BMJ Open

Interpretation of confidence intervals in clinical trials with non-significant results: always helpful, often inadequate

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-017288
Article Type:	Research
Date Submitted by the Author:	12-Apr-2017
Complete List of Authors:	Gewandter, Jennifer; University of Rochester Medical Center, University of Rochester McDermott, M; University of Rochester Kitt, Rachel; University of Rochester Medical Center, Anesthesiology Chaudari, Jenna; University of Rochester Medical Center, Anesthesiology Koch, James; University of Rochester Medical Center, Anesthesiology Evans, Scott; Harvard School of Public Heath, Department of Biostatistics and Center for Biostatistics in AIDS Research, Gross, Robert ; University of Rochester Medical Center, Neurology and Pharmacology & Physiology Markman, John; University of Rochester, Neurosurgery Turk, Dennis; University of Washington School of Medicine Dworkin, Robert; University of Rochester, Richard Bright Renal Unit
Primary Subject Heading:	Medical publishing and peer review
Secondary Subject Heading:	Medical education and training
Keywords:	Clinical trials < THERAPEUTICS, Confidence intervals, Clinical trial interpretation, Reporting

Interpretation of confidence intervals in clinical trials with non-significant results: always helpful, often inadequate

Jennifer S. Gewandter[a], Michael P. McDermott[b], Rachel A. Kitt[a], Jenna Chaudari[a], James G. Koch[a], Scott R. Evans[c], Robert A. Gross[d], John D. Markman[e], Dennis C. Turk[f], Robert H. Dworkin[a]

[a]*Department of Anesthesiology, University of Rochester, 601 Elmwood Ave, Rochester, NY, USA, 14642 (Jennifer S. Gewandter, Assistant Professor; Rachael A. Kitt, Research Assistant; Jenna Chaudari, Research Assistant; James G. Koch, Research Assistant; Robert Dworkin, Professor)*

[b]*Department of Biostatistics and Computational Biology, University of Rochester, 265 Crittenden Blvd, Rochester, NY, USA, 14642 (Michael P. McDermott, Professor)*

[c]*Department of Biostatistics, Harvard University, 651 Huntington Ave., Boston, MA, USA, 02115 (Scott R. Evans, Professor)*

[d]*Departments of Neurology and Pharmacology & Physiology, University of Rochester, 601 Elmwood Ave., Rochester, NY, USA, 14642 (Robert A. Gross, Professor)*

[e]*Department of Neurosurgery, University of Rochester, 2180 S. Clinton Ave., Rochester, NY, USA, 14618 (John D. Markman, Professor)*

[f]*Department of Anesthesiology and Pain Medicine, University of Washington, 1949 NE Pacific St., Seattle, WA, USA, 98195 (Dennis C. Turk, Professor)*

Correspondence to: J. Gewandter jennifer_gewandter@urmc.rochester.edu

Abstract

Objectives Interpretation of confidence intervals (CIs) in randomized clinical trials (RCTs) with treatment effects that are not statistically significant can distinguish between results that are "negative" (the data are not consistent with a clinically meaningful treatment effect) or "inconclusive" (the data remain consistent with the possibility of a clinically meaningful treatment effect). This interpretation is important to ensure that potentially beneficial treatments are not prematurely abandoned in future research or clinical practice based on invalid conclusions.

Design Systematic review of RCT reports published in 2014 in Annals of Internal Medicine, New England Journal of Medicine, JAMA, JAMA Internal Medicine, and Lancet (n = 247).

Results Eighty-five of 99 articles with statistically non-significant results reported CIs for the treatment effect. Only 17 of those 99 articles interpreted the CI. Of the 22 articles in which CIs indicated an inconclusive result, only 4 acknowledged that the study could not rule out a clinically meaningful treatment effect.

Conclusions Interpretation of CIs is important but occurs infrequently in study reports of trials with treatment effects that are not statistically significant. Increased author interpretation of CIs could improve application of RCT results. Reporting recommendations are provided

Article Summary

Strength and limitations of this study

Strengths

- Systematic review, including RCTs published in 6 high impact medical journals.

- Identified a large deficiency in interpretation of CIs in RCTs with results that are not statistically significant.
- Recommendations for reporting and interpreting CIs are provided.

Limitation

- Our interpretation of the CIs was based on the author-specified clinically relevant treatment effect or the treatment effect used in the sample size calculation. We did not attempt to evaluate these assumption.

Introduction

Randomized clinical trials (RCTs) are the gold standard for evaluating the efficacy of medical treatments. However, when a statistically significant treatment effect is not demonstrated (i.e., the p-value for the primary analysis is not less than or equal to the pre-specified significance level), the estimate of the treatment effect and the p-value alone do not allow the reader of an RCT report to distinguish between the following two possibilities: (1) the treatment does not have a clinically meaningful effect, or (2) the study is unable to rule out a clinically meaningful treatment effect with a high degree of confidence (i.e., the results of the trial would best be described as "inconclusive").[1-6] However, trials for which the effect of treatment on the primary outcome variable is not statistically significant have often been called "negative" and presented as though they support the conclusion that the experimental treatment lacks efficacy.[3] This can result in premature abandonment of potentially beneficial treatments clinically and in future research programs.

For decades, biostatisticians and others have encouraged the use of confidence intervals (CIs) as a means to present the range of treatment effects consistent with the observed data and to evaluate whether RCT results that are not statistically significant suggest that the experimental treatment is ineffective or instead that the trial results are inconclusive (Figure 1).[1-6] Inconclusive results should not be used to inform clinical practice or treatment guidelines.

Previous reviews have assessed CI reporting in publications of preclinical and clinical studies within specific medical specialties.[7-14] To our knowledge, no reviews have examined CI reporting *and* interpretation in RCTs published in high-impact general medical journals.

Methods

Data Sources and Searches

RCTs published in 2014 in Annals of Internal Medicine, British Medical Journal, Journal of the American Medical Association (JAMA), JAMA Internal Medicine, Lancet, and New England Journal of Medicine were identified using PubMed (Appendix 1). Relevant articles were identified following PRISMA guidelines.

Study Selection

Selected articles were primary reports of RCTs that compared the efficacy of at least 2 treatments (1 of which could be a placebo, active comparator, or a wait-list control) using frequentist inferential methods. Trials not evaluating treatments were excluded (e.g., comparison of two cancer screening techniques or the effect of two imaging techniques on surgical decision making). Trials utilizing a non-inferiority or super superiority design were excluded because CIs are interpreted differently for these trials than for standard superiority trials. Dose-finding studies, studies declared to be exploratory in nature, studies focused on safety, and cluster-randomized studies were also excluded. Two authors (RAK and JSG) independently screened all identified articles to determine whether they met the eligibility criteria.

Data Extraction and Quality Assessment

A coding manual was developed to evaluate the frequency with which CIs were reported for the treatment effects in RCTs (Appendix 2). In trials reporting only results for the primary outcome measure(s) that were not statistically significant, coders were asked to evaluate whether the CI for the treatment effect indicated that the data were consistent with the absence of a clinically relevant treatment effect or that the results were inconclusive. This subset of articles included those that reported a statistically significant treatment effect in a subgroup or in analyses that were identified as sensitivity analyses, which were all considered secondary analyses. A treatment effect was considered not statistically significant if the associated p-value was greater than 0.05

unless a different significance criterion was specified by the authors. The coders compared the CI for the treatment effect to a clinically relevant treatment effect declared by the authors at any point in the manuscript or the treatment effect specified in the sample size calculation if no clinically relevant treatment effect was described by the authors. If neither value was provided, the coders did not interpret the CIs for that manuscript. For this comparison, the coders considered the primary analysis if one was identified. If a primary analysis was not identified, the coders considered the first analysis of a primary outcome measure that was reported by the authors. Coders also recorded whether the authors used the CI to interpret any results that were not statistically significant. The coding manual was pre-tested and modified for clarity and content by JSG and RAK in five rounds of three articles each using RCTs published in 2013 that otherwise met the eligibility criteria.

In some cases, the absolute or relative differences in event rates to be detected between groups were reported in the sample size calculation and the results concerning the treatment effect were presented as either a hazard ratio (HR), odds ratio (OR), or relative risk (RR). In these cases, JSG attempted to convert the information provided in the sample size calculation to either the HR, OR, or RR, as appropriate, using some combination of the following: absolute risk reduction ($p_0 - p_1$), relative risk reduction ($(p_0 - p_1)/p_0$), assumed event rate in the control group (p_0), and assumed event rate in the treatment group (p_1). The following formulas were used: $HR = \ln(1 - p_1)/\ln(1 - p_0)$; $OR = (p_1(1 - p_0))/(p_0(1 - p_1))$; and $RR = p_1/p_0$. Such calculations were used to determine ratios representing the clinically relevant treatment effect for 26 articles. Note that the HR calculation yields an estimate that assumes an exponential distribution for the event times.

The data were extracted from each article independently by 2 authors (RAK coded all articles and JSG and JGK each coded approximately half). RAK reviewed the

data for discrepancies and fixed obvious oversights. JSG reviewed any discrepancies due to interpretation and made the final decision on their resolution. JSG also reviewed the final data relating to interpretation of CIs in all of the relevant articles to ensure accuracy.

Results

Trial Characteristics

The final sample included 247 articles (Figure 2). Trial characteristics are presented in Table 1. The articles covered a range of medical specialties; the most common were cardiovascular (22%), infectious disease (15%), and cancer (13%). A little over half of the trials were sponsored, at least in part, by industry (54%).

CI Reporting

Of the 247 included articles, 99 did not report any statistically significant treatment effects on the primary outcome measure. Of those 99, 85 (86%) reported the CI for the treatment effect. Of the 14 articles that did not report the CI for the treatment effect, 6 (42%) reported the CI for the parameter estimate (e.g., mean, event rate) for each group separately. The percentage of articles that reported a CI for the treatment effect in the whole sample (n=247) was similar (85%).

Seven of the 85 articles with no statistically significant treatment effect that also reported a CI for the treatment effect did not report the magnitude of the treatment effect used to estimate the sample size of the study or specify what they would consider to be a clinically relevant treatment effect. Of the remaining 78 articles, 18 specified a clinically relevant treatment effect (6 identified this as a minimal clinically meaningful or important treatment effect; 12 identified this as a clinically meaningful, relevant, significant, important, or worthwhile treatment effect) and in the other 60 articles we interpreted the

trial results based on the treatment effect used to estimate the sample size. We interpreted the nonsignificant results most commonly as falling into two categories: (1) the CI excluded the treatment effect used for the sample size calculation or the author-specified clinically relevant effect (i.e., the data were consistent with no clinically relevant treatment effect) (n=50, 64%), and (2) the CI included the treatment effect used for the sample size calculation or the author-specified clinically relevant effect in favor of the experimental treatment only (i.e., the data could not rule out a clinically meaningful effect of the experimental treatment) (n=20, 26%) (Figures 1 and 3).

