OPEN ACCESS

Sex based subgroup differences in randomized controlled trials: empirical evidence from Cochrane meta-analyses

Joshua D Wallach,[1] Patrick G Sullivan,[1] John F Trepanowski,[2] Ewout W Steyerberg,[3] John P A Ioannidis[4]

[1]Department of Health Research and Policy, and Meta-Research Innovation Center at Stanford (METRICS), Stanford, CA, USA
[2]Stanford Prevention Research Center, Stanford University, Stanford, CA, USA
[3]Department of Public Health, Erasmus MC, Rotterdam, Netherlands
[4]Departments of Medicine, Health Research and Policy, and Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, CA 94305, USA

Correspondence to: J P A Ioannidis
jioannid@stanford.edu

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2016;355:i5826
http://dx.doi.org/10.1136/bmj.i5826

Accepted: 20 October 2016

ABSTRACT

OBJECTIVE

To evaluate the frequency, validity, and relevance of statistically significant (P<0.05) sex-treatment interactions in randomized controlled trials in Cochrane meta-analyses.

DESIGN

Meta-epidemiological study.

DATA SOURCES

Cochrane Database of Systematic Reviews (CDSR) and PubMed.

ELIGIBILITY CRITERIA FOR STUDY SELECTION

Reviews published in the CDSR with sex-treatment subgroup analyses in the forest plots, using data from randomized controlled trials.

DATA EXTRACTION

Information on the study design and sex subgroup data were extracted from reviews and forest plots that met inclusion criteria. For each statistically significant sex-treatment interaction, the potential for biological plausibility and clinical significance was considered.

RESULTS

Among the 41 reviews with relevant data, there were 109 separate treatment-outcome analyses ("topics"). Among the 109 topics, eight (7%) had a statistically significant sex-treatment interaction. The 109 topics included 311 randomized controlled trials (162 with both sexes, 46 with males only, 103 with females only). Of the 162 individual randomized controlled trials that included both sexes, 15 (9%) had a statistically significant sex-treatment interaction. Of four topics where the first published randomized controlled trial had a statistically significant sex-treatment interaction, no meta-analyses that included other randomized controlled trials retained the statistical significance and no meta-analyses showed statistical significance when data from the first published randomized controlled trial were excluded. Of the eight statistically significant sex-treatment interactions from the overall analyses, only three were discussed by the CDSR reviewers for a potential impact on different clinical management for males compared with females. None of these topics had a sex-treatment interaction that influenced treatment recommendations in recent guidelines. UpToDate, an online physician-authored clinical decision support resource, suggested differential management of men and women for one of these sex-treatment interactions.

CONCLUSION

Statistically significant sex-treatment interactions are only slightly more frequent than what would be expected by chance and there is little evidence of subsequent corroboration or clinical relevance of sex-treatment interactions.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Subgroup analyses are often performed and differential treatment effects are sometimes claimed among patient subgroups

Subgroup differences might offer insights on how to optimize individualized treatment, but they might also be spurious

There is a lot of interest about sex differences in treatment effects, and stratification by sex is common in clinical trials

WHAT THIS STUDY ADDS

An assessment of 109 topics shows that significant sex-treatment interactions from Cochrane reviews are only slightly more common than what would be expected by chance

Meta-analyses rarely corroborate sex based subgroup findings from individual randomized trials

Statistically significant sex-treatment interactions typically have limited biological plausibility or clinical significance

Introduction

Subgroup analyses in randomized controlled trials are commonly used to determine whether treatment effects vary across certain patient characteristics, such as whether an effect is different between males and females.[1-5] It has been proposed that results from these analyses can be used to tailor patient care ("stratified medicine" and "precision medicine").[6-8] In particular, male and female subgroups are often compared for their responses to a broad range of interventions owing to differences that might exist between the sexes in physiology, pharmacokinetics, and pharmacodynamics.[9-12] For example, it is speculated that women might respond differently from men to some drugs and might have more adverse events in response to certain drugs.[9-15] Though there is substantial interest about sex differences in treatment effects, and stratification by sex is common in clinical trials, relatively little is known about how often subgroup differences are valid and how frequently the findings of sex based subgroups go on to affect specific clinical practice. Clearly, if subgroup differences are such that only men or only women deserve treatment with some intervention, this would have major implications for clinical practice.

There are many reasons to suspect a lack of validity in subgroup claims. Trials often perform numerous subgroup analyses without correcting for multiple testing, thereby increasing the probability of false positive claims of a true difference between subgroups.[1 16 17] Furthermore, subgroup analyses are often not prespecified, which can increase the reporting of spurious findings.[4 18-20] Even when an analysis plan is prespecified for subgroups in randomized controlled trials, more than 90% of the trials deviate from the protocol.[18]

Considering the likelihood of spurious findings from subgroup testing, external replication studies that would contribute evidence to meta-analyses are necessary to determine the validity of subgroup differences.[20] Though there is evidence that pure replication studies in the biomedical literature are rare,[21] it is unclear how often subgroup findings are externally corroborated by pooled evidence from meta-analyses. Lastly, formal tests of statistical interaction need to be done systematically to assess initial claims of subgroup differences and their corroboration.[4 20 22 23] We used data from the Cochrane Database of Systematic Reviews (CDSR) to evaluate the frequency, validity, and clinical impact of sex related subgroup differences for diverse outcomes across a large number of interventions assessed in randomized controlled trials.

Methods

See the supplementary file for the study protocol.

Identification of relevant meta-analyses

To identify the search terms necessary to locate sex based subgroup analyses in the forest plots of the meta-analyses published in the CDSR, we performed a pilot PubMed search using the terms: "The Cochrane database of systematic reviews"[Journal] AND (gender OR sex OR male OR female OR men OR women OR man OR woman AND subgroup*)". Two independent reviewers (JDW, JFT) then selected and reviewed a random sample of articles in full text to establish the most common terminology used for sex or gender subgroup analyses in the data and analysis section and the forest plots of Cochrane meta-analyses. Cochrane representatives then performed an automated search on 30 March 2016 for the terms "gender" OR "sex" OR "men" OR "women" OR "male" OR "female" in the "Data and Analysis" section of Cochrane reviews published between issue 1, 1995, and issue 2, 2016, and extracted the review group, title, and CD number for all of the search results.

