The meta-analyzed risk ratios were also attenuated to the null, with risk ratios of 0.98 (95% CI, 0.88-1.09) for the ventricular fibrillation group and 1.13 (95% CI, 0.74-1.74) for the non–ventricular fibrillation group. The meta-analyzed *P* value for interaction was not statistically significant (*P* = .52). There was no overall treatment benefit (risk ratio, 1.02; 95% CI, 0.89-1.16) (eFigure in the Supplement).

eAppendix in the Supplement. It should be noted that all SATIRE articles were published in 2007, and the DISCO articles with subgroup findings were published between 2002 and 2012. Three to four years may not be long enough for a new RCT to publish a corroboration attempt.

We also found modified corroboration attempts (ie, different subgroup levels, interventions, outcomes, or study population) for 4 subgroup findings in 3 index articles.[30-32] Two index articles[30,32] had citing meta-analyses; in both cases, the corroboration attempts used different subgroup definitions (changing from 3 subgroup levels to 2 subgroup levels). There were 2 findings from the same RCTs that had subsequent RCTs investigating similar treatment by insulin interaction status. However, none of these RCTs used insulin concentration at 30 minutes as the measurement of insulin status (eTable in the Supplement).

Two index articles with subgroup findings[30,31] mentioned and cited previous RCTs that had evaluated subgroup analyses for similar comparisons, outcomes, and subgroup levels and also had evidence of statistically significant interaction. These index articles could themselves be viewed as potential corroboration attempts. However, in one case, the outcome of mortality had been assessed at a different time point (day 28 instead of day 140).[33] In the other one, a different study population and different dietary interventions were involved.[34]

Discussion

Our empirical evaluation of subgroup claims from the abstracts of RCTs revealed that most claims (71 [60.7%] of 117) failed to have underlying evidence of statistical significance based on a test for interaction. Formal testing for interactions

is not done (or reported) routinely. In addition, most subgroup findings reported in the abstracts of RCTs fail to meet other best practices for subgroup tests, including prespecification, stratified randomization, and adjustment for multiple testing. Rarely are attempts made to corroborate statistically significant subgroup findings in subsequent trials and meta-analyses. Moreover, none of the subsequent meta-analyses or individual RCTs successfully corroborated the subgroup findings. When effect sizes were available (n = 3), we found that the effect sizes were attenuated toward the null.[23-25]

Recent evaluations of RCTs have found that almost half of the publications report subgroup analyses.[8,9] Furthermore, one-third of RCTs that claimed a subgroup effect for a primary outcome reported a corresponding interaction *P* value or information that allowed for calculation of the *P* value for the primary outcome.[9] We found that less than one-third (33 of 117 [28.2%]) of the 117 abstract-level subgroup claims had a corresponding *P* value anywhere in the text. When interaction *P* values were reported, they were often (24 of 33 [69.7%]) statistically significant, but when it was necessary to extract data to evaluate statistical interaction, only a minority (18 of 47 [38.3%]) of these claims were statistically significant. Novel claims for scientific discoveries typically receive more credit than external validation attempts, which may explain why the latter type of research occurs so infrequently. While reproducibility efforts are essential to ensure that trial findings are complete and unbiased, replication studies across the biomedical literature are rare.[14] Previous research suggests that more than one-third of published reanalyses of RCTs lead to different conclusions than those presented by original articles.[35] Herein, we provide additional evidence that most subgroup findings reported in abstracts of RCTs are not subsequently corroborated.

Limitations

Our study has some limitations. First, it is possible that the window of opportunity for corroboration was too short for some index articles. While the SATIRE articles were all published in 2007, the DISCO articles with subgroup findings were published between 2002 and 2012. A minimum of 3 to 4 years may not be long enough for a new RCT to publish a corroboration attempt. We acknowledge that it takes time for the research community to digest the findings from individual RCTs and then plan subsequent RCTs that may or may not evaluate the same subgroup analyses. By evaluating the meta-analyses citing the index articles, we expect to have identified the cumulative evidence related to the more recent DISCO publications. Second, when authors of the index articles presented evidence from tests for interaction or qualitatively stated that subgroup differences existed, we did not perform any additional calculations. We relied on the reported data in the index articles for our calculations. Furthermore, when we extracted data and tested for heterogeneity, we used the effect measures provided by the authors. Because tests for interaction are influenced by the effect measures considered, this limitation may have influenced the classification of certain subgroup findings as to their statistical significance.[36] Third, our Scopus search may not have identified all subgroup corroboration attempts. Some individual trials evaluating subgroup effects may not cite previous articles making the same subgroup claims. We believe that our search strategy was able to capture most corroboration attempts that could have occurred after the publication date of the index articles.

