Cochrane Library

Cochrane Database of Systematic Reviews

Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials (Review)

Anglemyer A, Horvath HT, Bero L

Anglemyer A, Horvath HT, Bero L.
Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials.
Cochrane Database of Systematic Reviews 2014, Issue 4. Art. No.: MR000034.
DOI: 10.1002/14651858.MR000034.pub2.

www.cochranelibrary.com

WILEY

TABLE OF CONTENTS

[Methodology Review]

Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials

Andrew Anglemyer[1], Hacsi T Horvath[1], Lisa Bero[2]

[1]Global Health Sciences, University of California, San Francisco, San Francisco, California, USA. [2]Department of Clinical Pharmacy and Institute for Health Policy Studies, University of California San Francisco, San Francisco, California, USA

Contact address: Lisa Bero, Department of Clinical Pharmacy and Institute for Health Policy Studies, University of California San Francisco, Suite 420, Box 0613, 3333 California Street, San Francisco, California, 94143-0613, USA. berol@pharmacy.ucsf.edu.

Editorial group: Cochrane Methodology Review Group.
Publication status and date: New, published in Issue 4, 2014.

Citation: Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database of Systematic Reviews* 2014, Issue 4. Art. No.: MR000034. DOI: 10.1002/14651858.MR000034.pub2.

ABSTRACT

Background

Researchers and organizations often use evidence from randomized controlled trials (RCTs) to determine the efficacy of a treatment or intervention under ideal conditions. Studies of observational designs are often used to measure the effectiveness of an intervention in 'real world' scenarios. Numerous study designs and modifications of existing designs, including both randomized and observational, are used for comparative effectiveness research in an attempt to give an unbiased estimate of whether one treatment is more effective or safer than another for a particular population.

A systematic analysis of study design features, risk of bias, parameter interpretation, and effect size for all types of randomized and non-experimental observational studies is needed to identify specific differences in design types and potential biases. This review summarizes the results of methodological reviews that compare the outcomes of observational studies with randomized trials addressing the same question, as well as methodological reviews that compare the outcomes of different types of observational studies.

Objectives

To assess the impact of study design (including RCTs versus observational study designs) on the effect measures estimated.

To explore methodological variables that might explain any differences identified.

To identify gaps in the existing research comparing study designs.

Search methods

We searched seven electronic databases, from January 1990 to December 2013.

Along with MeSH terms and relevant keywords, we used the sensitivity-specificity balanced version of a validated strategy to identify reviews in PubMed, augmented with one term ("review" in article titles) so that it better targeted narrative reviews. No language restrictions were applied.

Selection criteria

We examined systematic reviews that were designed as methodological reviews to compare quantitative effect size estimates measuring efficacy or effectiveness of interventions tested in trials with those tested in observational studies. Comparisons included RCTs versus

observational studies (including retrospective cohorts, prospective cohorts, case-control designs, and cross-sectional designs). Reviews were not eligible if they compared randomized trials with other studies that had used some form of concurrent allocation.

Data collection and analysis

In general, outcome measures included relative risks or rate ratios (RR), odds ratios (OR), hazard ratios (HR). Using results from observational studies as the reference group, we examined the published estimates to see whether there was a relative larger or smaller effect in the ratio of odds ratios (ROR).

Within each identified review, if an estimate comparing results from observational studies with RCTs was not provided, we pooled the estimates for observational studies and RCTs. Then, we estimated the ratio of ratios (risk ratio or odds ratio) for each identified review using observational studies as the reference category. Across all reviews, we synthesized these ratios to get a pooled ROR comparing results from RCTs with results from observational studies.

Main results

Our initial search yielded 4406 unique references. Fifteen reviews met our inclusion criteria; 14 of which were included in the quantitative analysis.

The included reviews analyzed data from 1583 meta-analyses that covered 228 different medical conditions. The mean number of included studies per paper was 178 (range 19 to 530).

Eleven (73%) reviews had low risk of bias for explicit criteria for study selection, nine (60%) were low risk of bias for investigators' agreement for study selection, five (33%) included a complete sample of studies, seven (47%) assessed the risk of bias of their included studies,

Seven (47%) reviews controlled for methodological differences between studies,

Eight (53%) reviews controlled for heterogeneity among studies, nine (60%) analyzed similar outcome measures, and four (27%) were judged to be at low risk of reporting bias.

Our primary quantitative analysis, including 14 reviews, showed that the pooled ROR comparing effects from RCTs with effects from observational studies was 1.08 (95% confidence interval (CI) 0.96 to 1.22). Of 14 reviews included in this analysis, 11 (79%) found no significant difference between observational studies and RCTs. One review suggested observational studies had larger effects of interest, and two reviews suggested observational studies had smaller effects of interest.

Similar to the effect across all included reviews, effects from reviews comparing RCTs with cohort studies had a pooled ROR of 1.04 (95% CI 0.89 to 1.21), with substantial heterogeneity (I^2 = 68%). Three reviews compared effects of RCTs and case-control designs (pooled ROR: 1.11 (95% CI 0.91 to 1.35)).

No significant difference in point estimates across heterogeneity, pharmacological intervention, or propensity score adjustment subgroups were noted. No reviews had compared RCTs with observational studies that used two of the most common causal inference methods, instrumental variables and marginal structural models.

Authors' conclusions

Our results across all reviews (pooled ROR 1.08) are very similar to results reported by similarly conducted reviews. As such, we have reached similar conclusions; on average, there is little evidence for significant effect estimate differences between observational studies and RCTs, regardless of specific observational study design, heterogeneity, or inclusion of studies of pharmacological interventions. Factors other than study design *per se* need to be considered when exploring reasons for a lack of agreement between results of RCTs and observational studies. Our results underscore that it is important for review authors to consider not only study design, but the level of heterogeneity in meta-analyses of RCTs or observational studies. A better understanding of how these factors influence study effects might yield estimates reflective of true effectiveness.

PLAIN LANGUAGE SUMMARY

Comparing effect estimates of randomized controlled trials and observational studies

Researchers and organizations often use evidence from randomized controlled trials (RCTs) to determine the efficacy of a treatment or intervention under ideal conditions, while studies of observational designs are used to measure the effectiveness of an intervention in non-experimental, 'real world' scenarios. Sometimes, the results of RCTs and observational studies addressing the same question may have different results. This review explores the questions of whether these differences in results are related to the study design itself, or other study characteristics.

This review summarizes the results of methodological reviews that compare the outcomes of observational studies with randomized trials addressing the same question, as well as methodological reviews that compare the outcomes of different types of observational studies.

The main objectives of the review are to assess the impact of study design--to include RCTs versus observational study designs (e.g. cohort versus case-control designs) on the effect measures estimated, and to explore methodological variables that might explain any differences.

We searched multiple electronic databases and reference lists of relevant articles to identify systematic reviews that were designed as methodological reviews to compare quantitative effect size estimates measuring efficacy or effectiveness of interventions of trials with observational studies or different designs of observational studies. We assessed the risks of bias of the included reviews.

Our results provide little evidence for significant effect estimate differences between observational studies and RCTs, regardless of specific observational study design, heterogeneity, inclusion of pharmacological studies, or use of propensity score adjustment. Factors other than study design *per se* need to be considered when exploring reasons for a lack of agreement between results of RCTs and observational studies.

BACKGROUND

Researchers and organizations often use evidence from randomized controlled trials (RCTs) to determine the efficacy of a treatment or intervention under ideal conditions. Studies of observational design are used to measure the effectiveness of an intervention in non-experimental, 'real world' scenarios at the population level. The Institute of Medicine defines comparative effectiveness research (CER) as: “the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels" (Institute of Medicine 2009). Comparative effectiveness research has also been called "comparative clinical effectiveness research" and "patient centered outcomes research" (Kamerow 2011). Regardless of what this type of research is called, it should give an unbiased estimate of whether one treatment is more effective or safer than another for a particular population. Debate about the validity of observational studies versus randomized trials for estimating effectiveness of interventions has continued for decades.

Numerous study designs and modifications of existing designs, both randomized and observational, are used for comparative effectiveness research. These include, but are not limited to, head-to-head randomized trials, cluster-randomized trials, adaptive designs, practice/pragmatic/explanatory trials, PBE-CPI “practice based evidence for clinical practice improvement,” natural experiments, observational or cross-sectional studies of registries and databases including electronic medical records, meta-analysis, network meta-analysis, modeling and simulation. Modifications can often include newer observational study analysis approaches employing so-called causal inference techniques, which can include instrumental variables, marginal structural models, propensity scores, among others. Non-randomized experimental designs (e.g., non-randomized trials), also play a role in comparative effectiveness research, but this review focuses on comparing randomized trials with non-experimental observational designs. As noted in the *Cochrane Handbook for Systematic Reviews of Interventions*, potential biases for all non-randomized studies are likely to be greater than for randomized trials (Higgins 2011). A systematic analysis of study design features, risk of bias, and effect size for all the types of studies used for comparative effectiveness research is needed to identify specific differences in design types and potential biases.