Article and forest plot screening

Three reviewers (JDW, JFT, PGS) screened all of the articles located by the CDSR search to identify those that contained at least one sex based subgroup analysis. First, we excluded duplicate studies, withdrawn studies, and studies where the title and abstract indicated a clear sex specific outcome (eg, menopause for women). Next they examined all forest plots from the remaining articles to determine if they contained a sex based subgroup analysis, or any of the aforementioned search terms in their title, because reviews might present a sex based subgroup analysis across separate forest plots for men and women rather than a single subanalysis within the same forest plot. When a forest plot for males matched a forest plot for females on intervention and outcome within a given study, we considered the pair of forest plots as a single sex based subgroup analysis. We excluded forest plots with data from non-randomized controlled trials (eg, quasirandomized, observational studies). Two reviewers (JDW, JPAI) arbitrated all potential discrepancies.

Data extraction

For all eligible forest plots that passed both initial stages of screening, one reviewer (JDW) manually extracted several characteristics: study population; treatment interventions compared; outcome; number of randomized controlled trials on men overall, men only, women overall, women only, and both men and women; total sample size of individual randomized controlled trials (for men, women, and both separately); effect measure used in each separate analysis (eg, odds ratio, risk ratio, risk difference, rate ratio, mean difference, standardized mean difference, Peto odds ratio); methods used for data synthesis (eg, fixed effects (Mantel-Haenszel, inverse variance, Peto) or random effects (Mantel-Haenszel, inverse variance)); number of randomized controlled trials that included both men and women (trials with separate data on each sex); and, when available, P value from the χ^2 test for subgroup differences. Two additional independent reviewers (PGS, JFT) checked the extracted data for accuracy.

On compiling the eligible evidence, we came across instances when a single CDSR meta-analysis might present multiple forest plots with similar outcomes for the same intervention. Therefore to avoid redundancy we devised criteria to prevent the same data from being included across multiple analyses. When a single meta-analysis presented forest plots evaluating sex-treatment interactions with identical or nearly identical outcomes on the same intervention, we only selected one comparison (eg, frequency of drinking measured as quantity of drinking in grams per week, frequency of drinking measured as the number days drinking per week, and intensity of drinking measured as the number of grams per drinking day were deemed to be nearly identical and thus correlated outcomes[24]). When the population, outcome, or eligibility criteria for one eligible analysis was a subset of the population, outcome, or eligibility criteria from a separate forest plot representing a similar yet broader meta-analysis, we selected the primary analysis described in the text of the article (eg, we selected the outcome of "incidence lung cancer" because it was a primary outcome and a subset of "incidence of all cancers," which was excluded[25]). If there was no clear primary analysis, multiple outcomes were considered primary analyses, or primary analyses were not specified in the text, we retained the analysis with the most data (larger number of trials, or, when there was a tie, smaller variance in the summary effect). In cases where a subgroup analysis was presented for the same intervention but with multiple distinct outcomes, we kept all of the intervention-outcome pairs, but made note of how many outcomes there were for each comparison. All eligible intervention-outcome analyses (ie, at the forest plot level rather than the article level) will hereafter be referred to as "topics."

As suggested during peer review of our work, we also give the proportion of statistically significant sex-treatment interactions considering only one topic in each CDSR review with relevant data and the proportion of statistically significant sex-treatment interactions considering only one outcome for each

doi: 10.1136/bmj.i5826 | *BMJ* 2016;355:i5826 | the bmj

comparison in each CDSR review, when a review had data on multiple different comparisons of interventions. When multiple topics still existed for the same review or comparison in our sample, we further selected the one with the largest available sample size (or, in the case of ties, we selected the one with the largest number of events or the smallest variance in the summary effects, if counts were not provided). These analyses avoid the potential lack of independence among some topics and use the topics with maximal power to detect subgroup differences.

Statistical analysis

RevMan (version 5.4) was used to recreate the forest plots identified by our search and to test the sex-treatment interaction using the same meta-analysis methods as those presented in the original review in the CDSR. Because many of the eligible forest plots contained randomized controlled trials performed on one sex, we recreated the forest plots and evaluated the sex-treatment interaction including only the randomized controlled trials that include data for both men and women. The sex-treatment interaction was also evaluated for each individual randomized controlled trial with data for both sexes.

We used our best judgment to determine whether each individual study from a forest plot represented a single randomized controlled trial or information pooled from multiple randomized controlled trials. When a forest plot only provided pre-pooled information from multiple trials (eg, summary hazard ratios based on individual participant data from multiple trials), we only calculated the overall sex-treatment interaction for the pooled data. In cases where the entries in a forest plot did not include a study date, we checked the reference section of the CDSR review to determine whether multiple studies had been pre-pooled or whether a study date could be established. We treated multiple pre-pooled entries without a study date as "grouped studies" and tested the sex-treatment interaction for the grouped entries. When two or more randomized controlled trials from the same topic were published in the same journal issue, or data from multiple randomized controlled trials were grouped in a previous article, we counted the trials as one trial. When multiple randomized controlled trials from the same topic were published in the same year, we counted them together as representing the first trial evidence, since the window of opportunity for corroboration was too short.

For each topic we recorded whether a nominally statistically significant (P<0.05) sex-treatment interaction was seen in the overall meta-analyses and in the meta-analysis based only on the randomized controlled trials that included data for both men and women. Furthermore, we assessed whether the first published randomized controlled trial included in a meta-analysis had a significant sex-treatment interaction and whether this difference occurred in any other randomized controlled trial with data on both sexes. We then recorded whether the significant sex-treatment interaction that occurred in the first published randomized controlled trial, or in any individual randomized controlled trial, was corroborated by the summary of data from all randomized controlled trials in the same meta-analysis. We were particularly interested in the first published randomized controlled trial because this was likely the first time that a certain sex-treatment interaction was proposed. It is of interest to note whether individual subsequent randomized controlled trials corroborate the hypotheses proposed by the first published randomized controlled trial. Furthermore, if the first published randomized controlled trial had a statistically significant sex-treatment interaction it might be more likely that subsequent randomized controlled trials performed the same analyses.

To determine whether standardization of the forest plots affected the interpretation of the results, we recreated all of the forest plots using a random effects (DerSimonian and Laird) model with standard effect measures. The DerSimonian and Laird random effects model allows for treatment effects to vary across studies and is often used for clinical trials.[26 27] We selected the risk ratio and the mean difference because they were the most commonly reported effect measures for binary and continuous data, respectively. Hazard ratios and rate ratios were not transformed.

