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.
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 al5 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 studies1,3,6-9 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?
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 or 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. 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 explicit 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. 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.

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.
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 non-prespecification (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

<table>
<thead>
<tr>
<th>Characteristics of the Subgroup Findings (Index Articles)</th>
<th>Results of the Corroboration Attemptsa</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comparison (Year)</strong></td>
<td><strong>Subgroups</strong></td>
</tr>
<tr>
<td>Supportive expressive group therapy vs control 2007</td>
<td>Estrogen receptor status negative vs positive</td>
</tr>
<tr>
<td>Standard care vs standard care without intravenous cooling 2007</td>
<td>Patients with initial ventricular fibrillation vs no ventricular fibrillation</td>
</tr>
<tr>
<td>Dexamethasone sodium phosphate vs placebo 2007</td>
<td>Patients with confirmed bacterial meningitis vs probable meningitis</td>
</tr>
<tr>
<td>N-terminal brain natriuretic peptide-guided treatment vs symptom-guided treatment 2009</td>
<td>Patients aged 60–74 vs ≥75 y</td>
</tr>
</tbody>
</table>

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 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 infeild 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 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 and 2 randomized clinical trials 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) (Figure in the Supplement).

**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.6,9 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 corroborate 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.
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 Trierweiler is supported by grant T32 HL070034 from the NIH. Dr Steyerberg is partly supported by the PRISES project (grant U01 NS086294 from the NIH). Dr Ioannidis is supported by an unrestricted gift from Sue and Bob D’onnell 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

7. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting.

Copyright 2017 American Medical Association. All rights reserved.