Supplementary Online Content

Wallach JD, Sullivan PG, Trepanowski JF, Sainani KL, Steyerberg EW, Ioannidis JPA. Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. *JAMA Intern Med*. Published online February 13, 2017. doi:10.1001/jamainternmed.2016.9125

eAppendix. Trial Protocol

eTable. Characteristics of Subgroup Findings With Partial Corroboration Attempts

eFigure. Intravenous Cooling for Nontraumatic Out-of-Hospital Cardiac Arrest for the Outcome of Being Discharged Alive From Hospital

This supplementary material has been provided by the authors to give readers additional information about their work.

eAppendix. Trial Protocol

Additional information about subgroup corroborations

Protocol

Lack of statistical support and of corroboration of subgroup claims from the abstracts of randomized controlled trials

Joshua D Wallach, MS, PhD[1, 2]; Patrick G Sullivan, MS, MD[1, 2]; John F Trepanowski, PhD[3]; Kristin Sainani, MS, PhD[1]; Ewout W Steyerberg, MSc, PhD[4]; John PA Ioannidis, DSc, MD[1-3, 5,6]

[1]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California

[2] Meta-Research Innovation Center at Stanford (METRICS), Stanford University School of Medicine, Stanford, California

[3]Stanford Prevention Research Center, Stanford University, Stanford, California

[4]Department of Public Health, Erasmus MC, Rotterdam, the Netherlands

[5]Department of Statistics, Stanford University, Stanford, California

[6]Stanford University School of Medicine, Stanford, California

Correspondence to: JPA Ioannidis, 1265 Welch Road, Medical School Office Building, Room X306, Stanford, CA 94305, USA jioannid@stanford.edu

Background

Subgroup analyses in randomized controlled trials (RCTs) are used to determine whether treatment effects vary across certain patient characteristics.[1] Subgroup findings from RCTs may be used to tailor patient care ("stratified medicine"). However, spurious subgroup effects can lead to the withholding of treatment, or the provision of incorrect, ineffective, or harmful treatments, so it is essential that these effects are corroborated.[2]

Subgroup analyses are commonly reported in RCTs, with estimates ranging from 40-65%,[1,3-6] but there are many analytical challenges associated with this statistical approach.[7-11] Trials often perform multiple subgroup analyses without correcting for multiple testing, thereby increasing the probability of a false positive finding.[2,7-9] Furthermore, failure to pre-specify subgroup analyses can lead to selective reporting and the reporting of spurious finding given the multiplicity of subgroup analyses performed.[7] According to a recent evaluation of subgroup claims in RCTs, less than 40% of 207 RCTs reporting subgroup analyses pre-specified subgroup hypotheses.[4] Another study found that among RCTs that included pre-specified subgroup analysis plans in a protocol, the number of reported subgroup analyses reported in the publication matched the number planned in the protocol less than 10% of the time.[5] Studies have shown that only about half of RCTs with subgroup analyses use a statistical test for interaction.[4,7]

Current concerns about the inability to replicate published research and the lack of replication in the biomedical literature highlight the importance of validating previous findings.[12-14]Although previous empirical evaluations have quantified the number of subgroup claims in RCTs and the study characteristics that may influence the reporting of subgroup analyses,[3-5] little is known about how frequently subgroup claims are externally validated. Given the importance of applying subgroup information to optimize patient care, the objective of this project is to understand how frequently the most prominent of these claims are validated, i.e. those that appear in the abstracts of papers presenting the results of RCTs.

Aims

Aim 1: To determine how often subgroup claims reported in the abstracts of RCTs are subgroup findings (with formal statistical support).

Aim 2a: To determine how often subsequent RCTs are performed on the same treatment comparison and same disease and outcome when a subgroup finding has been published in the abstract of an RCT and how often these subsequent RCTs corroborate the specific subgroup claim.

Aim 2b: To determine how often meta-analyses citing RCTs with a subgroup claim at the abstract level include an outcome that pertains to corroborating a subgroup claim.

Data sources and Methods

We are using two data sources to identify subgroup claims that have been made in the abstracts of published RCTs:

1. The DISCO Study

A previous study investigated the agreement between the pre-specification of subgroup analyses at the publication level and the corresponding statements at the protocol level.[5] The authors defined a subgroup as "a subset of all trial participants with distinct characteristics at randomization (for example, age, sex, stage of disease)." Subgroup analyses were defined as "an analysis that explored whether intervention effects (experiment versus control) differed according to these characteristics."[5] For additional definitions and methods, please see the original manuscript.[5] We will scrutinize the abstracts of these 86 publications and independently identify the articles in which a claim for a subgroup difference is made in the abstract of the paper

After contacting the authors about their study, they shared the following data for the 86 articles they established as having made subgroup claims: author, journal, year of publication, first page of publication, patients randomized, setting (multicenter trial, single center trial), design (parallel, cross-over), label (non-inf/equiv trial, superiority trial, unclear), unit of randomization, journal impact (low, high), industry sponsored (yes, no), medical field, primary outcome significant, subgroup analysis pre-specified in publication, *ad hoc* subgroup analysis conducted, power calculation for subgroup analysis provided, subgroup hypothesis provided in publication, anticipated direction of subgroup effect provided in publication, interaction term reported, number of subgroup analyses reported, any subgroup claim in publication