This review summarizes the results of methodological reviews that compare the outcomes of observational studies with randomized trials addressing the same question, as well as methodological reviews that compare the outcomes of different types of observational studies. A number of reviews comparing the effect sizes and/or biases in RCTs and observational studies (or non-randomized controlled trials) have been conducted (Benson 2000; Britton 1998; Concato 2000; Deeks 2003; Ioannidis 2001; Kunz 1998; Kunz 2002; MacLehose 2000; Odgaard-Jensen 2011; Oliver 2010; Sacks 1982; Wilson 2001).These reviews examined whether certain types of study designs report smaller or larger treatment effects, or change the direction of effects. Some reviews found that a lack of randomization or inadequate randomization is associated with selection bias, larger treatment effects, smaller treatment effects, or reversed direction of treatment effects (Deeks 2003; Ioannidis 2001; Kunz 1998; Odgaard-Jensen 2011), while others found little to no difference in treatment effect sizes between study designs (Benson 2000; Britton 1998; Concato 2000; MacLehose 2000; Oliver 2010). However, there has been no systematic review of comparisons of all study designs currently being used for comparative effectiveness research. Reviews that compared RCTs with observational studies most often limited the comparison to cohort studies, or the types of observational designs included were not specified. In addition, most of the reviews were published between 1982 and 2003 and the methodology for observational studies has evolved since that time. One Cochrane review, first published in 2002 (Kunz 2002), has been archived and superseded by later versions. The most recent version of that review, published in 2011, compared random allocation versus non-random allocation or adequate versus inadequate/unclear concealment of allocation in randomized trials (Odgaard-Jensen 2011). This review included comparisons of randomized trials ("randomized controlled trials" or "RCTs"); non-randomized trials with concurrent controls, and non-equivalent control group designs. The review excluded comparisons of studies using historical controls (patients treated earlier than those who received the intervention being evaluated, frequently called "historically controlled trials" or "HCTs"); classical observational studies, including cohort studies, cross-sectional studies, case-control studies and ’outcomes studies’ (evaluations using large administrative or clinical databases). Another recent review assessing the relationship between randomized study designs and estimates of effect has focused only on policy interventions (Oliver 2010).

Why it is important to do this review

Despite the need for rigorous comparative effectiveness research, there has been no systematic comparison of effect measure estimates among all the types of randomized and non-experimental observational study designs that are being used to assess effectiveness of interventions. The findings of this review will inform the design of future comparative effectiveness research and help prioritize the types of context-specific study designs that should be used to minimize bias.

OBJECTIVES

To assess the impact of study design - to include RCTs versus observational study designs on the effect measures estimated.

To explore methodological variables that might explain any differences identified. Effect size estimates may be related to the underlying risk of bias (i.e., methodological variables) of the studies, and not the design *per se*. A flawed RCT may have larger effect estimates than a rigorous cohort study, for example. If the methodological reviews we included assessed the risk of bias of the study designs they included, we attempted to see if the differences in risk of bias explain any differences in effect size estimates.

To identify gaps in the existing research comparing study designs.

METHODS

Criteria for considering studies for this review

Types of studies

We examined systematic reviews that were designed as methodological reviews to compare quantitative effect size estimates measuring efficacy or effectiveness of interventions tested in trials with those tested in observational studies. For the purposes of this review, a methodological review is defined as a review that is designed to compare outcomes of studies that vary by a particular methodological factor (in this case, study design) and not to compare the clinical effect of an intervention to no intervention or a comparator. Comparisons included RCTs and observational studies (including retrospective cohorts, prospective cohorts, case-controls, and cross-sectional designs) that compared effect measures from different study designs or analyses. For this review, the only non-experimental studies we analyzed were observational in design. Therefore, we use the term "observational" in presenting the findings of our review. However, it should be noted that the terminology used in the literature to describe study designs is not consistent and can lead to confusion.

We included methodological reviews comparing studies described in the review as head to head randomized trials, cluster randomized trials, adaptive designs, practice / pragmatic / explanatory trials, PBE-CPI “practice based evidence for clinical practice improvement,” natural experiments, prospective and retrospective cohort studies, case-control studies, observational or cross-sectional studies of registries and databases including electronic medical records, or observational studies employing so-called causal inference techniques (e.g. briefly, analytical techniques that attempt to estimate a true causal relationship from observational data), which could include instrumental variables, marginal structural models, or propensity scores. Specifically, we included comparisons of estimates from RCTs with any of the above types of observational studies.

Our focus is on reviews of effectiveness or harms of health-related interventions. We included two types of reviews: a) systematic reviews of primary studies in which the review's main objective was pre-defined to include a comparison of study designs and not to answer one specific clinical research question; and b) methodological reviews of reviews that included existing reviews or meta-analyses that compared RCTs with observational designs. We excluded comparisons of study designs where the included studies were measuring the effects of putative harmful substances that are not health-related interventions, such as environmental chemicals, or diagnostic tests, as well as studies measuring risk factors or exposures to potential hazards. We excluded studies that compared randomized trials to non-randomized trials. For example, we excluded studies that compared studies with random allocation to those with non-random allocation or trials with adequate versus inadequate/unclear concealment of allocation. We also excluded studies that compared the results of meta-analyses with the results of single trials or single observational studies. Lastly, we excluded meta-analyses of the effects of an intervention that included both randomized trials and observational studies with an incidental comparison of the results.

Types of data

It was our intention to select reviews that quantitatively compared the efficacy or effectiveness of alternative interventions to prevent or treat a clinical condition or to improve the delivery of care. Specifically, our study sample included reviews that have effect estimates from RCTs or cluster-randomized trials and observational studies, which included, but were not limited to, cohort studies, case-control studies, cross-sectional studies.

Types of methods

We identified reviews comparing effect measures between trials and observational studies or different types of observational studies to include the following.

- RCTs/cluster-randomized trials versus prospective/ retrospective cohorts
- RCTs/cluster-randomized trials versus case-control studies
- RCTs/cluster-randomized trials versus cross-sectional studies
- RCTs/cluster-randomized trials versus other observational design
- RCTs/cluster-randomized trials versus observational studies employing so-called causal inference analytical methods

Types of outcome measures

The direction and magnitude of effect estimates (e.g. odds ratios, relative risks, risk difference) varied across meta-analyses included in this review. Where possible, we used odds ratios as the outcome measure in order to conduct a pooled odds ratio analysis.

Search methods for identification of studies

Electronic searches

To identify relevant methodological reviews we searched the following electronic databases, in the period from 01 January 1990 to 06 December 2013.

- Cochrane Methodology Register
- Cochrane Database of Systematic Reviews
- MEDLINE (via PubMed)
- EMBASE (via EMBASE.com)
- Literatura Latinoamericana y del Caribe en Ciencias de la Salud (LILACS)
- PsycINFO
- Web of Science/Web of Social Science

Along with MeSH terms and a wide range of relevant keywords, we used the sensitivity-specificity balanced version of a validated strategy to identify reviews in PubMed (Montori 2004), augmented with one term ("review" in article titles) so that it better targeted reviews. We anticipated that this strategy would retrieve all relevant reviews. See Appendix 1 for our PubMed search strategy, which was modified as appropriate for use in the other databases.

The search strategy was iterative, in that references of included reviews were searched for additional references. We used the "similar articles" and "citing articles" features of several of the databases to identify additional relevant articles. All languages were included.

Prior to executing the electronic searches, the search strategy was peer reviewed by a second information specialist, according to the Peer Review of Electronic Search Strategies (PRESS) guidance (Sampson 2009).

Data collection and analysis

The methodology for data collection and analysis was based on the guidance of *Cochrane Handbook of Systematic Reviews of Interventions* (Higgins 2011).

Selection of studies

After removing duplicate references, one review author (THH) screened the results, excluding those that were clearly irrelevant (e.g. animal studies, editorials, case studies).

Two review authors (AA and LB) then independently selected potentially relevant reviews by scanning the titles, abstracts, and descriptor terms of the remaining references and applying the inclusion criteria. Irrelevant reports were discarded, and the full article (or abstract if from a conference proceeding) was obtained for all potentially relevant or uncertain reports. The two review authors independently applied the inclusion criteria. Reviews were reviewed for relevance based on study design, types of methods employed, and a comparison of effects based on different methodologies or designs. THH adjudicated any disagreements that could not be resolved by discussion.

Data extraction and management

After an initial search and article screening, two review authors independently double-coded and entered information from each selected study onto standardized data extraction forms. Extracted information included the following.

- **Study details***:* citation, start and end dates, location, eligibility criteria, (inclusion and exclusion), study designs compared, interventions compared.
- **Comparison of methods details:** effect estimates from each study design within each publication.
- **Outcome details:** primary outcomes identified in each study.

Assessment of risk of bias in included studies

We included systematic reviews of studies therefore, The Cochrane Collaboration tool for assessing the risk of bias for individual studies does not apply. We used the following criteria to appraise the risk of bias of included reviews, which are similar to those used in the methodology review by Odgaard-Jensen and colleagues (Odgaard-Jensen 2011).

- Were explicit criteria used to select the studies?
- Did two or more investigators agree regarding the selection of studies?
- Was there a consecutive or complete sample of studies?
- Was the risk of bias of the included studies assessed?
- Did the review control for methodological differences of included studies (for example, with a sensitivity analysis)?
- Did the review control for heterogeneity in the participants and interventions in the included studies?
- Were similar outcome measures used in the included studies?
- Is there an absence of risk of selective reporting?
- Is there an absence of evidence of bias from other sources?