Sixty-one (78%) of the 78 articles did not provide any interpretation of the CI for the treatment effect. This percentage was similar for the subset of articles that explicitly identified a primary analysis (i.e., 38 (79%) of 48 such articles did not provide an interpretation of the CI for that analysis). In the 17 (22%) articles that did provide an interpretation of the CI for the treatment effect, the interpretations were of 5 types: (1) consistent with our interpretation, the authors stated that the CI suggested the absence of a clinically meaningful effect (n=8); (2) the authors highlighted the possible treatment effects that were consistent with the CI, but did not speculate on whether those effect sizes were clinically meaningful (n=4); (3) similar to our conclusions, the authors concluded that based on the CI, a clinically meaningful treatment effect could not be ruled out (n=2); (4) the authors conservatively stated that they could not rule out clinically meaningful treatment effects even though the CI excluded the effect size that the trial was designed to detect (n=2); and (5) the authors described the treatment as “modestly effective” and then went on to state that they “focused on the effect size and 95% CI whilst showing p-values, which is in line with the CONSORT 2010 guidelines” when the results were not statistically significant (n=1). We interpreted this trial’s results to be inconclusive (Figure 3).

Discussion

Consistent with widespread recommendations,[1-6] we found that the 85% of articles reporting RCTs published in 6 high-impact medical journals in 2014 reported the CIs for the treatment effect. The percentage of articles that reported CIs in our review was higher than the percentage of articles that reported CIs in previous reviews of RCTs in specialty journals (85% in our review vs. 5% to 66% in previous reviews).[7-14] This increase could be due to the earlier publication periods covered by the previous reviews (i.e., 1990 - 2008). It could also be due to the fact that the 6 journals included in our review require adherence to the CONSORT guidelines,[15] which promote transparent reporting, for publication of RCTs. Regardless of whether the increased reporting of CIs that we observed is in fact due to an effect of time or of the specific journals selected, our results suggest that relatively high quality reporting is possible when required by guidelines, reviewers, and/or editors.

Although reporting CIs provides the reader the ability to make a judgment regarding whether the results are “negative” or inconclusive, such interpretations require an understanding of CIs and knowledge of what should be considered a minimal clinically meaningful treatment effect with respect to the outcome variable used in the trial. Because it cannot be assumed that all readers and stakeholders will have this expertise, or necessarily agree on this point, best reporting practices should include careful interpretation of the CIs and their implications for the conclusions of the trial.

The percentage of articles in our sample that interpreted CIs was much lower than the percentage that simply reported them. Only 17 of the 99 articles that reported analyses of a primary outcome measure that were not statistically significant used a CI to (1) highlight the range of values of the treatment effect that were consistent with the data or (2) discuss whether the trial results were inconclusive or were consistent with the absence of a clinically meaningful treatment effect. Additionally, although the CIs of 22 articles included the treatment effect used for the sample size calculation or the author-

specified clinically relevant treatment effect, only 4 of these articles stated that the study could not rule out a clinically meaningful treatment effect. Our data suggest that many authors do not discuss that the results of their trial can be considered inconclusive on the basis of the CIs they report, perhaps because they believe that doing so might decrease the perceived importance of the RCT. Acknowledging that the study cannot rule out a clinically meaningful effect is important to ensure that clinicians, policy makers, and payers do not inappropriately use the trial results as evidence to suggest that the treatment is ineffective. It must be acknowledged, of course, that readers of RCT reports may not agree on the magnitude of a treatment effect that would be considered clinically meaningful, and that this value may depend on the setting of the trial (e.g., nature of the intervention, trial eligibility criteria, etc.). It is useful to pre-specify in the trial protocol what treatment effect is considered minimally clinically meaningful in order to provide clarity in the interpretation of the trial results. Furthermore, in situations where disagreement regarding what would constitute a clinically meaningful treatment effect may exist in the clinical community, it would be valuable for authors to acknowledge this in their interpretation of the results.

Another method that is sometimes used to interpret the results of RCTs is to present a *post hoc* power calculation. As many authors have correctly argued, however, such a calculation is irrelevant to trial interpretation.[16-18] Encouragingly, only 3 of the included articles with treatment effects on the primary outcome measures that were not statistically significant reported a post hoc power calculation. Three other articles stated that the trials had adequate power without any apparent justification. Post hoc power calculations should be avoided and interpretations regarding whether a trial is “negative” or “inconclusive” would be better based on CIs.

Interestingly, 8% of the 247 included articles reported a CI for the parameter of interest for each separate treatment group (e.g., mean or event rate), but not for the

between-group treatment effect. It is important to emphasize that CIs for the parameters of individual treatment groups are not informative with respect to evaluating whether the results of a trial with a statistically non-significant treatment effect are “negative” as opposed to inconclusive.

A limitation of our review is that we based our interpretation of the CIs reported in the studies with statistically non-significant treatment effects on the author-specified clinically relevant treatment effect or the magnitude of the treatment effect used in the sample size calculation. We did not attempt to evaluate the validity of these values as being of clinical importance because our intention was to evaluate the frequency with which authors used CIs in the interpretation of trial results and whether these interpretations were consistent with their assumptions regarding clinically meaningful treatment effects. Furthermore, the treatment effects used to determine the sample size of a trial are not necessarily what one would consider to be the minimal clinically meaningful treatment effect that investigators might still be pleased to demonstrate.[19] Thus, some trials we interpreted as “negative” might more appropriately have been considered inconclusive if a minimal clinically meaningful treatment effect had been considered. This potential limitation of our results suggests that an even larger percentage of authors may have failed to acknowledge properly that their studies were inconclusive.

It would have been interesting to determine whether the articles that concluded that the trials were “negative” without consideration of CIs actually reported CIs that did not exclude the clinically relevant treatment effect. Unfortunately, we were unable to categorize articles as claiming that the trial was “negative” because authors often had contradictory statements throughout the Discussion regarding whether the “negative” conclusion was definitive. These inconsistencies highlight the importance of using CIs to

interpret whether a trial with a treatment effect that is not statistically significant is "negative" or inconclusive.

In conclusion, the majority of the trials we reviewed reported the CI for the treatment effect, demonstrating relatively high-quality, transparent reporting of RCT results. In contrast, a substantially smaller percentage of articles reporting analyses of the primary outcome measure that were not statistically significant discussed the implications of the CIs of the treatment effect when interpreting the results of their study. We encourage all authors and reviewers to prioritize interpretation of RCT findings using CIs, especially when the CIs indicate that the data cannot rule out a clinically meaningful treatment effect (Table 2). We also encourage readers to consider the CIs when applying the results of RCTs with non-significant results to their clinical practice or research program.

Acknowledgements

We thank Dr. Stephen Senn for his thoughtful review and comments on our manuscript. This article was reviewed and approved by the Executive Committee of the ACTTION public-private partnership with the United States Food and Drug Administration.

Conflict of interest statement

The views expressed in this article are those of the authors and no official endorsement by the Food and Drug Administration (FDA) or the pharmaceutical and device companies that provided unrestricted grants to support the activities of the Analgesic, Anesthetic, and Addiction Clinical Trial Translations, Innovations, Opportunities, and Networks (ACTTION) public-private partnership should be inferred.

The authors have no conflicts of interest related to the work presented in this manuscript to disclose.

Funding Statement

Financial support for this project was provided by the ACTTION public-private partnership, which has received research contracts, grants, or other revenue from the FDA, multiple pharmaceutical and device companies, and other sources.

Contributorship

Jennifer Gewandter contributed to the design of the review and coding. She performed the analyses, wrote the first draft of the manuscript, and incorporated co-author revisions. Robert Dworkin, Michael McDermott, Scott Evans, Robert Gross, John Markman, and Dennis Turk contributed to the design of the study, interpretation of the results, and provided feedback on the manuscript. Rachel Kitt, Jenna Chaudari, and James Koch contributed to coding for the systematic review and provided feedback on the manuscript.

Data Sharing

No additional data available

Exclusive License agreement

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, *a worldwide license,* to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include

within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) license any third party to do any or all of the above.

References

1. Alderson P. Absence of evidence is not evidence of absence. BMJ 2004;328:476-7.
2. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311:485.
3. Freiman JA, Chalmers TC, Smith H, Jr., Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. NEJM 1978;299:690-4.
4. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ 1986;292:746-50.
5. Rothman KJ. A show of confidence. N Engl J Med 1978;299:1362-3.
6. Simon R. Confidence intervals for reporting results of clinical trials. Ann Intern Med 1986;105:429-35.
7. Anttila H, Malmivaara A, Kunz R, Autti-Ramo I, Makela M. Quality of reporting of randomized, controlled trials in cerebral palsy. Pediatrics 2006;117:2222-30.
8. Fan FF, Xu Q, Sun Q, Zhao SJ, Wang P, Guo XR. Assessment of the reporting quality of randomized controlled trials on treatment of coronary heart disease with traditional chinese medicine from the chinese journal of integrated traditional and Western medicine: a systematic review. PLoS One 2014;9:e86360.

9. Kiehna EN, Starke RM, Pouratian N, Dumont AS. Standards for reporting randomized controlled trials in neurosurgery. J Neurosurg 2011;114:280-5.
10. Kloukos D, Papageorgiou SN, Fleming PS, Petridis H, Pandis N. Reporting of statistical results in prosthodontic and implantology journals: p values or confidence intervals? Int J Prosthodont 2014;27:427-32.
11. Naunheim MR, Walcott BP, Nahed BV, Simpson AK, Agarwalla PK, Coumans JV. The quality of randomized controlled trial reporting in spine literature. Spine 2011;36:1326-30.
12. Polychronopoulou A, Pandis N, Eliades T. Appropriateness of reporting statistical results in orthodontics: the dominance of P values over confidence intervals. Eur J Orthod 2011;33:22-5.
13. Urbanic JJ, Lee WR. Confidence intervals and survival estimates: a systematic review of 3 oncology journals. Am J Clin Oncol 2006;29:405-7.
14. Wang P, Xu Q, Sun Q, Fan FF, Guo XR, Guo F. Assessment of the reporting quality of randomized controlled trials on the treatment of diabetes mellitus with traditional chinese medicine: a systematic review. PLoS One 2013;8:e70586.
15. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.
16. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 1994;121:200-6
17. Greenland S. Nonsignificance plus high power does not imply support for the null over the alternative. Ann Epidemiol 2012;22:364-8.
18. Hoenig JM, Heisey DM. An abuse of power: Ther pervasive fallacy of power calculations for data analysis. Am Stat 2001:55;19-24.