Our experience suggests that the authors of RCTs should avoid putting too much emphasis on subgroup findings. Research consumers, journal reviewers, and journal editors should be cautious about the credibility of subgroup analyses, even those reported prominently in abstracts. We also found examples of subsequent studies claiming to be corroboration attempts for subgroup findings from previous studies but which actually performed modified corroboration attempts. Interaction tests are sensitive to subgroup definitions (ie, 3 group levels or 2 group levels), the effect estimates used (ie, risk ratios or odds ratios), and the exact measurements used for the subgroup or outcome variables. When subsequent studies modify subgroup analyses, they increase the chances of spurious findings and lead research consumers to believe subgroup claims that actually lack adequate support.

Conclusions

Subgroup claims reported in the abstracts of RCTs are often vague, unaccompanied by information pertinent to a test for a significant interaction effect, and unclear regarding prespecification. Our results support the notion that individual subgroup analyses are often spurious and should be considered hypothesis generating. Furthermore, our research indicates that subsequent meta-analyses and RCTs may rarely attempt to corroborate the subgroup findings prominently reported in RCTs. Moreover, when subgroup corroborations are attempted, the initially observed subgroup differences are not demonstrated again.

ARTICLE INFORMATION

Accepted for Publication: November 7, 2016.

Published Online: February 13, 2017. doi:10.1001/jamainternmed.2016.9125

Author Affiliations: Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California (Wallach, Sullivan, Sainani, Ioannidis); Meta-Research Innovation Center at Stanford (METRICS), Stanford University School of Medicine, Stanford, California (Wallach, Sullivan, Ioannidis); Department of Medicine, Stanford University School of Medicine, Stanford, California (Sullivan, Ioannidis); Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, California (Trepanowski, Ioannidis); Department of Public Health, Erasmus MC, Rotterdam, the Netherlands (Steyerberg); Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, California (Ioannidis).

Author Contributions: Drs Wallach and Ioannidis had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Wallach, Sullivan, Steyerberg, Ioannidis.
Acquisition, analysis, or interpretation of data: All authors.
Drafting of the manuscript: Wallach, Sullivan, Sainani.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Wallach, Sullivan, Trepanowski, Sainani.
Study supervision: Wallach, Ioannidis.

Conflict of Interest Disclosures: None reported.

Funding/Support: The Meta-Research Innovation Center at Stanford (METRICS) is supported by a grant from the Laura and John Arnold Foundation. This work was conducted with support from the Stanford Clinical and Translational Science Award to Spectrum, an independent center within Stanford University that supports health-related research activities across Stanford University (grant UL1 TR001085 from the National Institutes of Health [NIH]). Dr Trepanowski is supported by grant T32 HL007034 from the NIH. Dr Steyerberg is partly supported by the PRICES project (grant U01 NS086294 from the NIH). Dr Ioannidis is supported by an unrestricted gift from Sue and Bob O'Donnell to the Stanford Prevention Research Center.

Role of the Funder/Sponsor: The funders were not involved in any aspect related to the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: Benjamin Kasenda, MD, PhD (Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Basel, Switzerland) (and the Discontinuation of Randomized Trials [DISCO] study group) shared the articles and some raw data for the DISCO articles with at least 1 subgroup claim anywhere in the text. Xi Sun, PhD (Chinese Evidence-Based Medicine Center, West China Hospital, Sichuan University, Chengdu, China) (and the Subgroup Analysis of Trials Is Rarely Easy [SATIRE] study group) provided the names of the SATIRE articles with at least 1 subgroup claim anywhere in the text. Drs Kasenda and Sun did not receive any compensation for sharing their data.

REFERENCES

1. Hamburg MA, Collins FS. The path to personalized medicine. *N Engl J Med*. 2010;363(4):301-304.

2. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med*. 2015;372(9):793-795.

3. Sun X, Briel M, Busse JW, et al. The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. *BMJ*. 2011;342:d1569.

4. Hernández 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(2):257-264.

5. 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(21):2189-2194.

6. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000;355(9209):1064-1069.

7. 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(19):2917-2930.

8. Kasenda B, Schandelmaier S, Sun X, et al; DISCO Study Group. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications. *BMJ*. 2014;349:g4539.

9. Sun X, Briel M, Busse JW, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ*. 2012;344:e1553. doi: 10.1136/bmj.e1553

10. Rothwell PM. Treating individuals, 2: subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*. 2005;365(9454):176-186.

11. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med*. 1992;116(1):78-84.

12. Sainani K. Misleading comparisons: the fallacy of comparing statistical significance. *PM&R*. 2010;2(6):559-562.

13. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.

14. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JP. Reproducible research practices and transparency across the biomedical literature. *PLoS Biol*. 2016;14(1):e1002333.