We noted how many P values from the test for subgroup differences were less than 0.05. As a sensitivity analysis for significant sex-treatment interactions, we also used the Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis in R using the metafor package,[28] because this method generates lower mean error rates than the standard DerSimonian-Laird method.[29]

Analysis of clinical relevance of significant sex-treatment interactions

For statistically significant (P<0.05) sex-treatment interactions, two independent investigators (JDW, PGS) extracted the comparison, population, outcome, overall effect size, effect size for males, effect size for females, number of randomized controlled trials, number of randomized controlled trials that included both sexes, and interaction P value. We also examined the full text of the respective CDSR review and noted those presenting evidence of biological plausibility and clinical relevance. As a non-prespecified objective, we then subjectively considered the biological plausibility and the potential clinical relevance of the sex-treatment effect, particularly whether it might translate to a difference in clinical management between subgroups. For eligible topics with biological plausibility and the potential for clinical relevance, we performed non-systematic searches and scrutinized recent guidelines for evidence suggesting differential clinical management based on the sex-treatment effects.

Analysis of general subgroup reporting practices within the CDSR

Finally, we used PubMed to identify and screen a random sample of 100 reviews drawn from the CDSR ("the Cochrane Database of Systematic Reviews"[Journal], search performed on 30 May 2016). This was done to

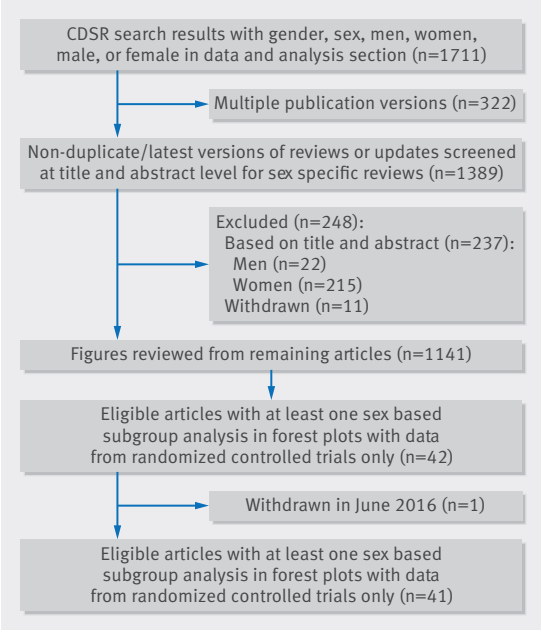

Fig 1 | Flow diagram of included studies. CDSR=Cochrane Database of Systematic Reviews

describe general sex related subgroup reporting practices by the CDSR and to account for the possibility that subgroup analyses might be planned in the data and analyses section, but not conducted, reported, or visualized as a forest plot. We excluded reviews that were withdrawn, included outcomes that only pertained to one specific sex, and did not list randomized trials as one of the selection criteria. Since all topics with a sex-treatment interaction from our Cochrane search came from forest plots that included only randomized trials, we made note of the reviews with data from only randomized trials in the forest plots. We included studies where some of the primary outcomes focused exclusively on females if any of the outcomes in the review also pertained to neonates or infants, since both sexes could be represented.

As a non-prespecified outcome suggested during peer review, we also investigated whether CDSR reviews that report some eligible sex subgroup data do this only for a subset of the relevant trials, while they report information without stratification by sex on all trials. We examined this possibility by focusing on one topic per review (selected as described) and estimating what proportion of total sample size was also represented in the sex subgroup analyses.

Patient involvement

No patients were involved in the development of the research question, development of outcome measures, design, or conduct of this study. There are no plans to involve patients in the dissemination of the results of this study.

Results

Search results

The search performed on the CDSR identified 1711 articles for review at the title and abstract level. Of these 1711 reviews, 322 were duplicate publications, 215 had outcomes that pertained only to females, 22 had outcomes that pertained only to males, and 11 were reviews that had been withdrawn, leaving 1141 articles requiring visual inspection of all forest plots. Out of these 1141 articles, 41 included data from randomized controlled trials that had sex based subgroups as determined by screening of the forest plots (fig 1). A total of 23 (56%) of the 41 reviews had more than one eligible topic, and there were a total of 109 eligible topics with sex based subgroup data.

Characteristics of included meta-analyses

Supplementary table 1 shows the characteristics of the reviews with sex based subgroup information contained in the forest plots. The 109 topics included a total of 311 trials (162 with data on both men and women, 46 with data only on men, and 103 with data only on women).

The median total sample size for each topic was 923 (interquartile range 225-2807), with 306 males (110-1600) and 483 females (115-1481). The median number of randomized controlled trials on males and females, among the 106 topics without individual patient data, were the same (n=1, interquartile range 1-2) (three topics reported hazard ratios based on pooled individual patient data, where information was not available at the level of single randomized controlled trials).

The most frequently reported effect measures among the 109 topics were the mean difference or mean change difference (n=42, 39%), risk ratio (n=25, 23%), and Peto odds ratio (n=21, 19%). Most analyses used inverse variance weighting (n=48, 44%) or Mantel-Haenszel methods (n=37, 34%). Just under half (n=54, 50%) of the 109 topics included only a single randomized controlled trial providing data for both males and females, and another four (4%) topics had only one randomized controlled trial with exclusively males and one with exclusively females.

Frequency of significant sex-treatment interactions

Since the results from the non-standardized and standardized forest plots were almost identical, we presented the results from the standardized calculations, using a random effects (inverse variance) model for a risk ratio or mean difference effect measure. When we recalculated the sex-treatment interactions for the topics with more than one randomized controlled trial using the Hartung-Knapp-Sidik-Jonkman method, the two eligible topics remained statistically significant.

Among the 109 topics, eight (7%) had a statistically significant sex-treatment interaction. When we recreated the forest plots using only the topics with at least one randomized controlled trial with data for both men and women or those with individual participant data, six (6%) out of 96 had a significant sex-treatment interaction. Overall, 162 individual randomized controlled trials could be tested for a sex-treatment interaction. Of these, 15 (9%) had a statistically significant sex-treatment interaction.