19. Senn SJ. Statistical Issues in Drug Development (2nd edition). John Wiley & Sons, Ltd., West Sussex. 2007

Table 1. Trial Characteristics

Characteristic	All articles (N=247)	Articles reporting a treatment effect (TE) that was not statistically significant, the CI of the TE, and a value for the TE that the authors considered to be clinically meaningful (N=78)
Journal		
New England Journal of Medicine	105 (43%)	31 (40%)
JAMA	61 (25%)	22 (28%)
The Lancet	50 (20%)	11 (14%)
British Medical Journal	13 (5%)	8 (10%)
JAMA Internal Medicine	11 (4%)	1 (1%)
Annals of Internal Medicine	7 (3%)	5 (6%)
Design		
Parallel group	245 (99%)	78 (100%)
Cross-over	2 (1%)	0 (0%)
Number randomized	480 (224 – 1195)	730 (311 – 1880)
Medical specialty		
Cardiovascular	55 (22%)	23 (29%)
Infectious disease	38 (15%)	12 (15%)
Cancer	31 (13%)	4 (5%)
Neurology (including pain)	22 (9%)	7 (9%)
Pulmonary	13 (5%)	6 (8%)
Psychiatry	12 (5%)	1 (1%)
Other*	76 (31%)	25 (32%)
Type of intervention		
Treatment	183 (74%)	52 (67%)
Prevention	64 (26%)	26 (33%)
Sponsor		
Industry	134 (54%)	36 (46%)
Other	113 (46%)	42 (54%)

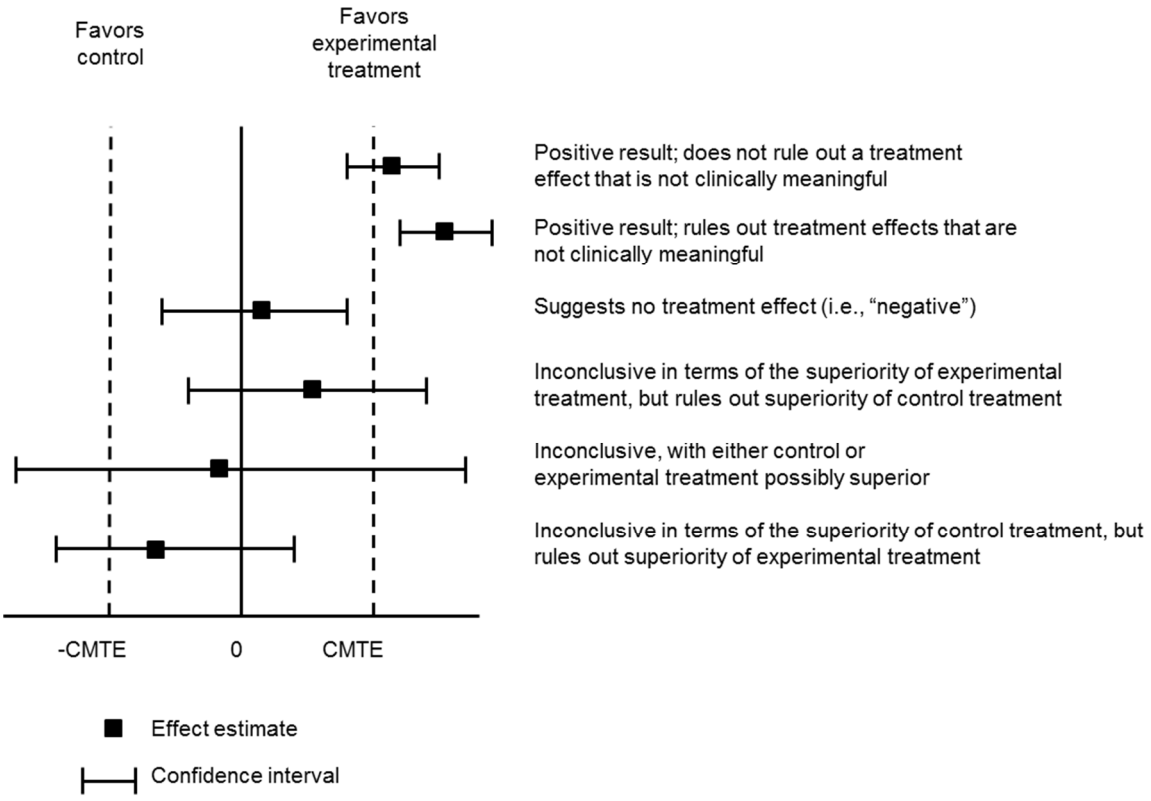
Values are N (%) or median (interquartile range)

*Other includes areas represented by fewer than 10 trials including urology, orthopedics, diabetes, immune disorders, etc.

Table 2. Confidence interval reporting recommendations for RCTs with statistically non-significant results

- Report confidence intervals (CIs) for the treatment effect.
- Discuss interpretation of the CI regarding the magnitude of effects that can be *ruled-out* with reasonable confidence.
- Discuss whether the results suggest a “negative” or “inconclusive” result
- Acknowledge any uncertainty regarding what is considered a clinically meaningful treatment effect on the outcome measure used in the trial

Figure 1. Using confidence intervals to interpret results of RCTs

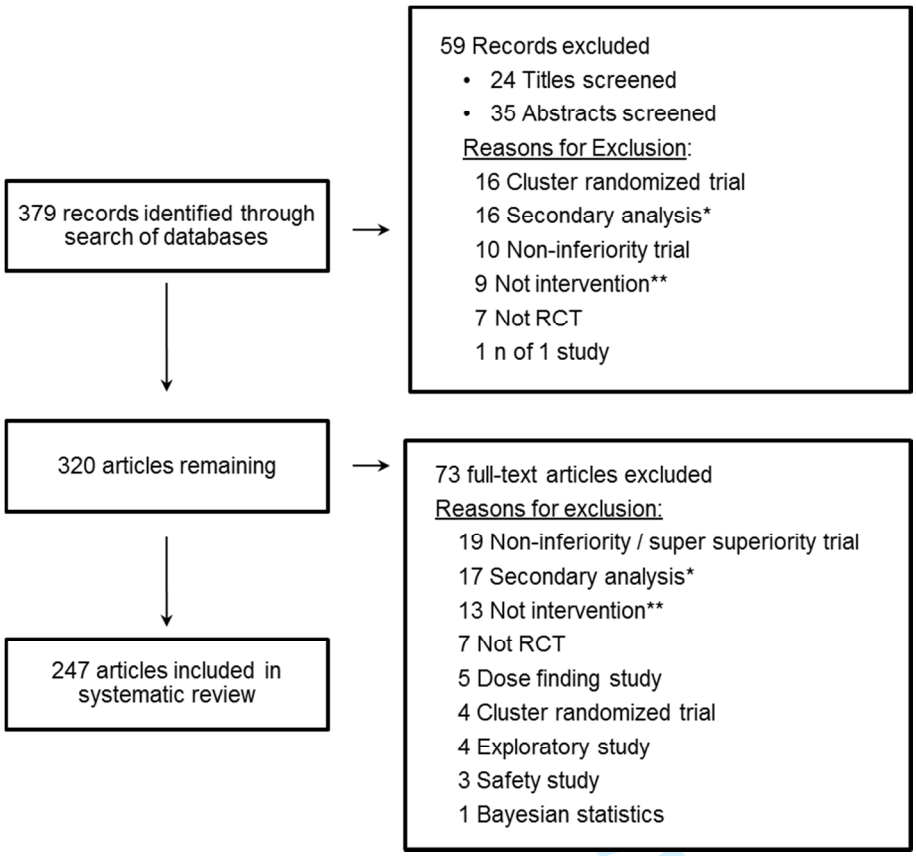


CMTE clinically meaningful treatment effect.

Note that a value of zero indicates no treatment effect in this case; in other cases such as when the treatment effect is quantified using, for example, an odds ratio, hazard ratio, or relative risk, a value of 1.0 would indicate no treatment effect

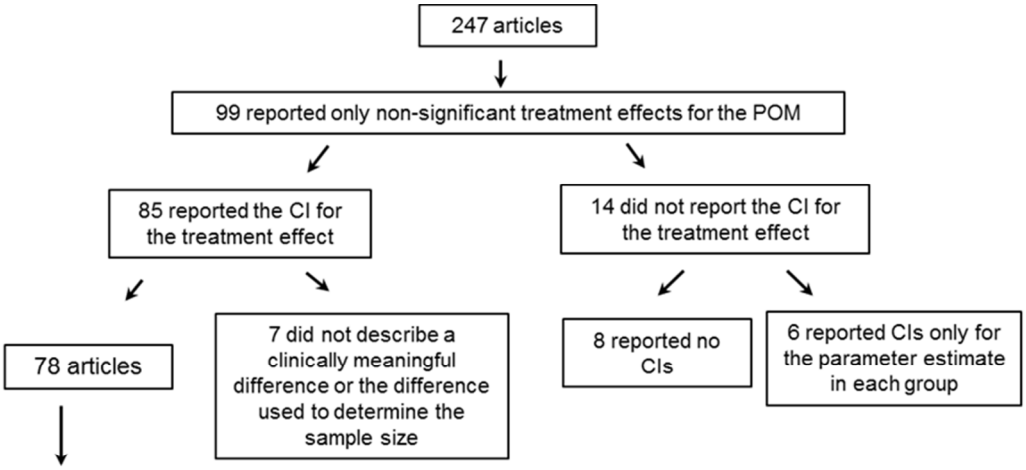
Adapted from reference 19.

Figure 2. PRISMA diagram



RCT Randomized clinical trial
*Secondary analysis of data from a previously reported trial
** RCT examines efficacy of something other than a medical or lifestyle intervention (e.g., a cancer screening method or a diagnostic decision making tool)

Figure 3. Confidence interval reporting and interpretation.



Interpretation based on specified clinically relevant treatment effect or magnitude of the treatment effect used in the sample size calculation	N (%)
CI excludes the clinically relevant effect in favor of both the experimental treatment and the control treatment (i.e., the study suggests no treatment effect)	50 (64%)
CI includes the clinically relevant effect in favor of the experimental treatment (and therefore the study is inconclusive regarding the superiority of the experimental treatment), with superiority of the control treatment excluded	20 (26%)
CI includes the clinically relevant effect in favor of the both the experimental treatment and the control treatment (and therefore the study is inconclusive regarding the superiority of either the control treatment or the experimental treatment)	2 (3%)
CI includes the clinically relevant effect in favor of the control treatment (and therefore the study is inconclusive regarding the superiority of the control treatment), with superiority of the experimental treatment excluded	5 (6%)
CI rules out benefit of 1 of 2 active treatments compared	1 (1%)

17 articles provided some interpretation of the CI

Article interpretation	N (%)
Correctly concludes that the CI suggests the absence of a clinically meaningful treatment effect	8 (47%)
Discusses the CI and points out that the data are consistent with a treatment effect within its bounds, but does not comment on whether the CI includes a clinically meaningful treatment effect	4 (24%)
Correctly concludes that based on the CI, a potentially meaningful treatment effect cannot be ruled out	2 (12%)
States that a clinically meaningful treatment effect cannot be ruled out, even though the CI did not include the treatment effect that the study was designed to detect	2 (12%)
Describes the treatment as “modestly effective” and then goes on to state that they “focused on the effect size and 95% CI whilst showing p-values, which is in line with the CONSORT 2010 guidelines” when the results were not statistically significant	1 (6%)

CI confidence interval; POM primary outcome measure

Appendix 1. Search Strategy.