Each criterion was rated as yes, no or unclear.

We summarized the overall risk of bias of each study as: low risk of bias, unclear risk of bias or high risk of bias.

Measures of the effect of the methods

In general, outcome measures included relative risks or rate ratios (RR), odds ratios (OR), hazard ratios (HR).

Dealing with missing data

This review is a secondary data analysis and did not incur the missing data issues seen in most systematic reviews. However, for a select, small number of reviews we needed more information from the publishing authors regarding methods or other details, therefore, we contacted the corresponding authors.

Assessment of heterogeneity

We synthesized data from multiple reviews to compare effects from RCTs with observational studies. We had a wide variety of outcomes and interventions synthesized, increasing the amount of heterogeneity between reviews. We assessed heterogeneity using the χ^2 statistic with a significance level of 0.10, and the I^2 statistic. Together with the magnitude and direction of the effect, we interpreted an I^2 estimate between 30% and 60% as indicating moderate heterogeneity, 50% to 90% substantial heterogeneity, and 75% to 100% as a high level of heterogeneity. Furthermore, if an included study was, in fact, a review article that already assessed heterogeneity, we reported the authors' original assessment of heterogeneity.

Assessment of reporting biases

We attempted to minimize the potential for publication bias by our comprehensive search strategy that included evaluating published and unpublished literature. In cases where we were missing specific information or data, we contacted authors and requested additional data.

Data synthesis

We examined the relationship between study design type and the affiliated estimates. Using results from observational studies as the reference group, we examined the published estimates to see whether there was a relative smaller or larger effect. We explored whether the RCT comparators showed about the same effects, larger treatment effects, or smaller treatment effects compared to the observational study reference group. Furthermore, in the text we qualitatively described the reported results from each included review. Within each identified review, if an estimate comparing results from RCTs with observational studies was not provided, we pooled the estimates for observational studies and RCTs. Then, using methods described by Altman (Altman 2003), we estimated the ratio of ratios (hazard ratio or risk ratio or odds ratio) for each included review using observational studies as the reference group. Across all reviews, we synthesized these ratios to get a pooled ratio of odds ratios (ROR) comparing results from RCTs to results from observational studies. Our results varied considerably by comparison groups, outcomes, interventions, and study design, which contributed greatly to heterogeneity. To avoid overlap of

data between included studies, we did not include data previously included in another included review.

Subgroup analysis and investigation of heterogeneity

Reducing bias in comparative effectiveness research is particularly important for studies comparing pharmacological interventions with their implications for clinical care and health care purchasing. Since a number of the studies comparing study designs used for comparative effectiveness research focused on pharmacological comparisons, we decided, *a priori*, to conduct a subgroup analysis of these pharmacological studies. Specifically, we hypothesized that studies of pharmacological comparisons in a randomized design may have smaller effect estimates than studies of pharmacological comparisons in an observational study.

Additionally, we performed a subgroup analysis by heterogeneity of the included methodological reviews to compare the differences between RCTs and observational studies from the subgroup of methodological reviews with high heterogeneity (as measured in their respective meta-analysis) to those with moderate-low heterogeneity. As such, we stratified the reviews by the heterogeneity *within* each methodology review.

RESULTS

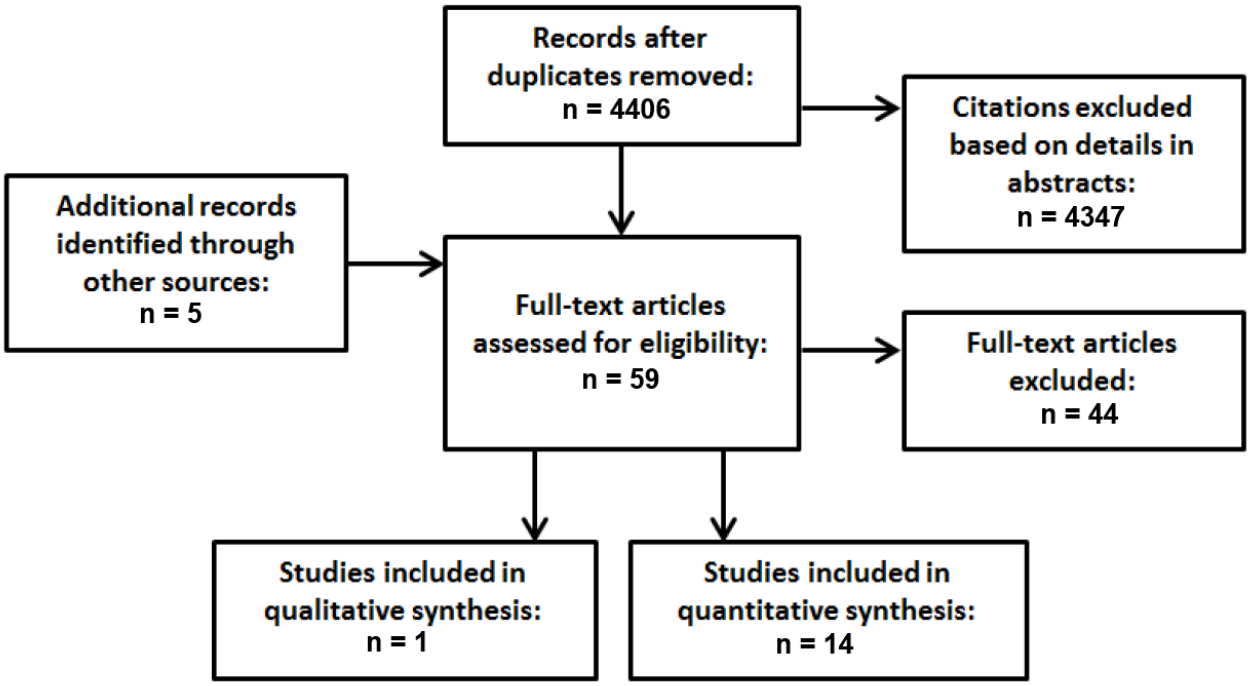
Description of studies

See Characteristics of included studies; Characteristics of excluded studies.

Results of the search

Our initial search yielded 4406 unique references. An additional five references were identified from checking the reference lists of included publications. We selected 59 full-text articles for further review, of which 44 were excluded because they did not meet our inclusion criteria. Fifteen reviews met our inclusion criteria for this review; 14 of these reviews were included in the quantitative analysis. See Figure 1 for study selection chart.

Figure 1. Flow chart depicting screening process



Included studies

See Characteristics of included studies. Fifteen reviews, published between 01 January 1990 and 06 December 2013, met the inclusion criteria for this review. Fourteen papers compared RCTs with observational designs; two reviews focused exclusively on pharmacological interventions (Beynon 2008; Naudet 2011), while four focused on pharmacological and other interventions, but provided data on drugs that could be analyzed separately (Benson 2000; Concato 2000; Golder 2011; Ioannidis 2001).

The included reviews analyzed data from 1583 meta-analyses that covered 228 different medical conditions. The mean number of included studies per paper was 178 (range 19 to 530).

Of the 15 reviews, 14 were included in the quantitative analysis and had data, or we were able to obtain quantitative data from the authors, that allowed us to calculate RORs. One study (Papanikolauo 2006) was included in a previously published review (Golder 2011), therefore we have described it, but did not include it in the meta-analysis.

Benson 2000 et al searched the Abridged Index Medicus and Cochrane databases for observational studies published between 1985 and 1998 that compared two or more treatments. To identify RCTs and observational studies comparing the same treatment, the researchers searched MEDLINE and Cochrane databases. One hundred and thirty-six publications were identified that covered 19

different treatments. Benson 2000 et al found little evidence that treatment effect estimates obtained from observational studies were consistently larger than estimates from RCTs.

Beynon 2008 et al attempted to identify all observational and randomized studies with all-cause mortality as the outcome for a sample of topics selected at random from the medical literature. One hundred and fourteen RCTs and 19 observational studies on 19 topics were included. The ratio of RRs for RCTs compared to observational studies was 0.88 (0.8 to 0.97), suggesting that observational studies had larger treatment effects by 12% on average.

Bhandari 2004 et al conducted a MEDLINE search for both observational and randomized studies comparing internal fixation and arthroplasty in patients with femoral neck fractures in publications between 1969 and 2002. The authors found 27 studies that met the criteria. Bhandari 2004 et al found that observational studies underestimated the relative benefit of arthroplasty by 19.5%.

Concato 2000 et al searched MEDLINE for meta-analyses of RCTs and observational studies of the same intervention published in five major journals between 1991 and 1995. From 99 reports on five clinical topics, observational studies, on average, were similar to RCTs. The authors concluded that well-designed observational studies generally do not have larger effects of treatment when compared to results of RCTs.