15. Ioannidis JP, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*. 2014;383(9912):166-175.

16. Sun X, Briel M, Busse JW, et al. Subgroup Analysis of Trials Is Rarely Easy (SATIRE): a study protocol for a systematic review to characterize the analysis, reporting, and claim of subgroup effects in randomized trials. *Trials*. 2009;10:101.

17. Girerd N, Rabilloud M, Pibarot P, Mathieu P, Roy P. Quantification of treatment effect modification on both an additive and multiplicative scale. *PLoS One*. 2016;11(4):e0153010.

18. VanderWeele TJ. On the distinction between interaction and effect modification [published corrections appear in *Epidemiology*. 2011;22(5):752 and 2010;21(1):162]. *Epidemiology*. 2009;20(6):863-871.

19. VanderWeele TJ, Robins JM. Four types of effect modification: a classification based on directed acyclic graphs. *Epidemiology*. 2007;18(5):561-568.

20. DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. *Contemp Clin Trials*. 2015;45(pt A):139-145.

21. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ*. 2003;326(7382):219.

22. Spiegel D, Butler LD, Giese-Davis J, et al. Effects of supportive-expressive group therapy on survival of patients with metastatic breast cancer: a randomized prospective trial. *Cancer*. 2007;110(5):1130-1138.

23. Kim F, Olsufka M, Longstreth WT Jr, et al. Pilot randomized clinical trial of prehospital induction of mild hypothermia in out-of-hospital cardiac arrest patients with a rapid infusion of 4°C normal saline. *Circulation*. 2007;115(24):3064-3070.

24. Nguyen TH, Tran TH, Thwaites G, et al. Dexamethasone in Vietnamese adolescents and adults with bacterial meningitis. *N Engl J Med*. 2007;357(24):2431-2440.

25. Pfisterer M, Buser P, Rickli H, et al; TIME-CHF Investigators. BNP-guided vs symptom-guided heart failure therapy: the Trial of Intensified vs Standard Medical Therapy in Elderly Patients With Congestive Heart Failure (TIME-CHF) randomized trial. *JAMA*. 2009;301(4):383-392.

26. Hunter BR, O'Donnell DP, Allgood KL, Seupaul RA. No benefit to prehospital initiation of therapeutic hypothermia in out-of-hospital cardiac arrest: a systematic review and meta-analysis. *Acad Emerg Med*. 2014;21(4):355-364.

27. Huang FY, Huang BT, Wang PJ, et al. The efficacy and safety of prehospital therapeutic hypothermia in patients with out-of-hospital cardiac arrest: a systematic review and meta-analysis. *Resuscitation*. 2015;96:170-179.

28. Kim F, Nichol G, Maynard C, et al. Effect of prehospital induction of mild hypothermia on survival and neurological status among adults with cardiac arrest: a randomized clinical trial. *JAMA*. 2014;311(1):45-52.

29. Castrén M, Nordberg P, Svensson L, et al. Intra-arrest transnasal evaporative cooling: a randomized, prehospital, multicenter study (PRINCE: Pre-ROSC Intranasal Cooling Effectiveness). *Circulation*. 2010;122(7):729-736.

30. Corwin HL, Gettinger A, Fabian TC, et al; EPO Critical Care Trials Group. Efficacy and safety of epoetin alfa in critically ill patients. *N Engl J Med*. 2007;357(10):965-976.

31. Ebbeling CB, Leidig MM, Feldman HA, Lovesky MM, Ludwig DS. Effects of a low-glycemic load vs low-fat diet in obese young adults: a randomized trial. *JAMA*. 2007;297(19):2092-2102.

32. Löwenberg B, Ossenkoppele GJ, van Putten W, et al; Dutch-Belgian Cooperative Trial Group for Hemato-Oncology (HOVON); German AML Study Group (AMLSG); Swiss Group for Clinical Cancer Research (SAKK) Collaborative Group. High-dose daunorubicin in older patients with acute myeloid leukemia [published correction appears in *N Engl J Med*. 2010;362(12):1155]. *N Engl J Med*. 2009;361(13):1235-1248.

33. Corwin HL, Gettinger A, Pearl RG, et al; EPO Critical Care Trials Group. Efficacy of recombinant human erythropoietin in critically ill patients: a randomized controlled trial. *JAMA*. 2002;288(22):2827-2835.

34. Pittas AG, Das SK, Hajduk CL, et al. A low-glycemic load diet facilitates greater weight loss in overweight adults with high insulin secretion but not in overweight adults with low insulin secretion in the CALERIE Trial. *Diabetes Care*. 2005;28(12):2939-2941.

35. Ebrahim S, Sohani ZN, Montoya L, et al. Reanalyses of randomized clinical trial data. *JAMA*. 2014;312(10):1024-1032.

36. Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. *Int J Epidemiol*. 2012;41(2):514-520.