("Lancet (London, England)"[Jour] OR "BMJ (Clinical research ed.)"[Jour] OR "JAMA"[Jour] OR "The New England journal of medicine"[Jour] OR "JAMA internal medicine"[Jour] OR "Annals of internal medicine"[Jour])

Filter: RCT

Date Search performed: 1/26/2015

“Create Date” Range : 2013-2014

2013 was included to ensure that all articles published in 2014 but created in Pubmed “ahead of print” in 2013 were captured. 744 articles were returned in the search and of those 379 were published in 2014.

Confidence Interval

* 1. Coder

- ◯ Jen
- ◯ Rachel

* 2. Article

* 3. Journal

* 4. Design

- ◯ Parallel
- ◯ Crossover
- ◯ Other (please specify)

* 5. Sponsor

- ☐ Industry
- ☐ Government
- ☐ Professional association / foundation
- ☐ Academic Institution / department funds
- ☐ None (explicitly stated)
- ☐ Not reported
- ☐ Other (please specify)

* 6. Condition

- ◯ Cancer
- ◯ Neurological (e.g., Movement disorders, Cognitive issues)
- ◯ Pain (excluding orthopedic associated pain such as OA)
- ◯ Respiratory
- ◯ Cardiovascular
- ◯ Infectious diseases (e.g., viral and bacterial infections)
- ◯ Obesity
- ◯ Pregnancy outcomes
- ◯ Smoking cessation
- ◯ Gastro/intestinal disorders other than cancer (i.e., GERD, IBS)
- ◯ Diabetes
- ◯ Orthopedic
- ◯ Other (please specify)

7. Is the study for prevention or treatment?

- ◯ Prevention
- ◯ Treatment

Confidence Interval

* 8. Did the authors identify a primary outcome measure?

- ◯ Yes
- ◯ No
- ◯ uncertain

notes

* 9. Did the authors identify a primary analysis or analyses (including primary outcome measure, time of comparison, groups to be compared, and statistical test used?)

- ◯ Yes
- ◯ No
- ◯ uncertain

notes

* 10. Did they identify multiple primary analyses

- ◯ Yes
- ◯ No
- ◯ N/A, no primary analysis identified
- ◯ uncertain

notes

* 11. Do authors report the confidence interval for the estimate of the treatment effect (e.g., difference between group means, difference between group proportions or percentages, treatment group odds ratio)?

- ◯ Yes
- ◯ No

* 12. Did you have to look in the Supplementary materials to find the confidence interval for question 9?

- ◯ Yes
- ◯ No
- ◯ N/A, the answer to question 9 was no

* 13. If the answer to question 9 was yes, what confidence coefficient is reported? Please fill in box below.

* 14. Does the significance level (alpha) used to declare a treatment effect significant match the confidence interval coefficient that was reported (e.g., alpha = 0.05, CI = 95%; alpha = 0.1, CI = 90%; alpha = 0.025, CI = 97.5%)

- Yes
- No
- N/A, no confidence interval was reported
- No significance level (alpha) reported for individual significance tests (this will occur if they state that the family-wise alpha was set to a certain point, but don't say what the individual alpha levels were)

NOTE: For **question 12** , if they do not specifically state something like "a p-value (or alpha) below 0.05 was considered significant", but are obviously interpreting the trial with this cut-off in mind, please assume that the alpha was set to 0.05 for the above question.

* 15. Did the authors report a confidence interval for each treatment group (including a confidence interval for the final outcome value, the change from baseline, or the percentage change from baseline, etc.)

- Yes
- No

* 16. Was the study ended early for a reason<u>other than</u> a planned interim analysis with specific stopping rules?

- Yes
- No

Confidence Interval

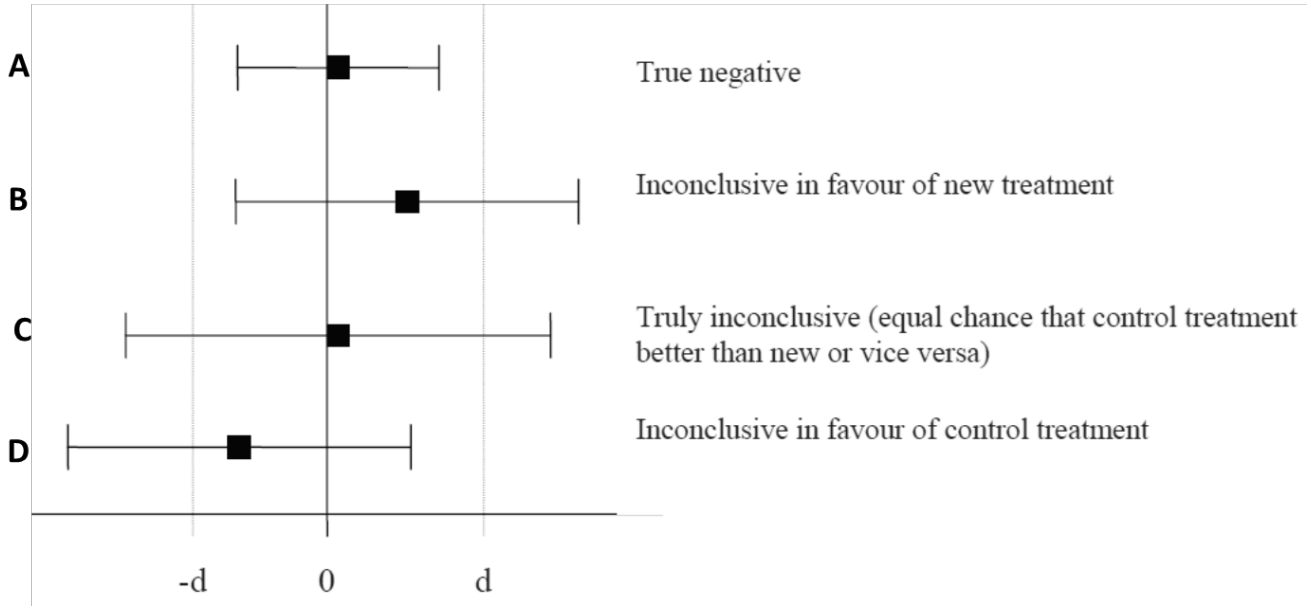
* 17. Was the result of any analysis performed with a primary outcome measure not significant? (if they do not state a significance level (i.e., alpha) used to determine significance, please assume 0.05 was significant)

- Yes
- No
- N/A, no primary outcome measure identified

Confidence Interval

Please only answer the next set of questions if ANY analysis used to compare treatments using a primary outcome measure did NOT demonstrate a statistically significant treatment effect (if they do not state a significance level (i.e., alpha) used to determine significance, please assume 0.05 was significant). If there is a primary analysis identified, please answer the following questions using that analysis If there is more than one analysis of a primary outcome measure and none are identified as primary or there are multiple primary analyses, please answer the following questions pertaining to the first NOT SIGNIFICANT analysis for the primary outcome measure presented in the Results section that compares the treatment groups .

d represents the CMTE or the difference on which the trial was powered; control = placebo in coding manual



* 18. If the article (1) defines a clinically meaningful treatment effect (CMTE) or (2) reports the magnitude of the treatment effect that was used to justify the sample size, please indicate which of the following is true about the confidence interval for the treatment effect.

See figure above with corresponding illustrations to help clarify definitions for A-D.

Please note that CMTE is used in the below options to indicate either the author-specified CMTE or difference the study was powered on

- ◯ A. The confidence interval does NOT include the CMTE or the magnitude of the treatment effect that was used to justify the sample size (i.e., we can conclude that a clinically meaningful treatment effect can be ruled out with whatever confidence was specified (e.g., 95%))
- ◯ B. The confidence interval includes the CMTE or the magnitude of the treatment effect that was used to justify the sample size in favor of the investigational treatment, but not in favor of placebo. Thus, a clinically meaningful treatment effect in favor of placebo can be ruled out with whatever confidence was specified (e.g., 95%). It cannot be ruled out, however, that the treatment is efficacious compared to placebo.
- ◯ C. The confidence interval includes the CMTE or the magnitude of the treatment effect that was used to justify the sample size in favor of both the investigational treatment and the placebo. Thus, a clinically meaningful treatment effect cannot be ruled out in either direction (i.e., in favor of treatment or in favor of placebo)
- ◯ D. The confidence interval includes the CMTE or the magnitude of the treatment effect that was used to justify the sample size in favor of the placebo, but not in favor of the investigational treatment. Thus, a clinically meaningful treatment effect in favor of treatment can be ruled out with whatever confidence was specified (e.g., 95%), but it cannot be ruled out, however, that the treatment is worse than placebo.
- ◯ Authors provide multiple estimations of the CMTE and/ or magnitude of the treatment effect that was used to justify the sample size **AND** depending on which one is considered, the answer to this question changes. Please explain in the box provided below .
- ◯ N/A, neither the CMTE nor the treatment effect that was used to justify the sample size were reported.
- ◯ N/A, no confidence interval was reported for the treatment effect
- ◯ Uncertain, please explain in the notes box

notes

* 19. If the article reports a clinically meaningful treatment effect (CMTE) or reports the treatment effect that was used to justify the sample size, do the authors explain what the confidence intervals suggest regarding whether the trial is negative or inconclusive?

- ◯ Yes
- ◯ No
- ◯ N/A, authors do not report an CMTE or treatment effect used to justify the sample size
- ◯ N/A, no confidence interval for the treatment effect was reported
- ◯ Uncertain, please make a note

notes

* 20. If the answer to question 17 is yes, is the authors' explanation appropriate regarding whether the trial could conclusively determine that the treatment was not efficacious (e.g., You would answer yes if you concluded that the results could conclusively determine the treatment was not efficacious and so did the authors.)

NOTE: the authors may only comment on whether the trial could conclusively determine that the investigational treatment was not efficacious, and not whether the trial could rule out that the investigational treatment was worse than placebo. In this case the authors' conclusions may match yours even if they do not report as much detail as is outlined in the answers in question 12.