2. The SATIRE Group

A second study group characterized the analysis, reporting, and claim of subgroup effects in randomized trials. The Subgroup Analysis of Trials Is Rarely Easy (SATIRE) group conducted a systematic review of 464 randomized controlled human trials published in 2007.[3,4,15]

The authors defined a subgroup as "a subset of a trial population that is identified on the basis of a patient or intervention characteristic that is either measured at baseline or after randomization." They defined a subgroup analysis as "a statistical analysis that explores whether effects of the intervention (i.e. experiment versus control) differ according to status of a subgroup variable. This includes a case in which investigators report a main result and analyze only a subset of patients." In the study protocol, a subgroup effect was defined as “a difference in the magnitude of a treatment effect across subgroups of a study population. The null hypothesis for a test of a subgroup effect (i.e. subgroup hypothesis) is that there is no difference in the magnitude of a treatment effect across subgroups.” For additional definitions, please see the original study protocol.[4,15]

From the 207 RCTs with subgroup analyses, the SATIRE group sent us a list of 83 studies that made subgroup claims. We will scrutinize the abstracts of these 83 publications and independently identify the articles in which a claim for a subgroup difference is made in the abstract of the paper.

Definitions

The RCTs previously established as making subgroup claims in these two previous empirical evaluations will form the basis of our empirical evaluation (hereafter referred to as “index articles”). Subgroup claims with clear evidence of statistical heterogeneity ($P < 0.05$) across subgroup levels will be referred to as “subgroup findings.” We will use the same definitions of subgroup, subgroup effect, and subgroup analyses as the SATIRE study.[15]

Review process

For the index articles, two reviewers (JDW, PGS) will independently determine the subset of the articles that make explicit subgroup claims at the abstract level. In cases where an ambiguous subgroup claim is made at the abstract level (e.g., where a subgroup difference is mentioned but an outcome has not been specified), we will consult the full text of the article to identify all potentially relevant subgroup findings that fit within the scope of the abstract level claim.

For each subgroup claim at the abstract level, we will determine whether a significant *P*-value ($P < 0.05$) is reported from a test for interaction. For claims without clear evidence of statistical heterogeneity, we will extract the relative or absolute effect sizes, confidence intervals, standard errors, or any other available data that can be used to calculate subgroup-level effect sizes and standard errors. The subgroup level effect sizes and standard errors will be entered into R (version 3.2.3) and the metafor package will be used to formally test for heterogeneity. A *P*-value less than 0.05 from Cochran’s Q test will be used to determine the presence of heterogeneity.

We will record the total number of subgroup claims, the total number of claims where a *P*-value is reported from a test for interaction, the total number of claims where a significant *P*-value ($P < 0.05$) is reported from a test for interaction, and the total number of claims with a significant *P*-value ($P < 0.05$) based on calculations. The number of studies with at least one finding will also be recorded.

We will record the first author, year of publication, and journal for each subgroup finding. We will also extract the compared interventions and population (disease/condition) assessed as well as the outcome pertaining to that subgroup claim.

Scopus will be used to scrutinize all citations to each of the eligible index articles with at least one subgroup finding from November to July 2016 (latest search). After recording the total number of citations, we will scan the title and abstract of all citing articles to determine the citing RCTs and meta-analyses. All RCTs and meta-analyses will be downloaded and screened at the full-text for any discussion related to the subgroup findings of interest. We will also record the date of the most recent citing meta-analysis that includes the index RCT in its calculations and the search date of the meta-analysis, when available. Any uncertainties will be discussed in detail with additional reviewers (JFT, KLS, JPAI).

For index articles where we do locate any citing meta-analyses, we will consider the date of the most recent meta-analysis that includes the index RCT in its calculations as the latest time on which we have information on whether a subgroup claim has been validated or not (e.g. if the index RCT was published in 2007, and a citing meta-analysis is published in 2014, with the search date of the meta-analysis being December 2011, then we will consider the literature scanned until December 2011). For these cases, we will perform PubMed searches to identify if any more recent randomized trials exist that examine the same comparison of interventions for the same disease. Search strategies will use the Cochrane Library search string for identifying RCTs coupled with the names of the compared interventions and disease/condition and limited to the time after the publication of the meta-analyses.

Data Abstraction

For RCTs and SR/meta-analyses that assess the subgroup claim from the index article, we will extract the following information:

- First author, journal, year of publication
- Whether any author of the index article with a subgroup finding is also an author of the citing RCT, meta-analysis, or systematic review.

For any eligible meta-analyses, we will consider a subgroup analysis for the main finding as having been reported if the authors report a subgroup-level average effect size for the same intervention(s), outcome(s) and study population(s) as the index article. In particular, we will look for evidence of a subgroup effect included in a forest plot. For eligible articles, we will also record the total number of studies and total number of participants included in the subgroup effect calculation, the total number of studies included in the calculation of the average effect size at the subgroup level were published after the index study, the summary effect size and 95% CI by fixed effects and by random effects in each pertinent subgroup level so as to be able to compare against the data on the subgroup difference in the index study.