Edwards 2012 et al performed a systematic review and meta-analysis comparing effect estimates evaluating the effects of surgical procedures for breast cancer in both RCTs and observational studies. A search of MEDLINE, EMBASE, and Cochrane Databases (2003 to 2008) yielded 12 RCTs covering 10 disparate outcomes. In two of 10 outcomes the pooled estimates from RCTs and observational studies differed, though none significantly. The authors conclude that RCTs comparing breast surgery procedures may yield different estimates in 20% to 40% of cases compared with estimates from observational studies.

Furlan 2008 et al searched for comparative studies of low-back pain interventions published in MEDLINE, EMBASE, or *The Cochrane Library* through May 2005 and included interventions with the highest numbers of non-randomised studies. Seventeen observational studies and eight RCTs were identified and, in general, results from observational studies either agreed with results from RCTs or underestimated the effects when compared to RCTs.

Golder 2011 et al performed a meta-analysis of meta-analyses comparing estimates of harm derived from meta-analysis of RCTs with meta-analyses of observational studies. Fifty-eight meta-analyses were identified. Pooled relative measures of adverse effect (odds ratio (OR) or risk ratio (RR)) suggested no difference in effect between study type (OR = 1.03; 95% confidence interval (CI) 0.93-1.15). The authors conclude that there is no evidence on average in effect estimate of adverse effect of interventions from meta-analyses of RCTs when compared to observational studies.

Ioannidis 2001 et al performed an analysis of meta-analyses comparing effect estimates evaluating medical interventions from meta-analysis of RCTs to meta-analyses of observational studies. A search of MEDLINE (1966to 2000) and *The Cochrane Library* (2000, Issue 3) and major journals yielded 45 diverse topics from 240 RCTs and 168 observational studies. Observational studies tended to show larger treatment effects (P = 0.009). The authors conclude that despite good correlation between RCTs and observational studies, differences in effect sizes are present.

Kuss 2011 et al performed a systematic review and meta-analysis comparing effect estimates from RCTs with observational studies employing propensity scores The included studies examined the effects of off-pump versus on-pump surgery in similar populations. A MEDLINE search yielded 29 RCTs and 10 propensity score analyses covering 10 different outcomes. For all outcomes, no differences were noted between RCTs and propensity score analyses.

The authors conclude that RCTs and propensity score analyses will likely yield similar results and propensity score analyses may have only a small remaining bias compared to RCTs.

Lonjon 2013 et al performed a systematic review and meta-analysis comparing effect estimates from RCTs with observational studies employing propensity scores studying the effects of surgery addressing the same clinical question. A MEDLINE search yielded 94 RCTs and 70 propensity score analyses covering 31 clinical questions. For all-cause mortality the authors noted no differences between RCTs and propensity score analyses (ROR = 1.07; 95% CI 0.87 to 1.33).

The authors conclude that RCTs and propensity score analyses will likely yield similar results in surgery studies.

Müeller 2010 et al searched PubMed for RCTs and observational studies comparing laparoscopic versus open cholecystectomy. A total of 162 studies were identified for inclusion (136 observational and 26 RCTs). Among the 15 outcomes of interest, three yielded significant discrepancies in effect sizes between study designs. As such, the authors conclude that the results from observational studies and RCTs differ significantly in at least 20% of outcomes variables.

Naudet 2011 et al identified published and unpublished studies from 1989 to 2009 that examined fluoxetine and venlafaxine as first line treatment for major depressive disorder. The authors identified 12 observational studies and 109 RCTs and produced meta-regression estimates for outcomes of interest. The standardized treatment response in RCTs was greater by a magnitude of 4.59 compared to observational studies and the authors conclude that the response to antidepressants is greater in RCTs than in observational studies.

Oliver 2010 et al identified systematic reviews that compared results of policy interventions, stratifying estimates by observational study and RCT study design published between 1999 and 2004. A total of 16 systematic reviews were identified, with a median of 11.5 RCTs and 14.5 observational studies in each systematic review. Observational studies published in systematic reviews were pooled separately from RCTs published in the same systematic reviews. Results that were stratified by study design were heterogeneous with no clear differences in magnitude of effects; the authors found no evidence for clear systematic differences in terms of results between RCTs and observational studies.

Shikata 2006 et al identified all meta-analyses of RCTs of digestive surgery published between 1966 and 2004. Fifty-two outcomes for 18 disparate topics were identified from 276 articles (96 RCTs and 180 observational studies). Pooled odds ratios and relative risks were extracted for each outcome, using the same indicator that had been used in the meta-analysis of interest and approximately 25% of all outcomes of interest yielded different results between observational studies and RCTs.

Papanikolauo 2006 et al compared evidence from RCTs with observational studies that explored the effects of interventions on the risk of harm. Harms of interest were identified from RCTs with more than 4000 patients. Observational studies of more than 4000 patients were also included for comparison. Fifteen harms of interest were identified and relative risks were extracted for 13 topics. Data from 25 observational studies were compared with results from RCTs. Relative risks for each outcome/harm were calculated for both study types. The estimated increase in RR differed by more than two-fold between observational studies and RCTs for 54% of the topics studied. The authors conclude that observational studies usually under-estimate the absolute risk of harms. These data were included in Golder 2011 and consequently were not re-analyzed in the current quantitative analysis.

Excluded studies

See Characteristics of excluded studies. Following full-text screening, 44 studies were excluded from this review. The main reasons for exclusion included the following: the studies were meta-analyses that did an incidental comparison of RCTs and observational studies, but were not designed for such a comparison (n = 14); the studies were methodological or statistical papers that did not conduct a full systematic review of the literature (n = 28); or the studies included quasi- or pseudo-randomized studies, or provided no numerical data that would allow a quantitative comparison of effect estimates (n = 7).

Risk of bias in included studies

Eleven reviews had low risk of bias for explicit criteria for study selection (Benson 2000; Beynon 2008; Bhandari 2004; Edwards 2012; Furlan 2008; Ioannidis 2001; Kuss 2011; Müeller 2010; Naudet 2011; Oliver 2010; Papanikolauo 2006); nine (60%) had low risk of bias for investigators' agreement for study selection (Bhandari 2004; Concato 2000; Edwards 2012; Golder 2011; Kuss 2011; Naudet 2011; Oliver 2010; Papanikolauo 2006; Shikata 2006); five (33%) included a complete sample of studies (Bhandari 2004; Müeller 2010; Naudet 2011; Oliver 2010; Shikata 2006); seven (47%) assessed the risk of bias of their included studies (Bhandari 2004; Furlan 2008; Golder 2011; Lonjon 2013; Müeller 2010; Naudet 2011; Oliver 2010); seven (47%) controlled for methodological differences between studies (Furlan 2008; Ioannidis 2001; Kuss 2011; Lonjon 2013; Müeller 2010; Naudet 2011; Oliver 2010); eight (53%) controlled for heterogeneity among studies (Beynon 2008; Edwards 2012; Furlan 2008; Ioannidis 2001; Lonjon 2013; Müeller 2010; Naudet 2011; Oliver 2010); nine (60%) analyzed similar outcome measures (Benson 2000; Beynon 2008; Bhandari 2004; Edwards 2012; Ioannidis 2001; Lonjon 2013; Müeller 2010; Oliver 2010; Shikata 2006); and only four (27%) were judged to be at low risk of reporting bias (Bhandari 2004; Furlan 2008; Ioannidis 2001; Naudet 2011).

We rated reviews that were coded as adequate for explicit criteria for study selection, complete sample of studies, and controlling for methodological differences and heterogeneity as having a low risk of bias and all others as having a high risk of bias. Two reviews, Müeller 2010 and Naudet 2011, met all four of these criteria and, thus, had an overall low risk of bias.

See Figure 2; Figure 3.

Figure 2. 'Risk of bias' graph: review authors' judgements about each risk of bias item presented as percentages across all included studies.