When considering only the 39 topics with multiple randomized controlled trials and at least one trial with

doi: 10.1136/bmj.i5826 | *BMJ* 2016;355:i5826 | the bmj

data for both men and women (eg, one randomized controlled trial with data for men and women, one with data for men only, and one with data for females only), three (8%) had significant sex-treatment interactions. When the 39 individual randomized controlled trials with data on only one sex were excluded and only the trials with data for both men and women were included, four (10.3%) had a significant sex-treatment interaction. Two of the topics had only one randomized controlled trial with data for both males and females, and one of these topics had two randomized controlled trials that were published in the same edition of a journal.[30]

Even when we selected only one topic with sex subgroup data in each CDSR review, there were only four (10%, 4/41) statistically significant sex-treatment interactions. When we allowed for multiple comparisons from each CDSR review, with only one outcome selected for each comparison, there were seven (11%, 7/61) statistically significant sex-treatment interactions.

Table 1 summarizes the proportion of statistically significant sex-treatment interactions when different eligibility criteria are used to select eligible topics.

Corroboration of sex-treatment interactions by subsequent trials

Among the 39 topics with more than one randomized controlled trial and at least one randomized controlled trial with data for both men and women, 106 trials (or grouping of randomized controlled trials) could be tested for a sex-treatment interaction. Of these 106 trials, 12 (11%) had a statistically significant sex-treatment interaction. Four of these 12 were the first published trial in the respective study. Of these four randomized controlled trials, three were from topics with more than one randomized controlled trial with data for both men and women (n=2 had only two randomized controlled trials with data for both men and women and n=1 had four randomized controlled trials with data for both men and women). Two of the statistically significant sex-treatment interactions from the first published randomized controlled trial were tested at a later date by randomized controlled trials with at least one of the same authors from the first published randomized controlled trial. None of the subsequent individual randomized controlled trials with data for both men and women corroborated the statistically significant sex-treatment interaction from the first published randomized controlled trial. For the five studies attempting to corroborate the statistically significant sex-treatment interaction from the first published randomized controlled trial from their respective review, the median sample size was 886 (interquartile range 734-1810).

When the four forest plots were recreated excluding the first randomized controlled trial with a statistically significant sex-treatment interaction, none of the remaining randomized controlled trials validated the statistically significant results. Supplementary table 2 provides additional information about the four statistically significant sex-treatment interactions from the first published randomized controlled trials that were not corroborated by cumulative meta-analyzed data, among the topics with more than one randomized controlled trial with data for both men and women.

Clinical relevance

The eight statistically significant sex-treatment interactions (table 2) did not present a consistent pattern in populations, disease of interest, interventions, or outcomes.[25 30-36] Three of the eight topics contained data from only a single randomized controlled trial. Five of the eight topics had effect sizes in the same direction for males and females. In four of the eight topics, the reviews explicitly stated that they planned to investigate gender or sex subgroups. There were two topics[31 35] where both biological rationale and clinical relevance were discussed in the review, and one topic[34] where only clinical relevance was discussed. There was one topic where the authors suggested a potential for biological rationale, but noted that no conclusions should be made based on one study.[36] However, none of these three topics with a discussion on clinical relevance found sex-treatment interactions that were consistent with recent guidelines. UpToDate, a physician-authored clinical decision online support resource, suggested differential management for men and women for one of the topics (perform surgery in men but not in women with symptomatic 50-69% carotid stenosis).[37]

For example, in the topic where systematic screening was compared with routine practice for the primary outcome of detection of new cases of atrial fibrillation among patients aged more than 40, the authors included some discussion of both biological rationale and clinical

Table 1 | Summary results for proportion of statistically significant sex-treatment interactions based on different eligible criteria

Eligibility criteria	No of topics (No of trials)	No (%) of statistically significant sex-treatment interactions
All	109 (311)	8/109 (7)
Only topics with data for both men and women	96 (162)	6/96 (6)
Only topics with >1 RCT and at least one RCT with data for both men and women*	39 (209)	3/39 (8)
As above, but excluding RCTs with data on only one sex*	39 (106)	4/39 (10)
One topic per review (most inclusive topic with the most data)	41† (164)	4/41 (10)
One topic per treatment comparison (most inclusive topic with the most data)	61‡ (194)	7/61 (11)

RCT=randomized controlled trial.
*Excluding topics based on individual patient data, where trial level information was not provided.
†One topic only had individual participant data and no trial level information.
‡Two topics only had individual participant data and no trial level information.

Table 2 | Topics with statistically significant sex-treatment interactions

Comparison	Population characteristics	Outcome	Primary outcome*	Effect size in women (95% CI)	Effect size in men (95% CI)	RCT women: men:both†	P value	P values for other outcomes	Biologic/clinical rationale‡	Plan§
Surgery (CEA) *v* no surgery	Recent symptomatic carotid stenosis (ie, TIA or minor ischemic stroke)	5 year risk of stroke and any stroke or death within 30 days	Unclear	RR 0.85 (0.58 to 1.23)	RR 0.54 (0.44 to 0.67)	0:0:2¶	0.04	NA	Both	Yes
Vitamin C *v* placebo	Healthy	Incidence of lung cancer	Yes	RR 1.84 (1.14 to 2.95)	RR 0.94 (0.64 to 1.38)	1:1:0	0.03	NA	Neither	Unclear
Sildenafil *v* placebo	Sexual dysfunction from antidepressant use	ASEX total scores	No	MD −0.50 (−2.24 to 1.24)	MD −4.62 (−6.29 to −2.95)	0:1:1	0.001	0.49	Unclear clinical rationale	Yes
Non-latex *v* latex condom**	Sexually active couples††	Medical event‡‡	Unclear	RR 0.50 (0.35 to 0.71)	MD 0.92 (0.57 to 1.48)	0:0:1	0.04	0.50	Neither	No
Risperidone *v* olanzapine	Schizophrenia§§	Change in prolactin level from baseline	No	MD 41.40 (29.64 to 53.16)	MD 19.91 (13.64 to 26.18)	0:1:1	0.002	NA	Clinical	No
Fibrates *v* control	Previous cardiovascular disease	Non-fatal stroke, non-fatal MI, and vascular death	Yes	RR 0.30 (0.16 to 0.56)	RR 0.83 (0.73 to 0.94)	1:2:1	0.002	NA	Neither	Yes
Systematic screening *v* routine practice	Age ≥40 years	Detection of new cases of atrial fibrillation	Yes	RR 0.98 (0.59 to 1.61)	RR 2.64 (1.50 to 4.66)	0:0:1	0.01	NA	Both	Yes
All tea *v* control	Healthy or high risk of cardiovascular disease	HDL cholesterol	No	MD −0.19 (−0.42 to 0.04)	MD 0.27 (−0.11 to 0.65)	0:0:1	0.04	0.34, 0.39, 0.07	Some possible biologic rationale	No