For any eligible RCTs, we will extract if the authors report an estimate and an associated confidence interval or a *P*-value for the effect in each pertinent subgroup level and the difference between the subgroups, if the authors report the results from an interaction test, if the *P* value for the interaction test is statistically significant at the 0.05 level, and if the authors state that they performed the subgroup analysis, but do not report any of the data discussed above.

We will compare the subgroup difference in the index study against the subgroup difference seen in subsequent RCTs and/or meta-analyses. The index subgroup effect and the subsequent evidence effect will be compared in terms of direction of effect and magnitude of effect.

Additional definitions and methods (updated after initial study protocol)

Definitions

We will use the same definition of a "subgroup" as both the SATIRE[3,15] and DISCO[5] study groups. We defined a subgroup effect as "claimed" if, at the abstract of the paper, there was either a clear or an implied statement that the effects of an intervention (i.e., experiment versus control) differed according to the presence of a subgroup variable. We will regard a subgroup claim as reported in the abstract if one or more of the following are included: a statistically significant result from a test for interaction; a statement about a possible effect in one or more subgroup levels and not others (or a difference in effect estimates of different subgroups), a statement that a subgroup analysis had been undertaken that resulted in subgroup differences, or an effect estimate for one or more subgroups. We will regard the following as being a "null claim" and not a subgroup claim: a statement that an intervention effect did not differ, or was consistent, between or among subgroup levels. The RCTs from the DISCO and SATIRE studies classified as including a subgroup claim in the abstract will be referred to as "index articles". We further defined a "subgroup finding" as a subgroup claim with evidence of statistically significant heterogeneity *(P* < 0.05) across subgroup levels from an interaction test or where the authors qualitatively implied that there was evidence of statistical heterogeneity. We defined a "pure corroboration attempt" as a subsequent RCT or meta-analyses with an analysis for the exact same subgroup finding(s) as reported in the index article (i.e., for the same subgroup levels, intervention(s), outcome(s), and study population). A subgroup finding will be considered corroborated if a subsequent RCT or meta-analysis has evidence of statistically significant heterogeneity across subgroup levels from an interaction test and if the subgroup-level effect sizes are in the same direction as those reported in the index article. We defined statistical significance as a *P*-value < 0.05.

Methods

Article screening

Two reviewers (JDW, PGS) will independently screen all index articles provided by the SATIRE and DISCO study groups to determine the subset of the articles that made subgroup claims in the abstract. Three additional reviewers will arbitrate all potential discrepancies (JFT, KS, JPAI). In cases where an ambiguous subgroup claim was made at the abstract level (e.g., where a subgroup difference was mentioned but an outcome has not been specified), we will consult the full text of the article to identify all potentially relevant subgroup findings that fit within the scope of the abstract-level claim.

Evaluation of heterogeneity and tests of interaction

For each subgroup claim in the abstract, we will determine whether a statistically significant *P*-value is reported from a test for interaction in the abstract or only in the full text. When statistically significant *P*-values are provided or qualitative statements about a statistically significant subgroup difference are made, we will not perform any additional calculations. For claims without clear evidence of statistical heterogeneity, two reviewers (JDW, PGS) will extract the relative or absolute effect sizes, confidence intervals, standard errors, or any other available data that could be used to calculate subgroup-level effect sizes and standard errors. In cases where the index articles do not provide effect measures for the subgroups of interest, we will use 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 neither relative or absolute effect measures were reported in the text, we will calculate relative effect measures, since the measure of interaction a multiplicative scale are more often assessed and reported based on logistic and Cox models in randomized studies[16,17] An online digitizer will be used to extract approximate values from figures when needed. When exact calculations are not possible, two reviewers (KS, JDW) will discuss the information provided 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 subgroup findings with at least one corroboration attempt, we extracted the overall treatment effect, the subgroup level effects, and whether the interactions were quantitative (where the size of the effect differs in the subgroups but the direction is the same), or qualitative (where the intervention is beneficial in one subgroup and harmful in another).[11]

The subgroup-level effect sizes and standard errors for the index articles with subgroup claims will be entered into R (version 3.2.3) and the metafor package was used to formally test for heterogeneity using Cochran's Q-test with DerSimonian and Laird fixed effect weighting. When index articles reported hazard ratios, we used RevMan (version 5.4) to formally test for heterogeneity. A third investigator (KS) will review all of the subgroup claim classifications and re-evaluate the test for interactions applying the Altman and Bland method.[18]

Data extraction

For each index article with at least one subgroup claim in the abstract, we will record the first author, year of publication, journal, and sample size randomized. We will also extract the compared intervention(s), population assessed, and the outcome(s) pertaining to each individual subgroup claim. We will note 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 made in the abstracts of papers where a statistically significant *P*-value from a test for interaction was reported, the total number of claims where there was not enough information provided in the full text to formally test for subgroup heterogeneity, the total number of claims where there was a qualitative statement in the full text indicating that the subgroup claim was a subgroup finding (e.g., "the interaction term was statistically significant"), and the total number of subgroup findings based on our calculations using the metafor package in R or RevMan.