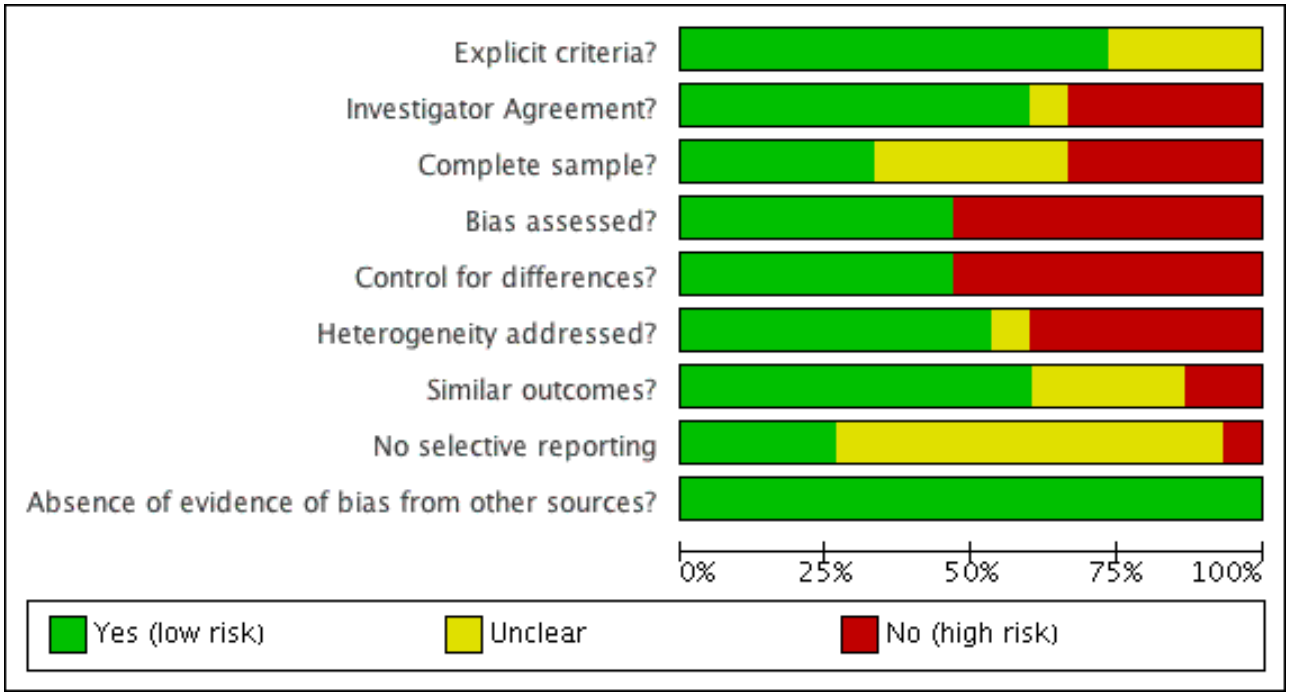
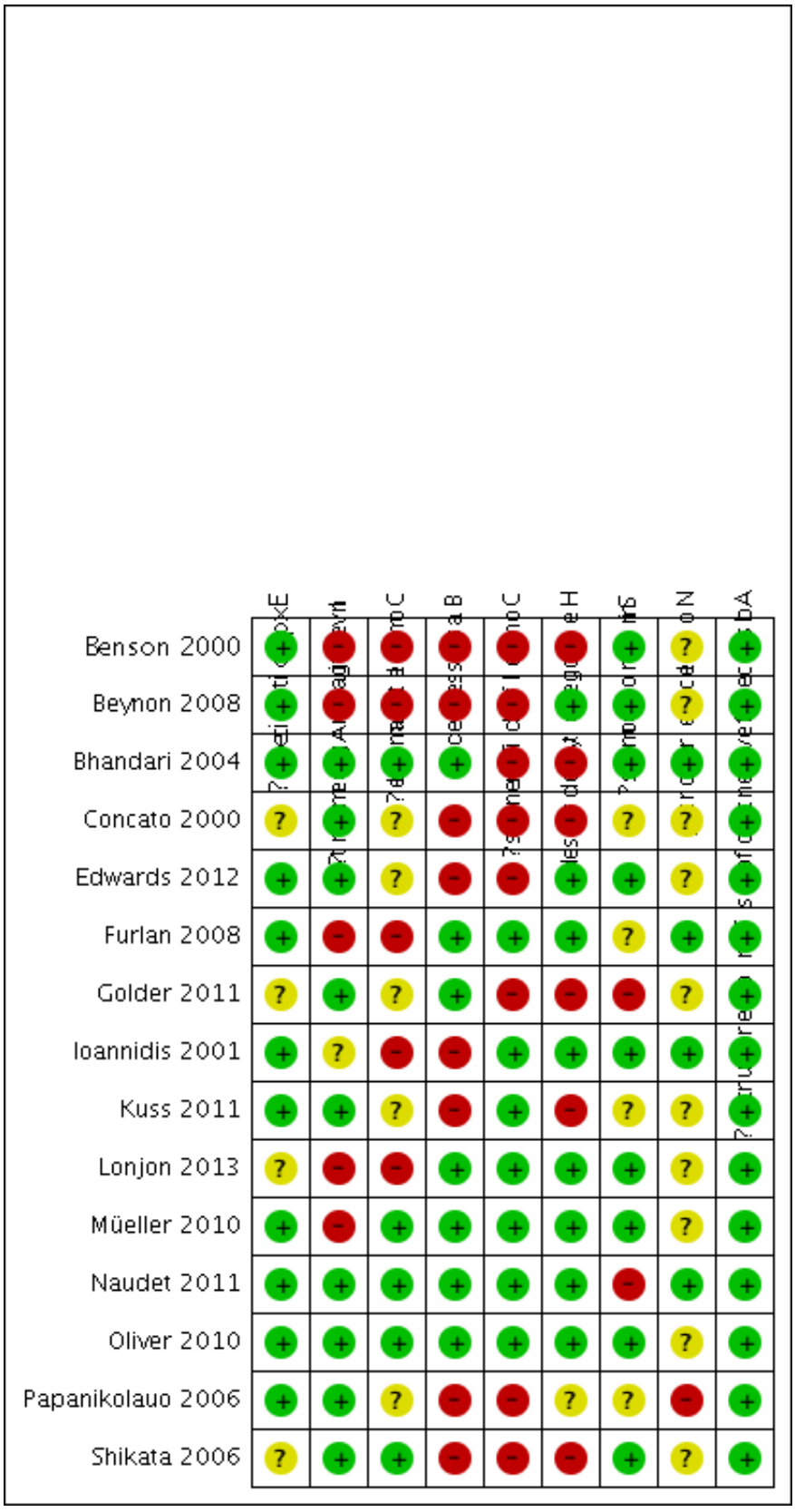


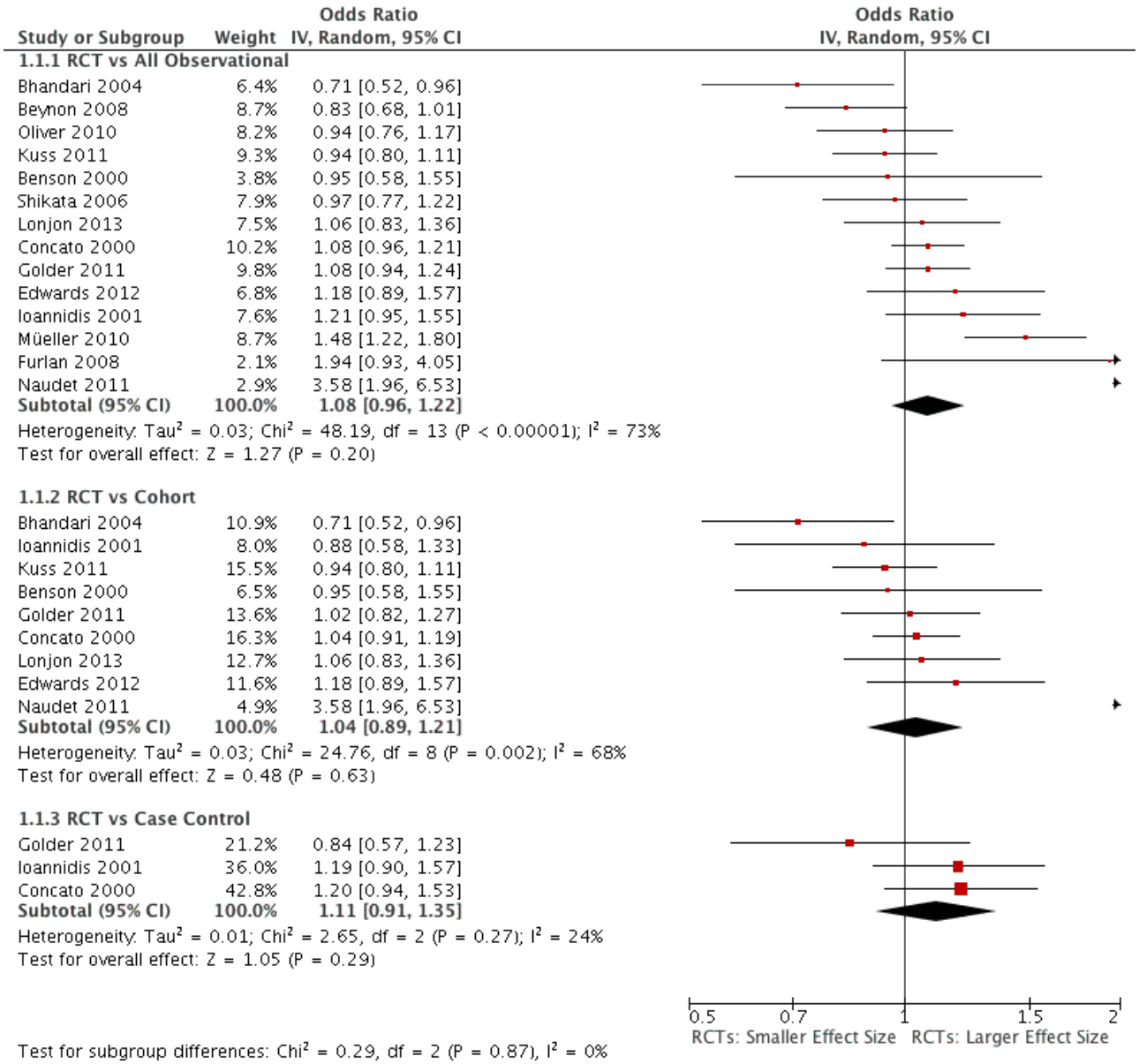
Figure 3. 'Risk of bias' summary: review authors' judgements about each risk of bias item for each included study.