RCT=randomized controlled trial; CEA=carotid endarterectomy; HDL=high density lipoprotein TIA=transient ischemic attack; RR=risk ratio; NA=not applicable; MD=mean difference; ASEX=Arizona Sexual Experience; MI=myocardial infarction.
*Is the outcome for the topic specified as the primary outcome in the text of the review?
†Ratio of number of RCTs with data for women only to number of RCTs with data for men only and number of RCTs with data on both sexes.
‡Does the Cochrane Database of Systematic Reviews (CDSR) article from which this topic was drawn present evidence about the biologic plausibility of the sex-treatment interaction or the clinical relevance of the sex-treatment interaction?
§Is there a statement in the CDSR review that the sex based subgroup analysis was planned a priori?
¶This topic had two trials where the final reports were published in the same year; thus they are both considered as "first trials" and we combined their results.
**Baggy Tactylon versus Aladan condom.
††Sexually active couples engaged in heterosexual, vaginal intercourse.
‡‡Medical event per condom, defined as any genital problem that remains for less than 24 hours.
§§Schizophrenia and schizophrenia-like psychosis.

significance for the management of men versus women. The authors noted that men are 1.5 times more likely than women to develop atrial fibrillation, which could make screening more effective in men.[38] Furthermore, the review notes that there could be differences in the uptake of screening programs among men and women, which could ultimately impact clinical outcomes. Overall, the authors noted that the observed difference could be a result of a higher prevalence of atrial fibrillation among men and a greater rate of participation.[35] However, the results are based on a single randomized controlled trial, and additional evidence is necessary. The 2014 guideline from the American Heart Association/American College of Cardiology/Heart Rhythm Society for the management of patients with atrial fibrillation does not even mention screening,[39] and the 2012 focused update of the European Society of Cardiology guidelines for the management of atrial fibrillation only recommends opportunistic screening in patients aged 65 years and older, and it does not differentiate between men and women.[40] UpToDate recommends against screening.[41]

In another example, a statistically significant subgroup finding was seen for surgery (carotid endarterectomy) compared with no surgery in patients with recent symptomatic carotid stenosis for the primary outcome of five year cumulative risk of ipsilateral carotid ischemic stroke or any stroke or death within 30 days.[31] The authors of the review suggest that sex is a clinically important effect modifier given the lower risk of ischemic stroke during medical treatment and higher operative risk in women.[31] Other authors have offered a mechanistic explanation as to why anatomical differences could affect carotid endarterectomy.[42] Although the most recent American Heart Association/American Stroke Association guidelines for the prevention of stroke do not make a distinction between the treatment of men and women,[43] UpToDate suggests differential management for men and women with recently symptomatic carotid stenosis of 50-69%.[37]

Finally, in the topic where risperidone was compared with olanzapine for the adverse effect of change in prolactin levels from baseline, the authors included a discussion about clinical significance.[34] Antipsychotic drug treatments are known to cause increases in plasma prolactin levels, and for women, hyperprolactinemia could be related to severe menstrual disorders.[44] Although the review does not discuss any biological rationale, the authors state that clinicians might want to consider the different tolerability profiles of risperidone and olanzapine.[34]

General sex based subgroup reporting practices within the CDSR

Among the first 100 randomly selected reviews drawn from the CDSR that did not include reviews that had been withdrawn, reviews with outcomes that pertained to one specific sex, and reviews that did not list randomized trials as one of the selection criteria, 12 (12%) did not identify any studies. Among the 88 remaining reviews, forest plots in 83 (94%) only included data from randomized trials.

doi: 10.1136/bmj.i5826 | *BMJ* 2016;355:i5826 | the bmj

Among these 83 reviews, three (4%) mentioned sex subgroup analyses under the data and analysis section and did not include any sex subgroup testing in the results or in a forest plot. Two reviews stated that the subgroup analyses could not be performed[45 46] and one review[47] separated trials with less than 45% female participants compared with other trials.

Among the 83 reviews, only one[48] had a forest plot with sex based subgroup data from individual patient data. Since this review did not discuss any sex or gender subgroup analyses in the data and analysis section, it was not captured by our search. The post hoc sex-treatment interaction reported was borderline statistically significant (P=0.049) and the authors stated that there was some suggestion of men benefitting more on a relative scale than women from chemotherapy for the outcome of survival.

Finally, we compared the total sample size in the sex stratified forest plots compared with the non-stratified analyses for the same comparison and outcome in the 41 topics (selecting the one with more evidence when multiple topics existed in the same CDSR review). We were able to gather information about sample size for 40 of the 41 topics. We found that the average proportion of the total sample size that was included in the subgroup analyses was 78% (interquartile range 53-100%). The four topics that had statistically significant sex interactions were based on 11%, 55%, 100%, and 100% of the respective total sample size for the same comparison and outcome.

Discussion

Our empirical evaluation of statistically significant sex-treatment interactions from the CDSR revealed only eight (7%) statistically significant sex-treatment interactions among 109 topics. This is not much beyond what would be expected by chance alone. With only eight statistically significant interactions, it is likely that the number of false positives outnumbered the number of true positives. Also, certain reviews had more than one topic, which could lead to an overlap of topics with non-independent data. However, even when we selected only one topic for each review or allowed for multiple comparisons with one outcome per review, the statistically significant sex-treatment interactions would still be uncommon (4/41 (10%) and 7/61 (12%), respectively), not far from what is expected from chance. None of the sporadically observed statistically significant sex-treatment interactions has resulted in a clear difference in clinical management for men compared with women.

Many of the assessments of sex related subgroup differences were based on a single trial for each topic. Thus, even when a significant difference was found, there was no evidence of independent corroboration. Previous research suggests that replication studies are rare,[21] and it is common for clinical topics of interest to have evidence that comes from only a single trial.[49] When multiple trials were published on the same topic, we found that sex based differences in the first trial were not corroborated by the meta-analysis combining the data from all relevant trials. Furthermore, they were never validated by the meta-analysis excluding the data from the first trial with a statistically significant sex-treatment interaction.