Evaluation of subgroup finding corroboration

We will use Scopus between November 2015 and July 2016 (all searches updated July 2016) to search for all English language reviews and research articles citing each of the eligible index articles with at least one subgroup finding. One reviewer (JDW) screened the title and abstract of all citing articles to determine the citing RCTs and meta-analyses. All RCTs and meta-analyses were downloaded and screened by two reviewers (JDW, PGS) for evidence of subgroup corroboration attempts. Three additional investigators (JFT, KS, JPA) will arbitrate any uncertainties.

Extractions for meta-analyses and RCTs with corroboration attempts

For individual RCTs and meta-analyses that evaluated the subgroup finding from the index article, we will extract the first author, journal, year of publication, and whether there is any overlap in authorship between the index article and a citing RCT or meta-analysis. When there are several meta-analyses and RCTs on the same corroboration, we will select the most inclusive one.

For any meta-analysis citing an index article with a subgroup finding, we will look for evidence of a subgroup effect included in a forest plot. When available, we will extract 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, and the summary effect size and 95% CI by fixed effects or by random effects in each pertinent subgroup level. In meta-analyses without detailed information about the evaluation of the subgroup finding, we will describe what information was provided related to the corroboration attempt.

For any subsequent RCT evaluating a subgroup finding established by the index article it was citing, we will look for evidence of a test for interaction and whether the results were statistically significant. We will also note whether the authors stated that they had performed the subgroup analysis, but did not report any additional data.

eText. Additional information about subgroup corroborations

Dexamethasone for suspected bacterial meningitis (BM)

In the 2007 index article comparing dexamethasone with placebo among patients with suspected bacterial meningitis (BM), the authors included a pre-specified subgroup analysis comparing the relative risk of death at 1 month grouped according to diagnosis (definite versus probable bacterial meningitis (BM)).[19] A statistically significant *P*-value for heterogeneity of the treatment effect was provided in the text, based on a Cox regression model with interaction terms (*P* = 0.01). The index article implied that there was a qualitative interaction, with lower risk of death at 1 month for patients with definite BM (relative risk 0.43 (95% CI 0.20 to 0.94) and a possibly increased risk of death for patients with probable BM (relative risk 2.65 (95% CI 0.73 to 9.63)). The reported overall relative risk was 0.79 (95% CI 0.45 to 1.39). The study only provided relative risks, total counts for both subgroup levels combined, and Kaplan-Meier (KM) survival curves according to the subgroup levels. Two meta-analyses made full corroboration attempts. In a 2010 meta-analysis[20], only individual patient data from five randomized, double-blind, placebo-controlled trials were used to establish whether any subgroups of patients with acute BM might benefit from adjunctive dexamethasone. The authors only provided combined odds ratios, using Mantel-Haenszel statistics, and a *P*-value for the test for subgroup differences *(P* = 0.23), but did not include any study-level data. Since only pooled odds ratios were provided, we compared the results from the 2010 meta-analysis to the odds ratio for the index article from the forest plot provided by the older 2009 meta-analysis[21] (definite BM: odds ratio 0.46 (95% CI 0.19 to 1.09), probable BM: odds ratio 1.47 (95% CI 0.46 to 4.67); interaction *P* = 0.1166). This suggested that the subgroup claim from the index article was a non-statistically significant qualitative interaction, when odds ratios are used instead of relative risks. When all of the 2010 meta-analysis data were considered, the subgroup level effect measures were attenuated (definite BM: odds ratio 0.90 (95% CI 0.72 to 1.14), probable BM: odds ratio 1.29 (95% CI 0.75 to 2.21) with interaction *P* = 0.28). There was also no evidence of an overall treatment effect (odds ratio 0.95 (95% CI 0.77 to 1.17). Among the three individual RCTs included in the earlier meta-analysis, none of the interaction *P*-values were statistically significant.

Intravenous cooling for nontraumatic out-of-hospital cardiac arrest

In the 2007 index article comparing standard care with intravenous cooling to standard care without intravenous cooling among patients resuscitated by paramedics for nontraumatic out-of-hospital cardiac arrest, the authors reported that the secondary outcome of awakening tended towards improvement in ventricular fibrillation (VF) patients versus non-VF patients randomized to in-field cooling.[22] A subsequent RCT, with at least one of the same co-authors, also included information for the outcome of awakening, but there was no evidence of an interaction test and there were no differences between the VF and non-VF subgroups.[23]