- ◯ Yes, the authors' conclusions match mine
- ◯ No, the authors' conclusions do not match mine
- ◯ Author discussed multiple interpretations based on different assumptions for the CMTE or magnitude of the treatment effect that was used to justify the sample size. Please explain in the box provided below.
- ◯ Uncertain
- ◯ N/A, the answer to question 17 was no or N/A

Please explain in the box below how your interpretation differs from the authors' or why you are uncertain

21. Did they report a post hoc power calculation?

- ◯ Yes (please explain below)
- ◯ No

notes

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	4
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	No protocol
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	N/A not efficacy review
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix 1
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Appendix 2
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	N/A not efficacy review
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	N/A not

PRISMA 2009 Checklist

			efficacy review
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	N/A not meta analysis

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	N/A not efficacy review
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A not efficacy review
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	19
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	N/A not efficacy review
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	N/A not efficacy review
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	N/A not efficacy review
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/A not efficacy review
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	N/A not efficacy review

PRISMA 2009 Checklist

Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A not efficacy review
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	N/A not efficacy review
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	N/A not efficacy review
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	9-12
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	12-13

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.

BMJ Open

Interpretation of confidence intervals in clinical trials with non-significant results: systematic review and recommendations

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-017288.R1
Article Type:	Research
Date Submitted by the Author:	01-Jun-2017
Complete List of Authors:	Gewandter, Jennifer; University of Rochester Medical Center, University of Rochester McDermott, M; University of Rochester Kitt, Rachel; University of Rochester Medical Center, Anesthesiology Chaudari, Jenna; University of Rochester Medical Center, Anesthesiology Koch, James; University of Rochester Medical Center, Anesthesiology Evans, Scott; Harvard School of Public Heath, Department of Biostatistics and Center for Biostatistics in AIDS Research, Gross, Robert ; University of Rochester Medical Center, Neurology and Pharmacology & Physiology Markman, John; University of Rochester, Neurosurgery Turk, Dennis; University of Washington School of Medicine Dworkin, Robert; University of Rochester, Richard Bright Renal Unit
Primary Subject Heading:	Medical publishing and peer review
Secondary Subject Heading:	Medical education and training
Keywords:	Clinical trials < THERAPEUTICS, Confidence intervals, Clinical trial interpretation, Reporting

Interpretation of confidence intervals in clinical trials with non-significant results: systematic review and recommendations

Jennifer S. Gewandter[a], Michael P. McDermott[b], Rachel A. Kitt[a], Jenna Chaudari[a], James G. Koch[a], Scott R. Evans[c], Robert A. Gross[d], John D. Markman[e], Dennis C. Turk[f], Robert H. Dworkin[a]

[a]Department of Anesthesiology, University of Rochester, 601 Elmwood Ave, Rochester, NY, USA, 14642 (Jennifer S. Gewandter, Assistant Professor; Rachael A. Kitt, Research Assistant; Jenna Chaudari, Research Assistant; James G. Koch, Research Assistant; Robert Dworkin, Professor)

[b]Department of Biostatistics and Computational Biology, University of Rochester, 265 Crittenden Blvd, Rochester, NY, USA, 14642 (Michael P. McDermott, Professor)

[c]Department of Biostatistics, Harvard University, 651 Huntington Ave., Boston, MA, USA, 02115 (Scott R. Evans, Professor)

[d]Departments of Neurology and Pharmacology & Physiology, University of Rochester, 601 Elmwood Ave., Rochester, NY, USA, 14642 (Robert A. Gross, Professor)

[e]Department of Neurosurgery, University of Rochester, 2180 S. Clinton Ave., Rochester, NY, USA, 14618 (John D. Markman, Professor)

[f]Department of Anesthesiology and Pain Medicine, University of Washington, 1949 NE Pacific St., Seattle, WA, USA, 98195 (Dennis C. Turk, Professor)

Correspondence to: J. Gewandter jennifer_gewandter@urmc.rochester.edu

Abstract

Objectives Interpretation of confidence intervals (CIs) in randomized clinical trials (RCTs) with treatment effects that are not statistically significant can distinguish between results that are “negative” (the data are not consistent with a clinically meaningful treatment effect) or “inconclusive” (the data remain consistent with the possibility of a clinically meaningful treatment effect). This interpretation is important to ensure that potentially beneficial treatments are not prematurely abandoned in future research or clinical practice based on invalid conclusions.

Design Systematic review of RCT reports published in 2014 in Annals of Internal Medicine, New England Journal of Medicine, JAMA, JAMA Internal Medicine, and Lancet (n = 247).

Results Eighty-five of 99 articles with statistically non-significant results reported CIs for the treatment effect. Only 17 of those 99 articles interpreted the CI. Of the 22 articles in which CIs indicated an inconclusive result, only 4 acknowledged that the study could not rule out a clinically meaningful treatment effect.

Conclusions Interpretation of CIs is important but occurs infrequently in study reports of trials with treatment effects that are not statistically significant. Increased author interpretation of CIs could improve application of RCT results. Reporting recommendations are provided.

Article Summary

Strength and limitations of this study

Strengths

- Systematic review, including RCTs published in 6 high impact medical journals.
- Recommendations for reporting and interpreting CIs are provided.

Limitation

- Our interpretation of the CIs was based on the author-specified clinically relevant treatment effect or the treatment effect used in the sample size calculation. We did not attempt to evaluate the validity of these interpretations.

Introduction

Randomized clinical trials (RCTs) are the gold standard for evaluating the efficacy of medical treatments. However, when a statistically significant treatment effect is not demonstrated (i.e., the p-value for the primary analysis is not less than or equal to the pre-specified significance level), the estimate of the treatment effect and the p-value alone do not allow the reader of an RCT report to distinguish between the following two possibilities: (1) the treatment does not have a clinically meaningful effect, or (2) the study is unable to rule out a clinically meaningful treatment effect with a high degree of confidence (i.e., the results of the trial would best be described as "inconclusive").[1-6] However, trials for which the effect of treatment on the primary outcome variable is not statistically significant have often been called "negative" and presented as though they support the conclusion that the experimental treatment lacks efficacy.[3] This can result in premature abandonment of potentially beneficial treatments clinically and in future research programs.

For decades, biostatisticians and others have encouraged the use of confidence intervals (CIs) as a means to present the range of treatment effects consistent with the observed data and to evaluate whether RCT results that are not statistically significant suggest that the experimental treatment is ineffective or instead that the trial results are inconclusive (Figure 1).[1-6] Inconclusive results should not be used to inform clinical practice or treatment guidelines.

Previous reviews have assessed CI reporting in publications of preclinical and clinical studies within specific medical specialties.[7-14] To our knowledge, no reviews have examined CI reporting *and* interpretation in RCTs published in high-impact general medical journals.

Methods

Data Sources and Searches

RCTs published in 2014 in Annals of Internal Medicine, British Medical Journal, Journal of the American Medical Association (JAMA), JAMA Internal Medicine, Lancet, and New England Journal of Medicine were identified using PubMed (Appendix 1). The year 2014 was selected to evaluate the most recent reporting practices at the time the project was initiated. Relevant articles were identified following PRISMA guidelines.

Study Selection

Selected articles were primary reports of RCTs that compared the efficacy of at least 2 treatments (1 of which could be a placebo, active comparator, or a wait-list control) using frequentist inferential methods. Trials not evaluating treatments were excluded (e.g., comparison of two cancer screening techniques or the effect of two imaging techniques on surgical decision making). Trials utilizing a non-inferiority or super superiority design were excluded because CIs are interpreted differently for these trials than for standard superiority trials. Dose-finding studies, studies declared to be exploratory in nature, studies focused on safety, and cluster-randomized studies were also excluded. Two authors (RAK and JSG) independently screened all identified articles to determine whether they met the eligibility criteria.

Data Extraction and Quality Assessment

A coding manual was developed to evaluate the frequency with which CIs were reported for the treatment effects in RCTs (Appendix 2). In the subset of articles that reported results that were not statistically significant for the primary outcome measure, coders were asked to evaluate whether the CI for the treatment effect indicated that the data were consistent with the absence of a clinically relevant treatment effect or that the results were inconclusive (i.e., the coders compared the CI for the treatment effect to a clinically relevant treatment effect declared by the authors at any point in the manuscript or the treatment effect specified in the sample size calculation if no clinically relevant

treatment effect was described by the authors). Articles were excluded from this subset if they reported results that were both significant and not significant for the primary outcome measure (i.e., when multiple analyses were reported for the primary outcome measure). Articles were, however, included in this subset even if they reported a statistically significant treatment effect in a subgroup analysis or in analyses that were identified as sensitivity analyses because these analyses were considered secondary. A treatment effect was considered not statistically significant if the associated p-value was greater than 0.05 unless a different significance criterion was specified by the authors. If neither value was provided, the coders did not interpret the CIs for that manuscript. For the comparison of the confidence interval to the author-declared clinically meaningful treatment effect or the effect size used in the sample size calculation, the coders considered the primary analysis if one was identified. If a primary analysis was not identified, the coders considered the first analysis of a primary outcome measure that was reported by the authors. Coders also recorded whether the authors used the CI to interpret any results that were not statistically significant. The coding manual was pre-tested and modified for clarity and content by JSG and RAK in five rounds of three articles each using RCTs published in 2013 that otherwise met the eligibility criteria.

In some cases, the absolute or relative differences in event rates to be detected between groups were reported in the sample size calculation and the results concerning the treatment effect were presented as either a hazard ratio (HR), odds ratio (OR), or relative risk (RR). In these cases, JSG attempted to convert the information provided in the sample size calculation to either the HR, OR, or RR, as appropriate, using some combination of the following: absolute risk reduction ($p_0 - p_1$), relative risk reduction ($(p_0 - p_1)/p_0$), assumed event rate in the control group (p_0), and assumed event rate in the

treatment group (p_1). The following formulas were used: HR = $\ln(1 - p_1)/\ln(1 - p_0)$; OR = $(p_1(1 - p_0))/(p_0(1 - p_1))$; and RR = p_1/p_0. Such calculations were used to determine ratios representing the clinically relevant treatment effect for 26 articles. Note that the HR calculation yields an estimate that assumes an exponential distribution for the event times.

The data were extracted from each article independently by 2 authors (RAK coded all articles and JSG and JGK each coded approximately half). RAK reviewed the data for discrepancies and fixed obvious oversights. JSG reviewed any discrepancies due to interpretation and made the final decision on their resolution. JSG also reviewed the final data relating to interpretation of CIs in all of the relevant articles to ensure accuracy.

Results

Trial Characteristics

The final sample included 247 articles (Figure 2). Trial characteristics are presented in Table 1. The articles covered a range of medical specialties; the most common were cardiovascular (22%), infectious disease (15%), and cancer (13%). A little over half of the trials were sponsored, at least in part, by industry (54%).

CI Reporting

Of the 247 included articles, 99 did not report any statistically significant treatment effects on the primary outcome measure. Of those 99, 85 (86%) reported the CI for the treatment effect. Of the 14 articles that did not report the CI for the treatment effect, 6 (42%) reported the CI for the parameter estimate (e.g., mean, event rate) for each group separately. The percentage of articles that reported a CI for the treatment effect in the whole sample (n=247) was similar (85%).