The importance of corroborating subgroup claims is illustrated by an example from one topic that we assessed. In a meta-analysis comparing home based with clinic based specimen collection in the management of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* infections among sexually active people for the outcome of uptake, there was no overall sex-treatment interaction based on the two randomized controlled trials that included data for both men and women.[50] Both trials had found statistically significant sex-treatment interactions, but the stratified results were in opposite directions. This serves as an important reminder that healthcare providers should exercise caution and clinical judgment when interpreting subgroup results from a single study. As our study shows, sex based subgroup findings are rarely corroborated at the meta-analysis level and they rarely are included in clinical guidelines or recommendations.

Limitations of this study

Our study has some limitations. First, although we believe we have captured the majority of the sex based subgroup analyses available in the CDSR, the number of reviews we found with these analyses was relatively small, and many of these analyses included only a single randomized controlled trial with data for men and women. The evidence for many topics is therefore underpowered to detect modest differences in effect sizes in men compared with women. However, many modest differences in effect sizes are unlikely to be clinically relevant. Another limitation is that it is possible that the authors of the CDSR reviews sometimes chose to simply present sex subgroup data for descriptive purposes or because it was easier to extract stratified data (eg, if mean differences were reported separately for men and women in the original randomized controlled trial), and they were not interested in testing or reporting a sex-treatment interaction. It is also possible that CDSR reviewers may not have been able to perform sex based subgroup analyses owing to selective reporting in individual randomized controlled trials. When individual trials choose not to present certain data, meta-analyses are prevented from pooling data from all available sources and some differences might be masked. In our survey we assumed that all forest plots with data stratified by sex were subgroup analyses. Furthermore, we did not examine sex based subgroup analyses that were not presented in forest plots of the CDSR articles that we screened. To deal with the possibility that some subgroup analyses might have been planned in a CDSR review, but not ultimately performed, reported in the text, or depicted as a forest plot, we took a random sample of reviews from Cochrane reviews in PubMed to evaluate subgroup practices in Cochrane reviews. In the random sample, all sex or gender subgroup analyses that were proposed in the data analysis section of the text were not included in the forest plots and the

authors stated in the results or discussion sections that subgroup analyses could not be performed. Among the 88 reviews with forest plots that only included data from randomized trials, there was only one study with a sex-treatment interaction in a forest plot, and our Cochrane search did not identify this meta-analysis. It is possible that several dozen such reviews might have been missed by our search. However, it is unlikely that sex based analyses that were not clearly referenced in the data and analysis section, and thus missed by our automatic Cochrane search, would have higher chances of being corroborated. It is possible that several non-statistically significant sex-treatment interactions are not being reported in the forest plots of the Cochrane reviews. Furthermore, it could be argued that individual trials might be more likely to present stratified results if the sex-treatment interaction is statistically significant. This could lead to more statistically significant sex-treatment interactions based on pooled data in the meta-analyses. It is also possible that only a subset of trials investigate sex subgroup differences, leading to a reduced power to detect any true subgroup differences. However, we found that this was not a likely explanation, because in most circumstances when sex-treatment interactions had been assessed in eligible CDSR meta-analyses, data had been included by all or almost all relevant trials assessing the comparison of interest. Among the 40 eligible topics, we found that the average proportion of total evidence available for each outcome that was included in the subgroup analyses was 78%. Lastly, in some cases, individual meta-analyses had more than one topic and these might not have been independent. However, analyses restricting to data where such dependence was not an issue showed largely similar results.

Implications for researchers and users of information from meta-analyses

Although subgroup analyses might often be reasonable to explore possible heterogeneity of treatment effects across patient populations or to generate new hypotheses, our study suggests that significant sex-treatment interactions, on the basis of data from single randomized controlled trials and meta-analyses, rarely occur beyond what would be expected by chance. It has been repeatedly shown that subgroup claims on relative effects are often not credible,[20] and we have provided additional evidence of the lack of corroboration and validation of significant sex-treatment interactions from individual randomized controlled trials by subsequent studies and meta-analyses. We should acknowledge that men and women might still have different absolute risks of outcomes of interest, and thus sex might sometimes need to be considered as a prognostic variable, even in the presence of equal treatment effects on the relative scale.[6 51] Both prognostic and predictive factors define the possible benefits of an intervention.[51-54] However, subgroup analyses from meta-analyses can lead to spurious findings.

The US National Institutes of Health has introduced a new policy that asks applicants to "explain how relevant biological variables, such as sex, are factored into research designs and analyses for studies in vertebrate animals and humans."[55 56] It is possible that as a result of this policy, more investigators will routinely perform exploratory sex subgroup analyses without considering biological or clinical relevance. However, the NIH also states that authors should describe how sex might influence the research questions being studied and that authors should perform a review of the relevant literature. If evidence of differences between men and women in previous studies is found by authors, it would then provide justification to consider sex in the research design and data analysis.[55 56] It remains to be seen whether this policy will improve the conduct and documentation of sex subgroup analyses in trials without generating a greater burden of false positive signals in subgroup differences. The Cochrane Sex/Gender Methods Group, which was established in 2005, also promotes the integration of sex or gender analysis in meta-analyses. A sex and gender in systematic review planning tool exists to help reviewers plan and interpret their results. The planning tool outlines that "questions about possible sex and gender differences should be asked and the particular relevance determined or ruled out."[57] Furthermore, authors are told to consider any previous evidence and whether interventions are likely to affect men and women differently.[57] In our sample we found little evidence of Cochrane reviews clearly justifying all sex based subgroup analyses. Our experience also suggests that individual Cochrane meta-analyses often perform multiple subgroup analyses for similar treatments or outcomes.

Individual randomized controlled trials and Cochrane reviews should clearly justify all sex based and other subgroup analyses performed. Many sex-treatment interactions in the forest plots from the CDSR were based on data from only one randomized controlled trial, and researchers and clinicians should view significant sex-treatment interactions from Cochrane reviews as exploratory, unless there is substantial evidence of corroboration, biological plausibility, and clinical relevance. We are not arguing that researchers should no longer undertake sex based subgroup analyses in general, but that sex based subgroups should be tested only when there is a priori credible biological rationale and some expectation for clinical relevance. This might not be a common occurrence.

Conclusion

Statistically significant sex-treatment interactions from Cochrane reviews do not occur much more often than what would be expected by chance alone, and meta-analyses rarely corroborate individual randomized controlled trials with significant sex-treatment interactions. Authors, research consumers, and journal reviewers and editors should carefully scrutinize the credibility of subgroup analyses.[58] Isolated sex based subgroup analyses are simply hypothesis generating.