N-terminal BNP-guided heart failure (HF) therapy

In the index article comparing the 18-month outcomes of N-terminal BNP-guided versus symptom guided heart failure (HF) therapy, the authors reported that HF therapy guided by N-terminal BNP improved outcomes in patients aged 60 to 75 years but not in patients older than 75 years.[24] Evidence from tests for interaction (from Cox regression models adjusted for baseline characteristics) were provided for the outcomes of survival free of any hospitalization (interaction $P = 0.02$), mortality (interaction $P = 0.01$), and survival free of hospitalization for HF (interaction $P = 0.01$). Hazard ratio's (HR) for overall survival were presented and indicated a quantitative interaction (Age <75 years HR = 0.41 (95% CI 0.19 to 0.87) versus ≥75 years HR = 0.88 (95% CI 0.54 to 1.44)). We when reevaluated subgroup heterogeneity using inverse variance fixed effects weighting, there was actually no evidence of a statistically significant quantitative interaction ($P = 0.10$). The overall treatment effect for the outcome of overall survival indicated a non-statistically significant HR that favors N-Terminal BNP-Guided Therapy over Symptom-Guided Therapy (HR = 0.70 (95% CI 0.47 to 1.06)). The most recent meta-analysis[25] from 2015 that focused specifically on the effects of N-terminal BNP-guided therapy on outcomes in subgroups based on individual patient data provided summary HRs and 95% CIs for both age groups and a non-statistically significant *P*-value from a test for interaction for the outcome of mortality (<75 years: HR 0.68 (95% CI 0.48 to 0.96), ≥75 years: HR 0.87 (95% CI 0.65 to 1.16); interaction $P = 0.22$). The authors did not include a forest plot with individual trial data. Both subgroup level effect measures were attenuated towards the null HR value of 1.0. When we calculated the overall HR, we found a marginal beneficial treatment effect (HR 0.78 (0.62 to 1.10)).

In an older meta-analysis[26] from a similar group of authors, a statistically significant interaction between age and treatment efficacy for mortality was reported (<75 years: HR 0.62 (0.45 to 0.85), ≥75 years: HR 0.98 (0.75 to 1.3); interaction $P = 0.028$). While the authors of the meta-analysis also stated that there was also no statistically significant interaction with age for the outcome of heart failure hospitalizations, both of these corroboration attempts cannot be considered "pure" because the meta-analysis also included patients with LVEF >45% while the index article excluded these patients. In a meta-analysis[27] published in 2015, subgroup analyses were performed for all-cause mortality, all-cause hospitalization, and heart failure related hospitalization, but the age groups were based on different cut-off values (<72 years versus ≥72 years of age). In a meta-analysis[28] from 2013, a separate subgroup analysis was performed for the same age groups, but based on a composite outcome of all-cause mortality and HF-related hospitalization. Lastly, in a 5-year follow-up from the index study, the long-term effects of N-terminal BNP-guided therapy were still more favorable in patients aged 60 to 74 years, however there was no evidence of a statistically significant interaction for hospital-free survival, HF hospital-free survival, and overall survival.[29]

Group therapy for metastatic or locally recurrent breast cancer

In the 2007 index article comparing supportive expressive group therapy (education materials plus weekly supportive expressive group therapy) for the primary outcome of survival among women with confirmed metastatic or locally recurrent breast cancer, the authors reported that a hormone estrogen receptor (ER) status (positive versus negative) by treatment interaction was statistically significant *(*$P = 0.002$), indicating that ER-negative participants randomized to treatment survived longer.[30] One commentary published in 2008 attempted to corroborate the subgroup finding using data from another RCT.[31] The authors of the commentary stated that the subgroup results from the index RCT were based on a small subsample (n = 25 ER-negative women), and an exact replication with a larger sample size (n = 70 ER-negative women) did not find a significant receptor status by treatment interaction (P = 0.71). The authors concluded that the subgroup results from the index article were likely a chance finding.[31]

Low-glycemic load for young adults with obesity

A SATIRE RCT compared low-glycemic load (40% carbohydrate and 35% fat) and low-fat (55% carbohydrate and 20% fat) diets for the primary outcomes of body weight, body fat percentage, and cardiovascular disease risk factors among young adults with obesity.[32] Statistically significant interactions for insulin concentration at 30 minutes during a 75-gram oral glucose tolerance test (above or below the median of 57.5 uIU/ml) by diet for the outcomes of body weight *(P*=0.02) and body fat percentage at 18 months *(P* = 0.01) were reported. A few subsequent RCTs tested for a diet by insulin status interaction for weight loss.[33,34] However, none of these RCTs used insulin concentration at 30 minutes as the measurement of insulin status.

Epoetin alfa to prevent red-cell transfusion

In the topic where epoetin alfa was compared to placebo for the primary outcome of percentage of patients who received a red-cell transfusion, the authors reported two similar subgroup findings without providing *P*-values from interaction tests.[35] Based on our own calculations using data extracted from the text to perform the test for heterogeneity, mortality at day 140 had a statistically significant *P*-value (0.03) and mortality at day 29 had a non-statistically significant *P*-value of 0.05. Two meta-analyses from 2007[36] and 2013[37] had subgroup analyses that were partial corroboration attempts for the outcome of mortality, but they both used data at day 29 instead of day 140. In the index RCT, the *a priori* established subgroup analyses were based on three mutually exclusive admission subgroups (trauma, surgical non-trauma, and medical non-trauma patients). Both meta-analyses attempting to corroborate the subgroup finding from the index article based their calculations on a binary comparison of trauma versus non-trauma patients. While the 2013 meta-analysis reported a statistically significant *P*-value (0.002) for subgroup differences from a *post hoc* analysis of trauma versus non-trauma patients, there was no forest plot, study level effect measures, or further information about the studies contributing data to the non-trauma subgroup level.[37] Among the five studies for the trauma subgroup, one was a Russian article that could not be located, one was an earlier RCT with the same first author as the first author of the index article reporting the subgroup finding, one was an abstract, one was a matched case control study, and one was an RCT from 2008 that only included trauma patients.[37] The 2007 meta-analysis attempting to corroborate the subgroup finding from the index article provided an effect size for the trauma subgroup, a *P*-value from a test for heterogeneity was not reported. Lastly, this meta-analysis only included data from the two RCTs with the same first author as the first of the index article.[36] Both meta-analyses warned that the results from the subgroup analyses were driven by the data from the two RCTs by the first author of the index article and that the results should be interpreted with caution.[36,37]