Seven of the 85 articles with no statistically significant treatment effect that also reported a CI for the treatment effect did not report the magnitude of the treatment effect used to estimate the sample size of the study or specify what they would consider to be a clinically relevant treatment effect. Of the remaining 78 articles, 18 specified a clinically relevant treatment effect (6 identified this as a minimal clinically meaningful or important treatment effect; 12 identified this as a clinically meaningful, relevant, significant, important, or worthwhile treatment effect) and in the other 60 articles we interpreted the trial results based on the treatment effect used to estimate the sample size. We interpreted the nonsignificant results most commonly as falling into two categories: (1) the CI excluded the treatment effect used for the sample size calculation or the author-specified clinically relevant effect (i.e., the data were consistent with no clinically relevant treatment effect) (n=50, 64%), and (2) the CI included the treatment effect used for the sample size calculation or the author-specified clinically relevant effect in favor of the experimental treatment only (i.e., the data could not rule out a clinically meaningful effect of the experimental treatment) (n=20, 26%) (Figures 1 and 3).

Eight-two (83%) of the 99 articles with statistically nonsignificant results did not provide any interpretation of the treatment effect using CIs. Sixty-one (78%) of the 78 articles that reported confidence intervals did not interpret them. The number of articles that provided an interpretation of the CI for each journal is provided in Supplemental Table 1. In the 17 (17%) articles that did provide an interpretation of the treatment effect using CIs, the interpretations were of 5 types: (1) consistent with our interpretation, the authors stated that the CI suggested the absence of a clinically meaningful effect (n=8); (2) the authors highlighted the possible treatment effects that were consistent with the CI, but did not speculate on whether those effect sizes were clinically meaningful (n=4); (3) similar to our conclusions, the authors concluded that based on the CI, a clinically meaningful treatment effect could not be ruled out (n=2); (4) the authors conservatively

stated that they could not rule out clinically meaningful treatment effects even though the CI excluded the effect size that the trial was designed to detect (n=2); and (5) the authors described the treatment as “modestly effective” and then went on to state that they “focused on the effect size and 95% CI whilst showing p-values, which is in line with the CONSORT 2010 guidelines” when the results were not statistically significant (n=1). We interpreted this trial’s results to be inconclusive (Figure 3).

Discussion

Consistent with widespread recommendations,[1-6] we found that the 85% of articles reporting RCTs published in 6 high-impact medical journals in 2014 reported the CIs for the treatment effect. The percentage of articles that reported CIs in our review was higher than the percentage of articles that reported CIs in previous reviews of RCTs in specialty journals (85% in our review vs. 5% to 66% in previous reviews).[7-14] This increase could be due to the earlier publication periods covered by the previous reviews (i.e., 1990 - 2008). It could also be due to the fact that the 6 journals included in our review require adherence to the CONSORT guidelines,[15] which promote transparent reporting, for publication of RCTs. Regardless of whether the increased reporting of CIs that we observed is in fact due to an effect of time or of the specific journals selected, our results suggest that relatively high quality reporting is possible when required by guidelines, reviewers, and/or editors.

Although reporting CIs provides the reader the ability to make a judgment regarding whether the results are “negative” or inconclusive, such interpretations require an understanding of CIs and knowledge of what should be considered a minimal clinically meaningful treatment effect with respect to the outcome variable used in the trial. Because it cannot be assumed that all readers and stakeholders will have this expertise, or necessarily agree on this point, best reporting practices should include careful interpretation of the CIs and their implications for the conclusions of the trial.

The percentage of articles in our sample that interpreted CIs was much lower than the percentage that simply reported them. Only 17 of the 99 articles that reported analyses of a primary outcome measure that were not statistically significant used a CI to (1) highlight the range of values of the treatment effect that were consistent with the data or (2) discuss whether the trial results were inconclusive or were consistent with the absence of a clinically meaningful treatment effect. Additionally, although the CIs of 22 articles included the treatment effect used for the sample size calculation or the author-specified clinically relevant treatment effect, only 4 of these articles stated that the study could not rule out a clinically meaningful treatment effect. Our data suggest that many authors do not discuss that the results of their trial can be considered inconclusive on the basis of the CIs they report, perhaps because they believe that doing so might decrease the perceived importance of the RCT. Acknowledging that the study cannot rule out a clinically meaningful effect is important to ensure that clinicians, policy makers, and payers do not inappropriately use the trial results as evidence to suggest that the treatment is ineffective.

It must be acknowledged, of course, that the magnitude of a treatment effect that would be considered clinically meaningful can differ depending on many factors, including the setting of the trial and perspective of the reader.[16-19] For example, if a treatment has very few side effects or no treatments currently exist for the condition, the minimal clinically meaningful treatment effect is likely to be relatively small compared to a treatment with greater safety risk. This may be especially true from an individual patient's perspective. On the other hand, the minimal clinically meaningful treatment effect may be larger from a funder or researcher's perspective when considering whether to support or pursue a line of research. It is important that these potential differences in perspective are acknowledged when interpreting CIs and that the authors

present the rationale for the minimal clinically meaningful treatment effect that they used to interpret the results of the trial.

Another method that is sometimes used to interpret the results of RCTs is to present a *post hoc* power calculation. As many authors have correctly argued, however, such a calculation is irrelevant to trial interpretation.[20-22] Encouragingly, only 3 of the included articles with treatment effects on the primary outcome measures that were not statistically significant reported a post hoc power calculation. Three other articles stated that the trials had adequate power without any apparent justification. Post hoc power calculations should be avoided and interpretations regarding whether a trial is "negative" or "inconclusive" would be better based on CIs.

Interestingly, 8% of the 247 included articles reported a CI for the parameter of interest for each separate treatment group (e.g., mean or event rate), but not for the between-group treatment effect. It is important to emphasize that CIs for the parameters of individual treatment groups are not informative with respect to evaluating whether the results of a trial with a statistically non-significant treatment effect are "negative" as opposed to inconclusive.

A limitation of our review is that we based our interpretation of the CIs reported in the studies with statistically non-significant treatment effects on the author-specified clinically relevant treatment effect or the magnitude of the treatment effect used in the sample size calculation. We did not attempt to evaluate the validity of these values as being of clinical importance because our intention was to evaluate the frequency with which authors used CIs in the interpretation of trial results and whether these interpretations were consistent with their assumptions regarding clinically meaningful treatment effects. Furthermore, the treatment effects used to determine the sample size of a trial are not necessarily what one would consider to be the minimal clinically meaningful treatment effect that investigators might still be pleased to demonstrate.[23]

For example, investigators may justify the sample size using a larger effect than would be considered minimally clinically meaningful if an effect of this magnitude were anticipated based on existing data.

It would have been interesting to determine whether the articles that concluded that the trials were “negative” without consideration of CIs actually reported CIs that did not exclude the clinically relevant treatment effect. Unfortunately, we were unable to categorize articles as claiming that the trial was “negative” because authors often had contradictory statements throughout the Discussion regarding whether the “negative” conclusion was definitive. These inconsistencies highlight the importance of using CIs to interpret whether a trial with a treatment effect that is not statistically significant is “negative” or inconclusive.

In conclusion, the majority of the trials we reviewed reported the CI for the treatment effect, demonstrating relatively high-quality, transparent reporting of RCT results. In contrast, a substantially smaller percentage of articles reporting analyses of the primary outcome measure that were not statistically significant discussed the implications of the CIs of the treatment effect when interpreting the results of their study. We encourage all authors and reviewers to prioritize interpretation of RCT findings using CIs, especially when the CIs indicate that the data cannot rule out a clinically meaningful treatment effect (Table 2). We also encourage readers to consider the CIs when applying the results of RCTs with non-significant results to their clinical practice or research program.

Acknowledgements

We thank Dr. Stephen Senn for his thoughtful review and comments on our manuscript. This article was reviewed and approved by the Executive Committee of the

ACTTION public-private partnership with the United States Food and Drug Administration.

Conflict of interest statement

The views expressed in this article are those of the authors and no official endorsement by the Food and Drug Administration (FDA) or the pharmaceutical and device companies that provided unrestricted grants to support the activities of the Analgesic, Anesthetic, and Addiction Clinical Trial Translations, Innovations, Opportunities, and Networks (ACTTION) public-private partnership should be inferred. The authors have no conflicts of interest related to the work presented in this manuscript to disclose.

Funding Statement

Financial support for this project was provided by the ACTTION public-private partnership, which has received research contracts, grants, or other revenue from the FDA, multiple pharmaceutical and device companies, and other sources.

Contributorship

Jennifer Gewandter contributed to the design of the review and coding. She performed the analyses, wrote the first draft of the manuscript, and incorporated co-author revisions. Robert Dworkin, Michael McDermott, Scott Evans, Robert Gross, John Markman, and Dennis Turk contributed to the design of the study, interpretation of the results, and provided feedback on the manuscript. Rachel Kitt, Jenna Chaudari, and James Koch contributed to coding for the systematic review and provided feedback on the manuscript.

Data Sharing

No additional data available

Exclusive License agreement

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, *a worldwide license,* to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) license any third party to do any or all of the above.

References

1. Alderson P. Absence of evidence is not evidence of absence. BMJ 2004;328:476-7.
2. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311:485.
3. Freiman JA, Chalmers TC, Smith H, Jr., Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. NEJM 1978;299:690-4.
4. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ 1986;292:746-50.
5. Rothman KJ. A show of confidence. N Engl J Med 1978;299:1362-3.

6. Simon R. Confidence intervals for reporting results of clinical trials. Ann Intern Med 1986;105:429-35.
7. Anttila H, Malmivaara A, Kunz R, Autti-Ramo I, Makela M. Quality of reporting of randomized, controlled trials in cerebral palsy. Pediatrics 2006;117:2222-30.
8. Fan FF, Xu Q, Sun Q, Zhao SJ, Wang P, Guo XR. Assessment of the reporting quality of randomized controlled trials on treatment of coronary heart disease with traditional chinese medicine from the chinese journal of integrated traditional and Western medicine: a systematic review. PLoS One 2014;9:e86360.
9. Kiehna EN, Starke RM, Pouratian N, Dumont AS. Standards for reporting randomized controlled trials in neurosurgery. J Neurosurg 2011;114:280-5.
10. Kloukos D, Papageorgiou SN, Fleming PS, Petridis H, Pandis N. Reporting of statistical results in prosthodontic and implantology journals: p values or confidence intervals? Int J Prosthodont 2014;27:427-32.
11. Naunheim MR, Walcott BP, Nahed BV, Simpson AK, Agarwalla PK, Coumans JV. The quality of randomized controlled trial reporting in spine literature. Spine 2011;36:1326-30.
12. Polychronopoulou A, Pandis N, Eliades T. Appropriateness of reporting statistical results in orthodontics: the dominance of P values over confidence intervals. Eur J Orthod 2011;33:22-5.
13. Urbanic JJ, Lee WR. Confidence intervals and survival estimates: a systematic review of 3 oncology journals. Am J Clin Oncol 2006;29:405-7.
14. Wang P, Xu Q, Sun Q, Fan FF, Guo XR, Guo F. Assessment of the reporting quality of randomized controlled trials on the treatment of diabetes mellitus with traditional chinese medicine: a systematic review. PLoS One 2013;8:e70586.

15. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.
16. Dworkin RH, Turk DC, McDermott MP, Pierce-Sander S, Burke LB, Cowan P, Farrar JT, Hertz S, Raja SN, Rappaport BA, Rauschkolb C, Sampaio C. Interpreting the clinical importance of group differences in chronic pain clinical trials: IMMPACT recommendations. Pain 2009;238-244.
17. Guyhatt GH, Juniper, EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomized trials. BMJ 316:1998;690-3.
18. Kraemer HC, Morgan GA, Leech NL, Gliner JA, Vaske JJ, Harmon RJ. Measures of clinical significance. J. Am. Child Adolesc. Psychiatry 2003;42:1524-9.
19. Ruyssen-Witrand A, Tubach F, Ravaud P, Systematic review reveals heterogeneity in definition of clinically relevant difference in pain. J. Clin. Epidemiol. 2011;64:463-70.
20. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 1994;121:200-6
21. Greenland S. Nonsignificance plus high power does not imply support for the null over the alternative. Ann Epidemiol 2012;22:364-8.
22. Hoenig JM, Heisey DM. An abuse of power: Ther pervasive fallacy of power calculations for data analysis. Am Stat 2001:55;19-24.
23. Senn SJ. Statistical Issues in Drug Development (2nd edition). John Wiley & Sons, Ltd., West Sussex. 2007.

Table 1. Trial Characteristics

Characteristic	All articles (N=247)	Articles reporting a treatment effect (TE) that was not statistically significant, the CI of the TE, and a value for the TE that the authors considered to be clinically meaningful (N=78)
Journal		
New England Journal of Medicine	105 (43%)	31 (40%)
JAMA	61 (25%)	22 (28%)
The Lancet	50 (20%)	11 (14%)
British Medical Journal	13 (5%)	8 (10%)
JAMA Internal Medicine	11 (4%)	1 (1%)
Annals of Internal Medicine	7 (3%)	5 (6%)
Design		
Parallel group	245 (99%)	78 (100%)
Cross-over	2 (1%)	0 (0%)
Number randomized	480 (224 – 1195)	730 (311 – 1880)
Medical specialty		
Cardiovascular	55 (22%)	23 (29%)
Infectious disease	38 (15%)	12 (15%)
Cancer	31 (13%)	4 (5%)
Neurology (including pain)	22 (9%)	7 (9%)
Pulmonary	13 (5%)	6 (8%)
Psychiatry	12 (5%)	1 (1%)
Other*	76 (31%)	25 (32%)
Type of intervention		
Treatment	183 (74%)	52 (67%)
Prevention	64 (26%)	26 (33%)
Sponsor		
Industry	134 (54%)	36 (46%)
Other	113 (46%)	42 (54%)

Values are N (%) or median (interquartile range)

*Other includes areas represented by fewer than 10 trials including urology, orthopedics, diabetes, immune disorders, etc.

Table 2. Confidence interval reporting recommendations for RCTs with statistically non-significant results

- Report confidence intervals (CIs) for the treatment effect.
- Discuss interpretation of the CI regarding the magnitude of effects that can be *ruled-out* with reasonable confidence.
- Discuss whether the results suggest a “negative” or “inconclusive” result
- Acknowledge any uncertainty regarding what is considered a clinically meaningful treatment effect on the outcome measure used in the trial

Figure Legends

Figure 1. Using confidence intervals to interpret results of RCTs

CMTE clinically meaningful treatment effect.

Note that a value of zero indicates no treatment effect in this case; in other cases such as when the treatment effect is quantified using, for example, an odds ratio, hazard ratio, or relative risk, a value of 1.0 would indicate no treatment effect

Adapted from reference 23.

Figure 2. PRISMA diagram

RCT Randomized clinical trial

*Secondary analysis of data from a previously reported trial
** RCT examines efficacy of something other than a medical or lifestyle intervention (e.g., a cancer screening method or a diagnostic decision making tool)

Figure 3. Confidence interval reporting and interpretation.
CI confidence interval; POM primary outcome measure

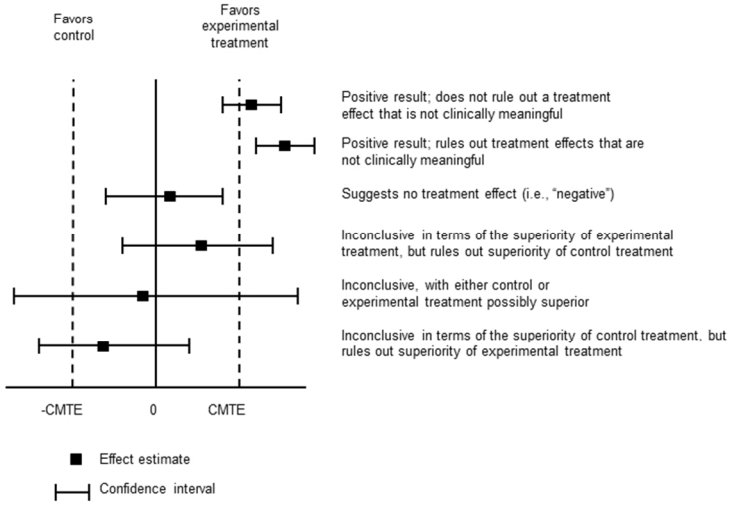


Figure 1

81x60mm (300 x 300 DPI)

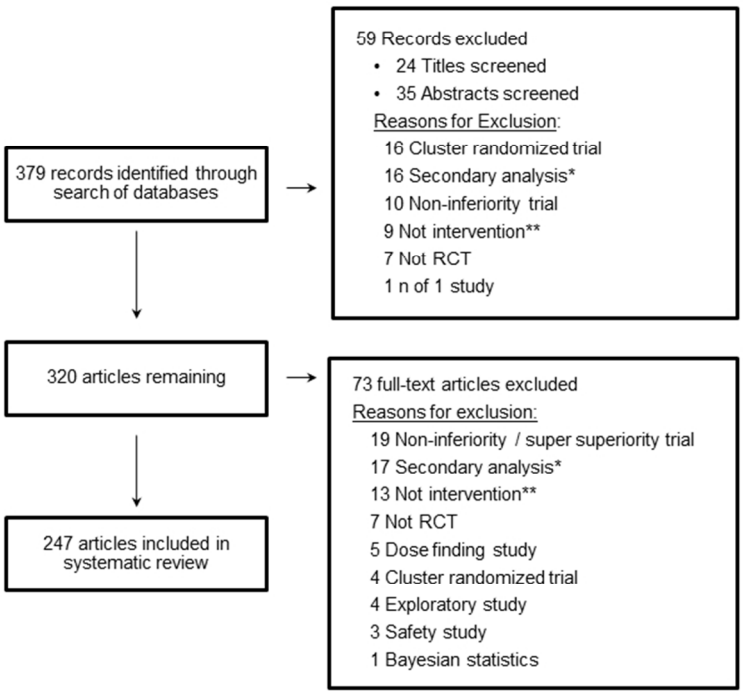
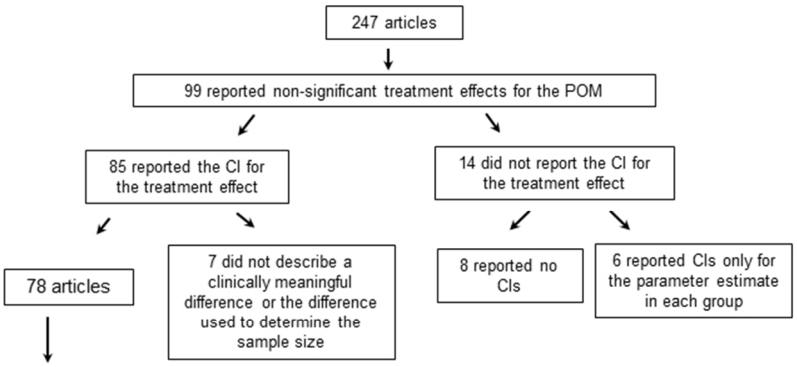


Figure 2

60x81mm (300 x 300 DPI)



Interpretation based on specified clinically relevant treatment effect or magnitude of the treatment effect used in the sample size calculation	N (%)
CI excludes the clinically relevant effect in favor of both the experimental treatment and the control treatment (i.e., the study suggests no treatment effect)	50 (64%)
CI includes the clinically relevant effect in favor of the experimental treatment (and therefore the study is inconclusive regarding the superiority of the experimental treatment), with superiority of the control treatment excluded	20 (26%)
CI includes the clinically relevant effect in favor of the both the experimental treatment and the control treatment (and therefore the study is inconclusive regarding the superiority of either the control treatment or the experimental treatment)	2 (3%)
CI includes the clinically relevant effect in favor of the control treatment (and therefore the study is inconclusive regarding the superiority of the control treatment), with superiority of the experimental treatment excluded	5 (6%)
CI rules out benefit of 1 of 2 active treatments compared	1 (1%)

17 articles provided some interpretation of the CI

Article interpretation	N (%)
Correctly concludes that the CI suggests the absence of a clinically meaningful treatment effect	8 (47%)
Discusses the CI and points out that the data are consistent with a treatment effect within its bounds, but does not comment on whether the CI includes a clinically meaningful treatment effect	4 (24%)
Correctly concludes that based on the CI, a potentially meaningful treatment effect cannot be ruled out	2 (12%)
States that a clinically meaningful treatment effect cannot be ruled out, even though the CI did not include the treatment effect that the study was designed to detect	2 (12%)
Describes the treatment as "modestly effective" and then goes on to state that they "focused on the effect size and 95% CI whilst showing p-values, which is in line with the CONSORT 2010 guidelines" when the results were not statistically significant	1 (6%)

Figure 3

60x81mm (300 x 300 DPI)

Appendix 1. Search Strategy.

("Lancet (London, England)"[Jour] OR "BMJ (Clinical research ed.)"[Jour] OR "JAMA"[Jour] OR "The New England journal of medicine"[Jour] OR "JAMA internal medicine"[Jour] OR "Annals of internal medicine"[Jour])

Filter: RCT

Date Search performed: 1/26/2015

“Create Date” Range : 2013-2014

2013 was included to ensure that all articles published in 2014 but created in Pubmed “ahead of print” in 2013 were captured. 744 articles were returned in the search and of those 379 were published in 2014.