We thank Ruth Foxlee (Information Specialists, Cochrane Editorial Unit, Cochrane Central Executive) and Rasmus Moustgaard (Cochrane

Informatics and Knowledge Management Department) for searching the Cochrane Database of Systematic Reviews (CDSR) and providing guidance on CDSR search procedures.

Contributors: JDW, PGS, JFT, EWS, and JPAI designed the study. JDW, PGS, and JT screened the Cochrane reviews for eligibility. JDW extracted the data. PGS and JFT checked all extracted data for accuracy. JDW and PGS performed the statistical analysis. JDW, PGS, JFT, and JPAI interpreted the results. JDW wrote the first draft of the manuscript, with major contributions from PGS, JFT, and JPAI. All authors made revisions to the manuscript and have read and approved the final version. JPAI is the guarantor. All authors had full access to all of the data.

Funding: METRICS (Meta-Research Innovation Center at Stanford) is supported by a grant from the Laura and John Arnold Foundation. PGS received support from the Stanford Clinical and Translational Science Award to Spectrum (NIH UL1 TR 001085). JFT is supported by the NIH (T32 HL007034). EWS is partly supported by the NIH (PRICES project, U01 NS086294). JPAI is supported by an unrestricted gift by Sue and Bob O'Donnell to Stanford Prevention Research Center. The research was conducted independent of any involvement of the funders. Funders were not involved in any aspect of the study design, data collection, data interpretation, writing, or the decision to submit the article for publication.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: METRICS (Meta-Research Innovation Center at Stanford) is supported by a grant from the Laura and John Arnold Foundation, PGS received support from the Stanford Clinical and Translational Science Award to Spectrum (NIH UL1 TR 001085), JFT is supported by the NIH (T32 HL007034), EWS is partly supported by the NIH (PRICES project, U01 NS086294), and JPAI is supported by an unrestricted gift by Sue and Bob O'Donnell to Stanford Prevention Research Center; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Data sharing: Data are available from the lead author on request.

Transparency: The guarantor (JPAI) affirms that the manuscript is a honest, accurate, and transparent account of the study bring reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/3.0/.

1 Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355:1064-9. doi:10.1016/S0140-6736(00)02039-0.
2 Bhandari M, Devereaux PJ, Li P, et al. Misuse of baseline comparison tests and subgroup analyses in surgical trials. *Clin Orthop Relat Res* 2006;447:247-51. doi:10.1097/01.blo.0000218736.23506.fe.
3 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;151:257-64. doi:10.1016/j.ahj.2005.04.020.
4 Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. *N Engl J Med* 1987;317:426-32. doi:10.1056/NEJM198708133170706.
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;357:2189-94. doi:10.1056/NEJMsr077003.
6 Hingorani AD, Windt DA, Riley RD, et al. PROGRESS Group. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ* 2013;346:e5793. doi:10.1136/bmj.e5793.
7 Hamburg MA, Collins FS. The path to personalized medicine. *N Engl J Med* 2010;363:301-4. doi:10.1056/NEJMp1006304.
8 Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;372:793-5. doi:10.1056/NEJMp1500523.
9 Cotreau MM, von Moltke LL, Greenblatt DJ. The influence of age and sex on the clearance of cytochrome P450 3A substrates. *Clin Pharmacokinet* 2005;44:33-60. doi:10.2165/00003088-200544010-00002.
10 Kashuba AD, Nafziger AN. Physiological changes during the menstrual cycle and their effects on the pharmacokinetics and pharmacodynamics of drugs. *Clin Pharmacokinet* 1998;34:203-18. doi:10.2165/00003088-199834030-00003.
11 Legato MJ. Gender and the heart: sex-specific differences in normal anatomy and physiology. *J Gend Specif Med* 2000;3:15-8.
12 Schwartz JB. The influence of sex on pharmacokinetics. *Clin Pharmacokinet* 2003;42:107-21. doi:10.2165/00003088-200342020-00001.
13 Drici MD, Clément N. Is gender a risk factor for adverse drug reactions? The example of drug-induced long QT syndrome. *Drug Saf* 2001;24:575-85. doi:10.2165/00002018-200124080-00002.
14 Makkar RR, Fromm BS, Steinman RT, Meissner MD, Lehmann MH. Female gender as a risk factor for torsades de pointes associated with cardiovascular drugs. *JAMA* 1993;270:2590-7. doi:10.1001/jama.1993.03510210076031.
15 Xhyheri B, Bugiardini R. Diagnosis and treatment of heart disease: are women different from men?*Prog Cardiovasc Dis* 2010;53:227-36. doi:10.1016/j.pcad.2010.07.004.
16 Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002;21:2917-30. doi:10.1002/sim.1296.
17 Lagakos SW. The challenge of subgroup analyses--reporting without distorting. *N Engl J Med* 2006;354:1667-9. doi:10.1056/NEJMp068070.
18 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. doi:10.1136/bmj.g4539.
19 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. doi:10.1136/bmj.d1569.
20 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.
21 Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JP. Reproducible Research Practices and Transparency across the Biomedical Literature. *PLoS Biol* 2016;14:e1002333. doi:10.1371/journal.pbio.1002333.
22 Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992;116:78-84. doi:10.7326/0003-4819-116-1-78.
23 Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001;357:1191-4. doi:10.1016/S0140-6736(00)04337-3.
24 Kaner EF, Beyer F, Dickinson HO, et al. Effectiveness of brief alcohol interventions in primary care populations. *Cochrane Database Syst Rev* 2007;(2):CD004148.
25 Cortés-Jofré M, Rueda JR, Corsini-Muñoz G, Fonseca-Cortés C, Caraballoso M, Bonfill Cosp X. Drugs for preventing lung cancer in healthy people. *Cochrane Database Syst Rev* 2012;10:CD002141.
26 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177-88. doi:10.1016/0197-2456(86)90046-2.
27 DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. *Contemp Clin Trials* 2015;45(Pt A):139-45.
28 Viechtbauer W. Conducting meta-analyses in R with the metaphor package. *J Stat Softw* 2010;36:1-48.
29 IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* 2014;14:25. doi:10.1186/1471-2288-14-25.
30 Wang D, Liu B, Tao W, Hao Z, Liu M. Fibrates for secondary prevention of cardiovascular disease and stroke. *Cochrane Database Syst Rev* 2015;(10):CD009580.
31 Rerkasem K, Rothwell PM. Carotid endarterectomy for symptomatic carotid stenosis. *Cochrane Database Syst Rev* 2011;(4):CD001081.
32 Taylor MJ, Rudkin L, Bullemor-Day P, Lubin J, Chukwujekwu C, Hawton K. Strategies for managing sexual dysfunction induced by antidepressant medication. *Cochrane Database Syst Rev* 2013;(5):CD003382.
33 Gallo MF, Grimes DA, Lopez LM, Schulz KF. Non-latex versus latex male condoms for contraception. *Cochrane Database Syst Rev* 2006;(1):CD003550.
34 Komossa K, Rummel-Kluge C, Schwarz S, et al. Risperidone versus other atypical antipsychotics for schizophrenia. *Cochrane Database Syst Rev* 2011;(1):CD006626.
35 Moran PS, Teljeur C, Ryan M, Smith SM. Systematic screening for the detection of atrial fibrillation. *Cochrane Database Syst Rev* 2016;6:CD009586.
36 Hartley L, Flowers N, Holmes J, et al. Green and black tea for the primary prevention of cardiovascular disease. *Cochrane Database Syst Rev* 2013;(6):CD009934.
37 Mohler E III, Fairman R. Management of symptomatic carotid atherosclerotic disease. In: Kasner S, Eidt, JF., Mills, JL., Dashe, JF., Collins, KA., eds. Waltham, MA: UpToDate; 2016.
38 Benjamin EJ, Levy D, Vaziri SM, D'Agostino RB, Belanger AJ, Wolf PA. Independent risk factors for atrial fibrillation in a population-based cohort. The Framingham Heart Study. *JAMA* 1994;271:840-4. doi:10.1001/jama.1994.03510350050036.