Escalated-dosing for acute myeloid leukemia or high-risk refractory anemia

In the index article comparing cytarabine plus daunorubicin at the conventional-dose or at an escalated-dose among patients newly diagnosed with acute myeloid leukemia or high-risk refractory anemia, the authors reported that patients in the escalated-treatment group who were 60 to 65 years of age had higher rates of remission, event-free survival, and overall survival compared with the patients in the same age group who received the conventional dose.[38] Although this was an unclear claim at the abstract level, the authors specified that *post hoc* tests for interaction were performed according to age in the methods (60-65 years, 66-70 years, >70 years). While a significant p-value for interaction was provided for rates of complete remission, the authors included a qualitative statement that tests for an interaction between age and treatment were also significant for overall survival and the primary outcome, event-free survival. The only meta-analysis that investigated the subgroup finding from the index article used a dichotomous age subgroup (>65 years versus <65 years) and could not be considered a full corroboration attempt.[39] Furthermore, there was one trial included in this meta-analysis that contributed data to >65 years age group for the outcomes of complete remission and overall survival.[39]

REFERENCES

1. 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.

2. Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. *Lancet.* 2005;365(9471):1657-1661.

3. Sun X, Briel M, Busse JW, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *Bmj.* 2012;344:e1553.

4. 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.

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

6. 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.

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. 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.

9. Lagakos SW. The challenge of subgroup analyses--reporting without distorting. *N Engl J Med.* 2006;354(16):1667-1669.

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

11. 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(1):93-98.

12. 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.

13. 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.

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

15. 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.

16. 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.

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

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

19. 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.

20. van de Beek D, Farrar JJ, de Gans J, et al. Adjunctive dexamethasone in bacterial meningitis: a meta-analysis of individual patient data. *Lancet Neurol.* 2010;9(3):254-263.

21. Vardakas KZ, Matthaiou DK, Falagas ME. Adjunctive dexamethasone therapy for bacterial meningitis in adults: a meta-analysis of randomized controlled trials. *Eur J Neurol.* 2009;16(6):662-673.

22. Kim F, Olsufka M, Longstreth WT, 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 degrees C normal saline. *Circulation.* 2007;115(24):3064-3070.

23. 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.

24. Pfisterer M, Buser P, Rickli H, et al. 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.

25. Brunner-La Rocca HP, Eurlings L, Richards AM, et al. Which heart failure patients profit from natriuretic peptide guided therapy? A meta-analysis from individual patient data of randomized trials. *Eur J Heart Fail.* 2015;17(12):1252-1261.

26. Troughton RW, Frampton CM, Brunner-La Rocca HP, et al. Effect of B-type natriuretic peptide-guided treatment of chronic heart failure on total mortality and hospitalization: an individual patient meta-analysis. *Eur Heart J.* 2014;35(23):1559-1567.

27. Xin W, Lin Z, Mi S. Does B-type natriuretic peptide-guided therapy improve outcomes in patients with chronic heart failure? A systematic review and meta-analysis of randomized controlled trials. *Heart Fail Rev.* 2015;20(1):69-80.

28. Savarese G, Trimarco B, Dellegrottaglie S, et al. Natriuretic peptide-guided therapy in chronic heart failure: a meta-analysis of 2,686 patients in 12 randomized trials. *PLoS One.* 2013;8(3):e58287.

29. Sanders-van Wijk S, Maeder MT, Nietlispach F, et al. Long-term results of intensified, N-terminal-pro-B-type natriuretic peptide-guided versus symptom-guided

treatment in elderly patients with heart failure: five-year follow-up from TIME-CHF. *Circ Heart Fail.* 2014;7(1):131-139.

30. 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.

31. Kissane D, Li Y. Effects of supportive-expressive group therapy on survival of patients with metastatic breast cancer: a randomized prospective trial. *Cancer.* 2008;112(2):443-444; author reply 444.

32. 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.

33. Gardner CD, Offringa LC, Hartle JC, Kapphahn K, Cherin R. Weight loss on low-fat vs. low-carbohydrate diets by insulin resistance status among overweight adults and adults with obesity: A randomized pilot trial. *Obesity (Silver Spring).* 2016;24(1):79-86.