Confidence Interval

* 1. Coder

- ◯ Jen
- ◯ Rachel

* 2. Article

* 3. Journal

* 4. Design

- ◯ Parallel
- ◯ Crossover
- ◯ Other (please specify)

* 5. Sponsor

- ☐ Industry
- ☐ Government
- ☐ Professional association / foundation
- ☐ Academic Institution / department funds
- ☐ None (explicitly stated)
- ☐ Not reported
- ☐ Other (please specify)

* 6. Condition

- ◯ Cancer
- ◯ Neurological (e.g., Movement disorders, Cognitive issues)
- ◯ Pain (excluding orthopedic associated pain such as OA)
- ◯ Respiratory
- ◯ Cardiovascular
- ◯ Infectious diseases (e.g., viral and bacterial infections)
- ◯ Obesity
- ◯ Pregnancy outcomes
- ◯ Smoking cessation
- ◯ Gastro/intestinal disorders other than cancer (i.e., GERD, IBS)
- ◯ Diabetes
- ◯ Orthopedic
- ◯ Other (please specify)

7. Is the study for prevention or treatment?

- ◯ Prevention
- ◯ Treatment

Confidence Interval

* 8. Did the authors identify a primary outcome measure?

- ◯ Yes
- ◯ No
- ◯ uncertain

notes

* 9. Did the authors identify a primary analysis or analyses (including primary outcome measure, time of comparison, groups to be compared, and statistical test used?)

- ◯ Yes
- ◯ No
- ◯ uncertain

notes

* 10. Did they identify multiple primary analyses

- ◯ Yes
- ◯ No
- ◯ N/A, no primary analysis identified
- ◯ uncertain

notes

* 11. Do authors report the confidence interval for the estimate of the treatment effect (e.g., difference between group means, difference between group proportions or percentages, treatment group odds ratio)?

- ◯ Yes
- ◯ No

* 12. Did you have to look in the Supplementary materials to find the confidence interval for question 9?

- ◯ Yes
- ◯ No
- ◯ N/A, the answer to question 9 was no

* 13. If the answer to question 9 was yes, what confidence coefficient is reported? Please fill in box below.

* 14. Does the significance level (alpha) used to declare a treatment effect significant match the confidence interval coefficient that was reported (e.g., alpha = 0.05, CI = 95%; alpha = 0.1, CI = 90%; alpha = 0.025, CI = 97.5%)

- Yes
- No
- N/A, no confidence interval was reported
- No significance level (alpha) reported for individual significance tests (this will occur if they state that the family-wise alpha was set to a certain point, but don't say what the individual alpha levels were)

NOTE: For **question 12** , if they do not specifically state something like "a p-value (or alpha) below 0.05 was considered significant", but are obviously interpreting the trial with this cut-off in mind, please assume that the alpha was set to 0.05 for the above question.

* 15. Did the authors report a confidence interval for each treatment group (including a confidence interval for the final outcome value, the change from baseline, or the percentage change from baseline, etc.)

- Yes
- No

* 16. Was the study ended early for a reason<u>other than</u> a planned interim analysis with specific stopping rules?

- Yes
- No

Confidence Interval

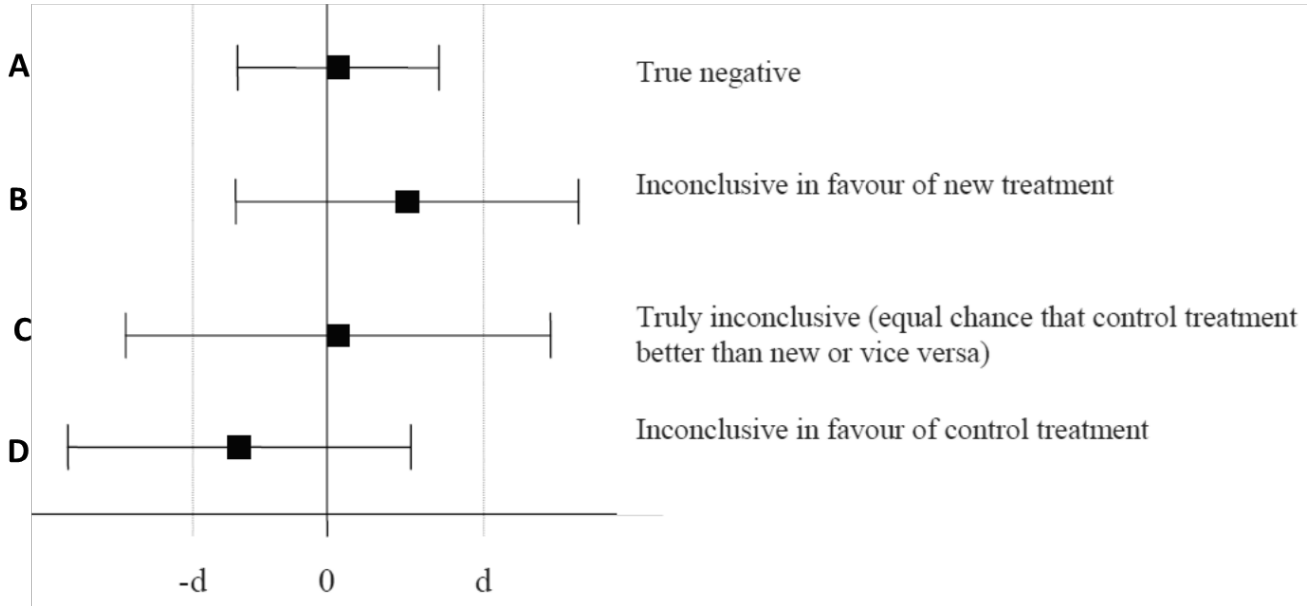
* 17. Was the result of any analysis performed with a primary outcome measure not significant? (if they do not state a significance level (i.e., alpha) used to determine significance, please assume 0.05 was significant)

- Yes
- No
- N/A, no primary outcome measure identified

Confidence Interval

Please only answer the next set of questions if ANY analysis used to compare treatments using a primary outcome measure did NOT demonstrate a statistically significant treatment effect (if they do not state a significance level (i.e., alpha) used to determine significance, please assume 0.05 was significant). If there is a primary analysis identified, please answer the following questions using that analysis If there is more than one analysis of a primary outcome measure and none are identified as primary or there are multiple primary analyses, please answer the following questions pertaining to the first NOT SIGNIFICANT analysis for the primary outcome measure presented in the Results section that compares the treatment groups .

d represents the CMTE or the difference on which the trial was powered; control = placebo in coding manual



* 18. If the article (1) defines a clinically meaningful treatment effect (CMTE) or (2) reports the magnitude of the treatment effect that was used to justify the sample size, please indicate which of the following is true about the confidence interval for the treatment effect.

See figure above with corresponding illustrations to help clarify definitions for A-D.

Please note that CMTE is used in the below options to indicate either the author-specified CMTE or difference the study was powered on

- ( ) A. The confidence interval does NOT include the CMTE or the magnitude of the treatment effect that was used to justify the sample size (i.e., we can conclude that a clinically meaningful treatment effect can be ruled out with whatever confidence was specified (e.g., 95%))
- ( ) B. The confidence interval includes the CMTE or the magnitude of the treatment effect that was used to justify the sample size in favor of the investigational treatment, but not in favor of placebo. Thus, a clinically meaningful treatment effect in favor of placebo can be ruled out with whatever confidence was specified (e.g., 95%). It cannot be ruled out, however, that the treatment is efficacious compared to placebo.
- ( ) C. The confidence interval includes the CMTE or the magnitude of the treatment effect that was used to justify the sample size in favor of both the investigational treatment and the placebo. Thus, a clinically meaningful treatment effect cannot be ruled out in either direction (i.e., in favor of treatment or in favor of placebo)
- ( ) D. The confidence interval includes the CMTE or the magnitude of the treatment effect that was used to justify the sample size in favor of the placebo, but not in favor of the investigational treatment. Thus, a clinically meaningful treatment effect in favor of treatment can be ruled out with whatever confidence was specified (e.g., 95%), but it cannot be ruled out, however, that the treatment is worse than placebo.
- ( ) Authors provide multiple estimations of the CMTE and/ or magnitude of the treatment effect that was used to justify the sample size **AND** depending on which one is considered, the answer to this question changes. Please explain in the box provided below .
- ( ) N/A, neither the CMTE nor the treatment effect that was used to justify the sample size were reported.
- ( ) N/A, no confidence interval was reported for the treatment effect
- ( ) Uncertain, please explain in the notes box

notes

* 19. If the article reports a clinically meaningful treatment effect (CMTE) or reports the treatment effect that was used to justify the sample size, do the authors explain what the confidence intervals suggest regarding whether the trial is negative or inconclusive?

- ( ) Yes
- ( ) No
- ( ) N/A, authors do not report an CMTE or treatment effect used to justify the sample size
- ( ) N/A, no confidence interval for the treatment effect was reported
- ( ) Uncertain, please make a note

notes

* 20. If the answer to question 17 is yes, is the authors' explanation appropriate regarding whether the trial could conclusively determine that the treatment was not efficacious (e.g., You would answer yes if you concluded that the results could conclusively determine the treatment was not efficacious and so did the authors.)

NOTE: the authors may only comment on whether the trial could conclusively determine that the investigational treatment was not efficacious, and not whether the trial could rule out that the investigational treatment was worse than placebo. In this case the authors' conclusions may match yours even if they do not report as much detail as is outlined in the answers in question 12.

- ◯ Yes, the authors' conclusions match mine
- ◯ No, the authors' conclusions do not match mine
- ◯ Author discussed multiple interpretations based on different assumptions for the CMTE or magnitude of the treatment effect that was used to justify the sample size. Please explain in the box provided below.
- ◯ Uncertain
- ◯ N/A, the answer to question 17 was no or N/A

Please explain in the box below how your interpretation differs from the authors' or why you are uncertain

21. Did they report a post hoc power calculation?

- ◯ Yes (please explain below)
- ◯ No

notes

Supplemental Table 1

Journal	N (%) providing an interpretation of the CI (of studies with primary analyses not yielding statistically significant results)
Annals of Internal Medicine	2 / 5 (40%)
British Medical Journal	2 / 9 (22%)
JAMA	6 / 29 (21%)
JAMA Internal Medicine	0 / 2 (0%)
New England Journal of Medicine	4 / 39 (10%)
The Lancet	3 / 15 (20%)

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	4
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	No protocol
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	N/A not efficacy review
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix 1
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Appendix 2
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	N/A not efficacy review
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	N/A not

PRISMA 2009 Checklist

			efficacy review
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	N/A not meta analysis

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	N/A not efficacy review
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A not efficacy review
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	20
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	N/A not efficacy review
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	N/A not efficacy review
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	N/A not efficacy review
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/A not efficacy review
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	N/A not efficacy review

PRISMA 2009 Checklist

Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A not efficacy review
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	N/A not efficacy review
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	N/A not efficacy review
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	9-12
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	13

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.