Effect of methods

Our primary quantitative analysis (Analysis 1.1), including 14 reviews, showed that the pooled ratio of odds ratios (ROR) comparing effects from RCTs with effects from observational studies was 1.08 (95% confidence interval (CI) 0.96 to 1.22) (see Figure 4). There was substantial heterogeneity for this estimate (I^2 = 73%). Of the 14 reviews included in this analysis, 11 (71%) found no significant difference between observational studies and RCTs. However, one review suggested observational studies have larger effects of interest (Bhandari 2004), while two other reviews suggested observational studies have smaller effects of interest (Müeller 2010; Naudet 2011).

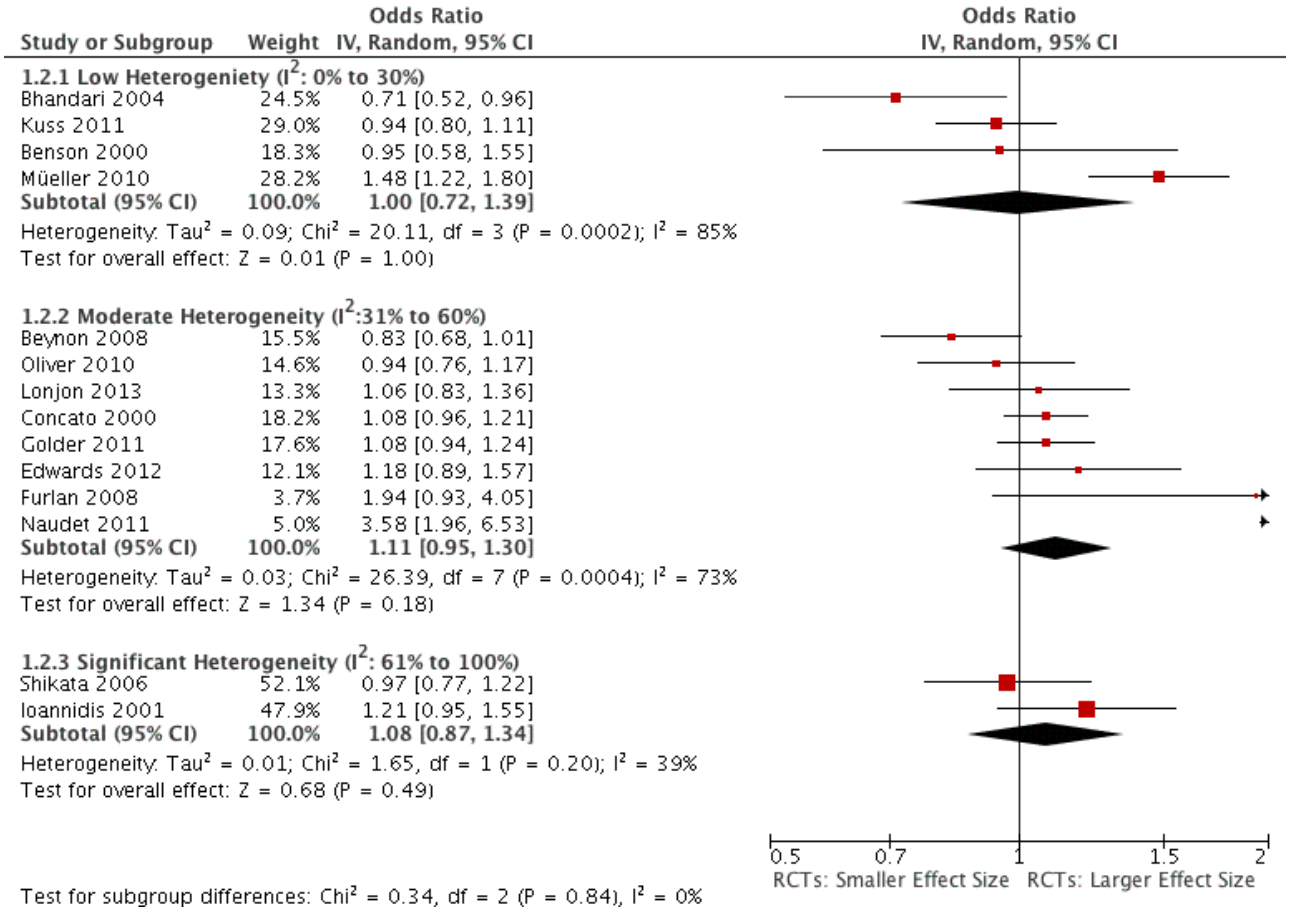
Figure 4. Forest plot of comparison: 1 RCT vs Observational, outcome: 1.2 Pooled Ratio of Odds Ratios--Study Design.



When possible or known, we isolated our results to reviews that specifically compared cohort studies and RCTs. Nine reviews either provided adequate data or performed these analyses in their publication (Benson 2000; Bhandari 2004; Concato 2000; Edwards 2012; Golder 2011; Ioannidis 2001; Kuss 2011; Lonjon 2013; Naudet 2011) Similar to the effect across all included reviews, the effects from RCTs compared with cohort studies was pooled ROR = 1.04 (95% CI 0.89 to 1.21), with substantial heterogeneity (I^2 = 68%) (Analysis 1.1.2). In lieu of a sensitivity analysis removing case-control studies, we performed a subgroup analysis of reviews that compared the effects of case-controls versus RCTs (Concato 2000; Golder 2011; Ioannidis 2001). The pooled ROR comparing RCTs with case-control studies was 1.11 (95% CI 0.91 to 1.35), with minor heterogeneity (I^2 = 24%). There was no significant difference between observational study design subgroups (P value = 0.61).

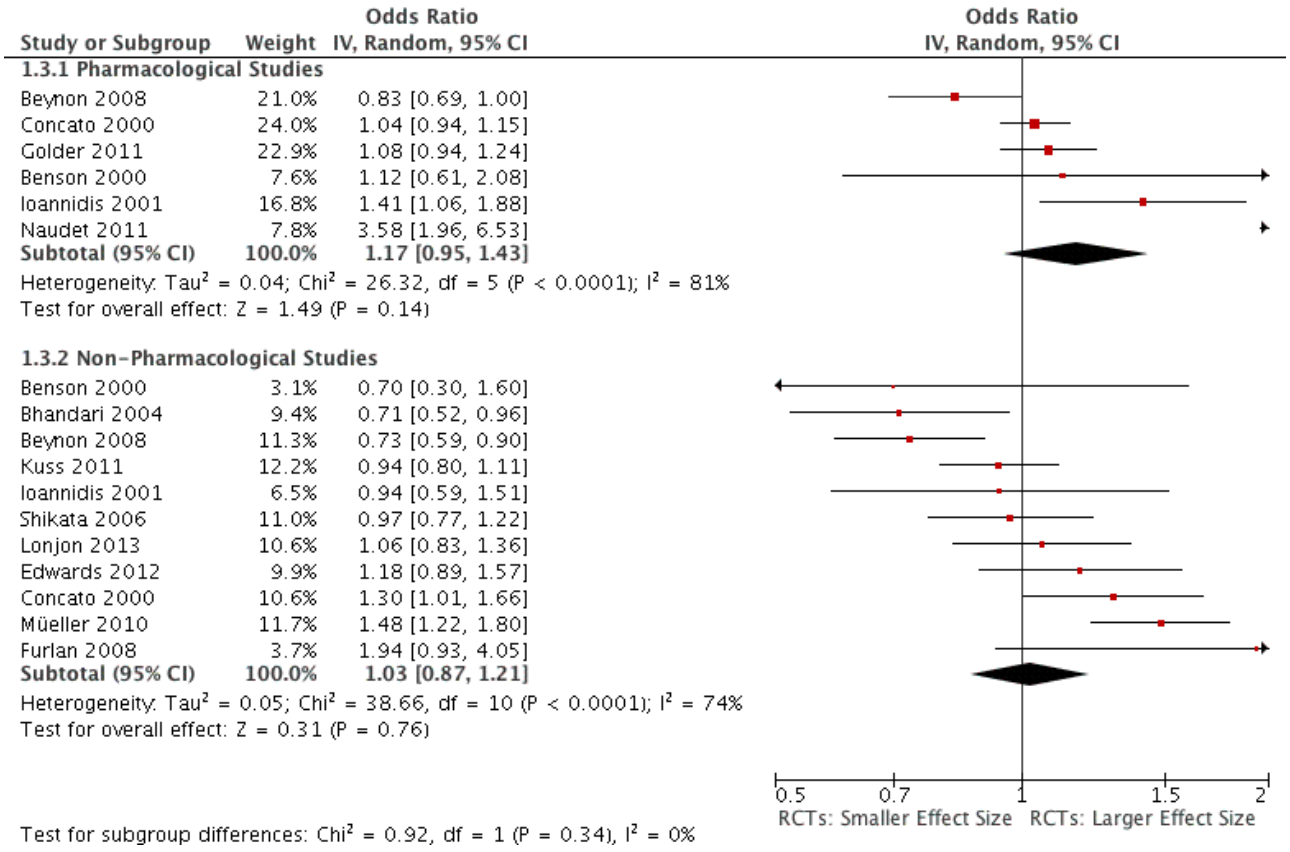
We also performed a subgroup analysis of all reviews stratified by levels of heterogeneity of the pooled RORs from the respective reviews (Analysis 1.2). No significant difference in point estimates across heterogeneity subgroups were noted (see Figure 5). Specifically, comparing RCTs with observational studies in the low heterogeneity subgroup yielded a pooled ROR of 1.00 (95% CI 0.72 to 1.39). The pooled ROR comparing RCTs with observational studies in the moderate heterogeneity group was also not significantly different (OR = 1.11; 95% CI 0.95 to 1.30). Similarly, the pooled ROR comparing RCTs with observational studies in the significant heterogeneity group was 1.08 (95% CI 0.87 to 1.34).