34. Klemsdal TO, Holme I, Nerland H, Pedersen TR, Tonstad S. Effects of a low glycemic load diet versus a low-fat diet in subjects with and without the metabolic syndrome. *Nutr Metab Cardiovasc Dis.* 2010;20(3):195-201.

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

36. Zarychanski R, Turgeon AF, McIntyre L, Fergusson DA. Erythropoietin-receptor agonists in critically ill patients: a meta-analysis of randomized controlled trials. *CMAJ.* 2007;177(7):725-734.

37. Mesgarpour B, Heidinger BH, Schwameis M, et al. Safety of off-label erythropoiesis stimulating agents in critically ill patients: a meta-analysis. *Intensive Care Med.* 2013;39(11):1896-1908.

38. Löwenberg B, Ossenkoppele GJ, van Putten W, et al. High-dose daunorubicin in older patients with acute myeloid leukemia. *N Engl J Med.* 2009;361(13):1235-1248.

39. Gong Q, Zhou L, Xu S, Li X, Zou Y, Chen J. High Doses of Daunorubicin during Induction Therapy of Newly Diagnosed Acute Myeloid Leukemia: A Systematic Review and Meta-Analysis of Prospective Clinical Trials. *PLoS One.* 2015;10(5):e0125612.

eTable. Characteristics of Subgroup Findings With Partial Corroboration Attempts

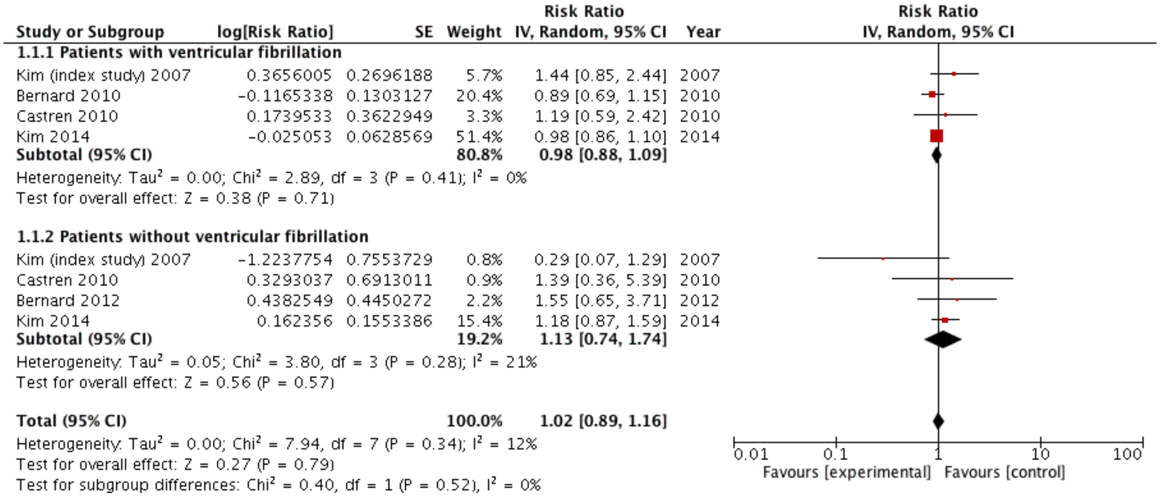
Comparison	**Subgroups index article**	**Population characteristics**	**Outcome (primary)**	**P-value for interaction**	**Corroboration characteristics**	**P-value for interaction significant from any corroboration**
Epoetinalfa vs. placebo (2007)[6]	Patients diagnosed with trauma vs. surgical non-trauma vs. medical non-trauma	Medical, surgical, or trauma patients between 48 and 96 hours after admission to the ICU.	Mortality day 29 (no)	0.05*	Both MAs, used data from mortality at day 29 and subgroups were trauma vs. non-trauma patients	MA 1: Provided *P*-value from interaction test (0.002) but only reported effect estimate for trauma patients.[7] MA2: Only reported effect estimate for trauma.[8]
			Mortality day 140 (no)	0.03*		
Low-glycemic load vs. low-fat diet (2007)[9]	Insulin concentration at 30 minutes after a dose of oral glucose (group x time x insulin concentration)	Obese young adults aged 18-35 y	Body weight (yes)	0.02	RCT 1: Used INS-AUC as instead of insulin-30	Authors stated no significant interaction between diet assignment and INS-AUC detected.[10]
					RCT 2: Different population , different intervention	Authors stated they found similar significant interaction when subjects divided according to median fasting insulin concentration but that data were not shown.[11]
Cytarabine plus daunorubicin at conventional dose vs. cytarabine + gaunorubicin at escalated dose[12]	60-65 y vs. 66-70 y vs. >70 y	Newly diagnosed acute myeloid leukemia or high risk refractory anemia patients 60 to 83 y of age	Complete remission	0.02	MA 1: Used age <65 y old vs. >65 y old	Authors provided non-significant heterogeneity *P*-value. The only study with data for both subgroup levels is the index article.[13]
			Overall survival	Qualitative statement of significant interaction	MA 1: Used age 18-65 y old vs. 65-83 y old	Authors provided non-significant heterogeneity *P*-value. The only study with data for both subgroup levels is the index article.[13]