39 January CT, Wann LS, Alpert JS, et al. ACC/AHA Task Force Members. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the Heart Rhythm Society. *Circulation* 2014;130:2071-104. doi:10.1161/CIR.0000000000000040.

40 Camm AJ, Lip GY, De Caterina R, et al. ESC Committee for Practice Guidelines (CPG). 2012 focused update of the ESC Guidelines for the management of atrial fibrillation: an update of the 2010 ESC Guidelines for the management of atrial fibrillation. Developed with the special contribution of the European Heart Rhythm Association. *Eur Heart J* 2012;33:2719-47. doi:10.1093/eurheartj/ehs253.

41 Kumar K. Overview of Atrial Fibrillation. In: Zimetbaum P, Saperia, GM., ed. Waltham, MA: UpToDate; 2016.

42 Bond R, Warlow CP, Naylor AR, Rothwell PM. European Carotid Surgery Trialists' Collaborative Group. Variation in surgical and anaesthetic technique and associations with operative risk in the European carotid surgery trial: implications for trials of ancillary techniques. *Eur J Vasc Endovasc Surg* 2002;23:117-26. doi:10.1053/ejvs.2001.1566.

43 Kernan WN, Ovbiagele B, Black HR, et al. American Heart Association Stroke Council, Council on Cardiovascular and Stroke Nursing, Council on Clinical Cardiology, and Council on Peripheral Vascular Disease. Guidelines for the prevention of stroke in patients with stroke and transient ischemic attack: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2014;45:2160-236. doi:10.1161/STR.0000000000000024.

44 Bargiota SI, Bonotis KS, Messinis IE, Angelopoulos NV. The Effects of Antipsychotics on Prolactin Levels and Women's Menstruation. *Schizophr Res Treatment* 2013; 2013:502697.

45 Klimas J, Tobin H, Field CA, et al. Psychosocial interventions to reduce alcohol consumption in concurrent problem alcohol and illicit drug users. *Cochrane Database Syst Rev* 2014;(12):CD009269.

46 Hoe VC, Urquhart DM, Kelsall HL, Sim MR. Ergonomic design and training for preventing work-related musculoskeletal disorders of the upper limb and neck in adults. *Cochrane Database Syst Rev* 2012;(8):CD008570.

47 Van de Laar FA, Lucassen PL, Akkermans RP, Van de Lisdonk EH, Rutten GE, Van Weel C. Alpha-glucosidase inhibitors for type 2 diabetes mellitus. *Cochrane Database Syst Rev* 2005;(2):CD003639.

48 Sarcoma Meta-analysis Collaboration (SMAC). Adjuvant chemotherapy for localised resectable soft tissue sarcoma in adults. *Cochrane Database Syst Rev* 2000;(2):CD001419.

49 Pereira TV, Horwitz RI, Ioannidis JP. Empirical evaluation of very large treatment effects of medical interventions. *JAMA* 2012;308:1676-84. doi:10.1001/jama.2012.13444.

50 Fajardo-Bernal L, Aponte-Gonzalez J, Vigil P, et al. Home-based versus clinic-based specimen collection in the management of Chlamydia trachomatis and Neisseria gonorrhoeae infections. *Cochrane Database Syst Rev* 2015;(9):CD011317.

51 Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials* 2010;11:85. doi:10.1186/1745-6215-11-85.

52 Riley RD, Hayden JA, Steyerberg EW, et al. PROGRESS Group. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med* 2013;10:e1001380. doi:10.1371/journal.pmed.1001380.

53 Italiano A.Prognostic or predictive? It's time to get back to definition!*J Clin Oncol* 2011;29:4718; author reply 4718-9. doi:10.1200/JCO.2011.38.3729.

54 van Klaveren D, Vergouwe Y, Farooq V, Serruys PW, Steyerberg EW. Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. *J Clin Epidemiol* 2015;68:1366-74. doi:10.1016/j.jclinepi.2015.02.012.

55 U.S. National Institutes of Health. Consideration of Sex as a Biological Variable in NIH-funded Research. http://orwh.od.nih.gov/resources/pdf/NOT-OD-15-102_Guidance.pdf

56 U.S. National Institutes of Health. Consideration of Sex as a Biological Variable in NIH-funded Research (NOT-OD-15-102). http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-102.html

57 Doull M, Runnels V, Tudiver S, Boscoe M. Sex and Gender in Systematic Reviews Planning Tool. 2011. methods.cochrane.org/sites/methods.cochrane.org.equity/files/public/uploads/SRTool_PlanningVersionSHORTFINAL.pdf

58 Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ* 2010;340:c117. doi:10.1136/bmj.c117.

Supplementary file: study protocol
Supplementary file: characteristics of the reviews
Supplementary file: additional information about four statistically significant sex-treatment interactions

No commercial reuse: See rights and reprints http://www.bmj.com/permissions Subscribe: http://www.bmj.com/subscribe