Figure 5. Forest plot of comparison: 1 RCT vs Observational, outcome: 1.3 Pooled Ratio of Odds Ratios--Heterogeneity Subgroups.



Additionally, we performed a subgroup analysis of all included reviews stratified by whether they compared pharmacological studies or not (Analysis 1.3). Though the pooled ROR for comparisons of pharmacological studies was higher than the pooled ROR for reviews of non-pharmacological studies, this difference was not significant (see Figure 6) (P value = 0.34). Namely, the pooled ROR comparing RCTs with observational studies in the pharmacological studies subgroup of six reviews was 1.17 (95% CI 0.95 to 1.43), with substantial heterogeneity (I^2 = 81%). The pooled ROR comparing RCTs with observational studies in the non-pharmacological studies subgroup of 11 reviews was 1.03 (95% CI 0.87 to 1.21), with substantial heterogeneity (I^2 = 74%).

Figure 6. Forest plot of comparison: 1 RCT vs Observational, outcome: 1.4 Pooled Ratio of Odds Ratios--Pharmacological Studies Subgroups.



Lastly, we performed an analysis of all included reviews that compared RCTs and observational studies that employed propensity score adjustments (Analysis 1.4). The pooled ROR comparing estimates from RCTs with the estimates from observational studies using propensity scores was not significant. Namely, the pooled ROR comparing RCTs with observational studies with propensity scores (two reviews) was 0.98 (95% CI 0.85 to 1.12), with no heterogeneity (I^2 = 0%). There was no difference between the pooled ROR of RCTs versus observational studies with propensity score adjustment and the pooled ROR of RCTs versus observational studies without propensity score adjustment (P value = 0.22).

DISCUSSION

Summary of main results

Our results showed that, on average, there is little difference between the results obtained from RCTs and observational studies. In addition, despite several subgroup analyses, no significant differences between effects of study designs were noted. However, due to high statistical heterogeneity, there may be important differences between subgroups of reviews that we were unable to identify, Our primary quantitative analysis showed that the pooled ROR comparing effects from RCTs with effects from observational studies was 1.08 (95% CI 0.96 to 1.22). The effects from RCTs compared with cohort studies only was pooled ROR = 1.04 (95% CI 0.89 to 1.21), while the pooled ROR comparing RCTs with only case-control studies was1.11 (95% CI 0.91 to 1.35).

Though not significant, the point estimates suggest that observational studies may have smaller effects than those obtained in RCTs, regardless of observational study design. Furthermore, it is possible that the difference between effects obtained from RCTs and observational studies has been somewhat attenuated in more recent years due to researchers' improved understanding of how to handle adjustments in observational studies. In the present study, it was not always very clear which observational studies included adjusted estimates and which did not in the included reviews. Bhandari et al reported that no observational study adjusted for all nine confounders the authors felt were important (Bhandari 2004). In fact, they adjusted for as few as two and as many as six. Mueller et al reported that of the 136 non-RCTs included in their review, 19 population-based studies and 22 other studies adjusted their results for baseline imbalances (Müeller 2010). Two reviews included only observational studies with propensity score adjustments (Kuss 2011; Lonjon 2013). Other included reviews note the importance of adjustment in the estimates from observational studies, but do not specifically list the studies with and without adjusted estimates. Our results suggest that although observational designs may be more biased than RCTs, this does not consistently result in larger or smaller intervention effects.

We also found that the effect estimate differences between observational studies and RCTs were potentially influenced by the heterogeneity within meta-analyses. Though subgroup analyses comparing heterogeneity groups were not statistically significant, meta-analyses comparing RCTs and observational studies may be

particularly influenced by heterogeneity and researchers should consider this when designing such comparisons. However, with so few reviews, spurious effects between heterogeneity subgroups cannot be ruled out.

The risks of bias in the included reviews were generally high. In particular, two-thirds of all included reviews either did not include a complete sample or there was not enough information provided to make a determination, and more than half of the reviews did not assess the risk of bias of their included studies. Furthermore, nearly three-quarters of the included reviews were judged to be at high or unclear risk of reporting bias.

We note that our results may be influenced by the different comparison arms in all the studies included in the reviews. Often the specific types of comparison arms in the meta-analyses were not identified in the review. However, among included reviews with reported details about comparison arms in the RCTs in the meta-analyses (n = 519 meta-analyses), 84% (n = 454) compared one intervention (e.g., drug or surgery) with another intervention (drug or surgery), 11% (n = 55) used a placebo or sham, 3% (n = 13) used an unspecified control arm, and 2% (n = 15) compared one intervention with no intervention or treatment.

Lastly, though not statistically significant, there appears to be a difference in effect comparing RCTs and observational studies when considering studies with pharmacological-only interventions or studies without pharmacological interventions. More specifically, the difference in point estimates between pharmacological RCTs and observational pharmacological studies is greater than the difference in point estimates from non-pharmacological studies. Perhaps this is a reflection of the difficulties in removing all potential confounding in observational pharmacological studies; or, perhaps this is an artifact of industry or selective reporting bias in pharmacological RCTs. The most recent study quantifying pharmaceutical industry support for drug trials found that the pharmaceutical industry funded 58% of drug trials in 2007 and this was the largest source of funding for these trials (Dorsey 2010). This is not surprising as RCTs must be submitted to regulatory agencies to obtain regulatory approval of drugs, whereas observational studies of drugs are conducted after drug approval. Funding and selective reporting bias have been well documented in industry-sponsored RCTs (Lundh 2012) and less is known about the extent of these biases in observational studies.

Potential biases in the review process

We reduced the likelihood for bias in our review process by having no language limits for our search and having two review authors independently screen abstracts and articles for selection. Nevertheless, we acknowledge the potential for introduction of unknown bias in our methods as we collected a myriad of data from 14 reviews (1583 meta-analyses covering 228 unique outcomes).

Agreements and disagreements with other studies or reviews

Our results across all reviews (pooled ROR 1.08; 95% CI 0.96 to 1.22) are very similar to results reported by Concato 2000 and Golder 2011. As such, we have reached similar conclusions--there is little evidence for significant effect estimate differences between observational studies and RCTs, regardless of specific observational study design, heterogeneity, or inclusion of drug studies.

Golder 2011 (and consequently, Papanikolauo 2006) and Edwards 2012) were the only reviews that focused on harm outcomes. Golder's findings do not support the notion that observational studies are more likely to detect harm than randomized controlled trials, as no differences in RCTs and observational studies were detected. However, this finding may be related to the short-term nature of the adverse events studied where one would expect shorter-term trials to be as likely to detect harm as longer-term observational studies.

AUTHORS' CONCLUSIONS

Implication for methodological research

In order to understand why RCTs and observational studies addressing the same question sometimes have conflicting results, methodological researchers must look for explanations other than the study design *per se*. Confounding is the greatest bias in an observational study compared to an RCT and methods for accounting for confounding in meta-analyses of observational studies should be developed (Reeves 2013). The Patient-Centered Outcomes Research Institute is finalizing methodological standards and calling for more research on measuring confounding in observational studies(PCORI 2012). PCORI has also called for empirical data to support the constitution of propensity scores and the validity of instrumental variables, two methods used to control for confounding in observational studies.

REFERENCES

References to studies included in this review

Benson 2000 *{published data only}*

Benson K, Hartz A. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine* 2000;**342**(25):1878-86.

Beynon 2008 *{published data only}*

* Beynon R, Harris R, Sterne JAC, et al. The quantification of bias in randomised and non-randomised studies: the BRANDO NRS Study [Poster]. 16th Cochrane Colloquium. Freiburg im Breisgau, Germany, 3-7 October, 2008.

Bhandari 2004 *{published data only}*

Bhandari M, Tornetta PIII, Ellis T, Audige L, Sprague S, Kuo JC, et al. Hierarchy of evidence: differences in results between non-randomized studies and randomized trials in patients with femoral neck fractures. *Archives of Orthopaedic and Trauma Surgery* 2004;**124**(1):10-6.

Concato 2000 *{published data only}*

Concato J, Shah N, Horwitz R. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine* 2000;**342**(25):1887-92.

Edwards 2012 *{published data only}*

Edwards J, Kelly E, Lin Y, Lenders T, Ghali W, Graham A. Meta-analytic comparison of randomized and nonrandomized studies of breast cancer surgery. *Canadian Journal of Surgery* 2012;**55**(3):155-62.