Abbreviations: ICU = intensive care unit, MA = meta-analysis, vs = versus, y = years
* Not provided by the authors, calculated by us based on odds ratios

REFERENCES

1. Kim F, Olsufka M, Longstreth WT, 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 degrees C normal saline. *Circulation.* 2007;115(24):3064-3070.
2. Bernard SA, Smith K, Cameron P, et al. Induction of therapeutic hypothermia by paramedics after resuscitation from out-of-hospital ventricular fibrillation cardiac arrest: a randomized controlled trial. *Circulation.* 2010;122(7):737-742.
3. 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.
4. 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.
5. Bernard SA, Smith K, Cameron P, et al. Induction of prehospital therapeutic hypothermia after resuscitation from nonventricular fibrillation cardiac arrest*. *Crit Care Med.* 2012;40(3):747-753.
6. Corwin HL, Gettinger A, Fabian TC, et al. Efficacy and safety of epoetin alfa in critically ill patients. *N Engl J Med.* 2007;357(10):965-976.
7. Mesgarpour B, Heidinger BH, Schwameis M, et al. Safety of off-label erythropoiesis stimulating agents in critically ill patients: a meta-analysis. *Intensive Care Med.* 2013;39(11):1896-1908.
8. Zarychanski R, Turgeon AF, McIntyre L, Fergusson DA. Erythropoietin-receptor agonists in critically ill patients: a meta-analysis of randomized controlled trials. *CMAJ.* 2007;177(7):725-734.
9. 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.
10. Gardner CD, Offringa LC, Hartle JC, Kapphahn K, Cherin R. Weight loss on low-fat vs. low-carbohydrate diets by insulin resistance status among overweight adults and adults with obesity: A randomized pilot trial. *Obesity (Silver Spring).* 2016;24(1):79-86.
11. Klemsdal TO, Holme I, Nerland H, Pedersen TR, Tonstad S. Effects of a low glycemic load diet versus a low-fat diet in subjects with and without the metabolic syndrome. *Nutr Metab Cardiovasc Dis.* 2010;20(3):195-201.
12. Löwenberg B, Ossenkoppele GJ, van Putten W, et al. High-dose daunorubicin in older patients with acute myeloid leukemia. *N Engl J Med.* 2009;361(13):1235-1248.
13. Gong Q, Zhou L, Xu S, Li X, Zou Y, Chen J. High Doses of Daunorubicin during Induction Therapy of Newly Diagnosed Acute Myeloid Leukemia: A Systematic Review and Meta-Analysis of Prospective Clinical Trials. *PLoS One.* 2015;10(5):e0125612.

eFigure. Intravenous Cooling for Nontraumatic Out-of-Hospital Cardiac Arrest for the Outcome of Being Discharged Alive From Hospital



Study or Subgroup	log[Risk Ratio]	SE	Weight	Risk Ratio IV, Random, 95% CI	Year
1.1.1 Patients with ventricular fibrillation					
Kim (index study) 2007	0.3656005	0.2696188	5.7%	1.44 [0.85, 2.44]	2007
Bernard 2010	−0.1165338	0.1303127	20.4%	0.89 [0.69, 1.15]	2010
Castren 2010	0.1739533	0.3622949	3.3%	1.19 [0.59, 2.42]	2010
Kim 2014	−0.025053	0.0628569	51.4%	0.98 [0.86, 1.10]	2014
Subtotal (95% CI)			**80.8%**	**0.98 [0.88, 1.09]**	
Heterogeneity: Tau^2 = 0.00; Chi^2 = 2.89, df = 3 (P = 0.41); I^2 = 0%					
Test for overall effect: Z = 0.38 (P = 0.71)					
1.1.2 Patients without ventricular fibrillation					
Kim (index study) 2007	−1.2237754	0.7553729	0.8%	0.29 [0.07, 1.29]	2007
Castren 2010	0.3293037	0.6913011	0.9%	1.39 [0.36, 5.39]	2010
Bernard 2012	0.4382549	0.4450272	2.2%	1.55 [0.65, 3.71]	2012
Kim 2014	0.162356	0.1553386	15.4%	1.18 [0.87, 1.59]	2014
Subtotal (95% CI)			**19.2%**	**1.13 [0.74, 1.74]**	
Heterogeneity: Tau^2 = 0.05; Chi^2 = 3.80, df = 3 (P = 0.28); I^2 = 21%					
Test for overall effect: Z = 0.56 (P = 0.57)					
Total (95% CI)			**100.0%**	**1.02 [0.89, 1.16]**	
Heterogeneity: Tau^2 = 0.00; Chi^2 = 7.94, df = 7 (P = 0.34); I^2 = 12%					
Test for overall effect: Z = 0.27 (P = 0.79)					
Test for subgroup differences: Chi^2 = 0.40, df = 1 (P = 0.52), I^2 = 0%					

Standard care with intravenous cooling versus standard care without intravenous cooling for the outcome of being discharged alive from hospital. (Kim 2007[1], Bernard 2010[2], Castren 2010[3], Kim 2014[4], Bernard 2012[5])