Furlan 2008 *{published data only}*

Furlan A, Tomlinson G, Jadad A, Bombardier C. Examining heterogeneity in meta-analysis: comparing results of randomized trials and nonrandomized studies of interventions for low back pain. *Spine* 2008;**33**(3):339-48.

Golder 2011 *{published data only}*

Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological review. *PLoS Medicine* 2011;**8**(5):e1001026.

Ioannidis 2001 *{published data only}*

Ioannidis J, Haidich A, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;**286**(7):821-30.

Kuss 2011 *{published data only}*

Kuss O, Legler T, Boergermann J. Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. *Journal of Clinical Epidemiology* 2011;**64**:1076-84.

Lonjon 2013 *{published data only}*

Lonjon G, Boutron I, Trinquart L, Ahmad N, Aim F, Nizard R, et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. *Annals of Surgery* 2013;**259**(1):18-25.

Müeller 2010 *{published data only}*

Müeller D, Sauerland S, Neugebauer EA, Immenroth M. Reported effects in randomized controlled trials were compared with those of nonrandomized trials in cholecystectomy. *Journal of Clinical Epidemiology* 2010;**63**(10):1082-90.

Naudet 2011 *{published data only}*

Naudet F, Maria AS, Falissard B. Antidepressant response in major depressive disorder: a meta-regression comparison of randomized controlled trials and observational studies. *PLoS One* 2011;**6**(6):e20811.

Oliver 2010 *{published data only}*

Oliver S, Bagnall AM, Thomas J, Shepherd J, Sowden A, White I, et al. Randomised controlled trials for policy interventions: a review of reviews and meta-regression. *Health Technology Assessment* 2010;**14**(16):1.

Papanikolauo 2006 *{published data only}*

Papanikolaou P, Christidi G, Ioannidis J. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *CMAJ: Canadian Medical Association Journal* 2006;**174**(5):635-41.

Shikata 2006 *{published data only}*

Shikata S, Nakayama T, Noguchi Y, Taji Y, Yamagishi H. Comparison of effects in randomized controlled trials with observational studies in digestive surgery. *Annals of Surgery* 2006;**244**(5):668-76.

References to studies excluded from this review

Ather 2011 *{published data only}*

Ather S, Bangalore S, Vemuri S, Cao LB, Bozkurt B, Messerli FH. Trials on the effect of cardiac resynchronization on arterial blood pressure in patients with heart failure. *American Journal of Cardiology* 2011;**107**(4):561-78.

Begg 1991 *{published data only}*

Begg C, Pilote L. A model for incorporating historical controls into a meta-analysis. *Biometrics* 1991;**47**(3):899-906.

Beyersmann 2008 *{published data only}*

Beyersmann J, Gastmeier P, Wolkewitz M, Schumacher M. An easy mathematical proof showed that time-dependent bias inevitably leads to biased effect estimation. *Journal of Clinical Epidemiology* 2008;**61**(12):1216-21.

Bosco 2010 *{published data only}*

Bosco J, Silliman R, Thwin S, Geiger AM, Buist DS, Prout MN, et al. A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. *Journal of Clinical Epidemiology* 2010;**3**(1):64-74.

Britton 1998 *{published data only}*

Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. *Health Technology Assessment* 1998;**2**(13):1-124.

Chambers 2010 *{published data only}*

Chambers D, Fayter D, Paton F, Woolacott N. Use of non-randomised evidence alongside randomised trials in a systematic review of endovascular aneurysm repair: strengths and limitations. *European Journal of Vascular and Endovascular Surgery* 2010;**39**(1):26-34.

Coulam 1994 *{published data only}*

Coulam CB, Clark DA, Collins J, Scott JR, Schlesselman JS, Aoki K, et al. Recurrent Miscarriage Immunotherapy Trialists Group. Worldwide collaborative observational study and meta-analysis on allogenic leukocyte immunotherapy for recurrent spontaneous abortion. *American Journal of Reproductive Immunology* 1994;**32**(2):55-72.

Dahabreh 2012 *{published data only}*

Dahabreh I, Sheldrick R, Paulus J, Chung M, Varvarigou V, Jafri H, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *European Heart Journal* 2012;**33**:1893-901.

Deeks 2002 *{published data only}*

Deeks JJ, D'Amico R, Sakarovitch C, et al. Are comparability of case-mix and the use of statistical adjustment markers of quality in non-randomised studies? An empirical investigation. 4th Symposium on Systematic Reviews: Pushing the Boundaries. Oxford, UK, 2002.

Deeks 2003 *{published data only}*

Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. International Stroke Trial Collaborative Group, European Carotid Surgery Trial Collaborative Group. Evaluating non-randomised intervention studies. *Health Technology Assessment* 2003;**7**(27):1-173.

Diehl 1986 *{published data only}*

Diehl L, Perry D. A comparison of randomized concurrent control groups with matched historical control groups: are historical controls valid?. *Journal of Clinical Oncology* 1986;**4**(7):1114-20.

Diez 2010 *{published data only}*

Diez P, Vogelius IS, Bentzen SM. A new method for synthesizing radiation dose-response data from multiple trials applied to prostate cancer:. *International Journal of Radiation Oncology, Biology, Physics* 2010;**77**(4):1066-71.

Flossmann 2007 *{published data only}*

Flossmann E, Rothwell P. Effect of aspirin on long-term risk of colorectal cancer: consistent evidence from randomised and observational studies. *Lancet* 2007;**369**:1603-13.

Hallstrom 2000 *{published data only}*

Hallstrom A, Anderson JI, Cobb La, Friedman PL, Herre JM, Klein RC, et al. Advantages and disadvantages of trial designs: a review of analysis methods for ICD studies. *Pacing and Clinical Electrophysiology: PACE* 2000;**23**(6):1029-38.

Henry 2001 *{published data only}*

Henry D, Moxey A, O'Connell D. Agreement between randomized and non-randomized studies: the effects of bias and confounding. 9th Cochrane Colloquium. Lyon, France, 9-13 October, 2001.

Hlatky 1988 *{published data only}*

Hlatky MA, Califf RM, Harrell FE Jr, Lee KL, Mark DB, Pryor DB. Comparison of predictions based on observational data with the results of randomized controlled clinical trials of coronary artery bypass surgery. *Journal of the American College of Cardiology* 1988;**11**(2):237-45.

Ioannidis 2005 *{published data only}*

Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005;**294**(2):218-28.

Labrarere 2006 *{published data only}*

Labarere J, Bosson JL, Sevestre MA, Delmas AS, Dupas S, Thenault MH, et al. Graduated compression stocking thromboprophylaxis for elderly inpatients. *Journal of General Internal Medicine* 2006;**21**(12):1282-7.

LaTorre 2009 *{published data only}*

LaTorre G, de Waure C, Specchia ML, Nicolotti N, Capizzi S, Bilotta A, et al. Does quality of observational studies affect the results of a meta-analysis?: the case of cigarette smoking and pancreatic cancer. *Pancreas* 2009;**38**(3):241-7.

Linde 2007 *{published data only}*

Linde K, Streng A, Hoppe A, Weidenhammer W, Wagenpfeil S, Melchart D. Randomized trial vs. observational study of acupuncture for migraine found that patient characteristics differed but outcomes were similar. *Journal of Clinical Epidemiology* 2007;**60**(3):280-7.

Lipsey 1993 *{published data only}*

Lipsey M, Wilson D. The efficacy of psychological, educational, and behavioral treatment. *American Psychologist* 1993;**48**(12):1181-209.

Loke 2011 *{published data only}*

Loke Y, Cavallazzi R, Singh S. Risk of fractures with inhaled corticosteroids in COPD: systematic review and meta-analysis of randomised controlled trials and observational studies. *Thorax* 2011;**66**:699-708.

MacLehose 2000 *{published data only}*

MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AM. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment* 2000;**4**(34):1-154.

Mak 2009 *{published data only}*

Mak A, Cheung MW, Chun-Man Ho R, Ai-Cia Cheak A, Chak Sing Lau C. Bisphosphonates and atrial fibrillation: Bayesian meta-analyses of randomized controlled trials and observational studies. *BMC Musculoskeletal Disorders* 2009;**10**:113.

McCarron 2010 *{published data only}*

McCarron CE, Pullenayegum EM, Thabane L, Goeree R, Tarride JE. The importance of adjusting for potential confounders in Bayesian hierarchical models synthesising evidence from randomised and non-randomised studies: an application comparing treatments for abdominalaortic aneurysms. *BMC Medical Research Methodology* 2010;**10**:64.

McKee 1999 *{published data only}*

McKee M, Britton A, Black N, McPherson K, Sanderson C, Bain C. Interpreting the evidence: choosing between randomised and non-randomised studies. *BMJ* 1999;**319**:312-5.

Moreira 2012 *{published data only}*

Moreira RF, Foltran FA, Albuquerque-Sendin F, Mancini MC, Coury HJCG. Comparison of randomized and non- randomized controlled trials evidence regarding the effectiveness of workplace exercise on musculoskeletal pain control. *Work* 2012;**41**(Suppl 1):4782-9.

Ni Chroinin 2013 *{published data only}*

Ni Chroinin D, Asplund K, Asberg S, Callaly E, Cuadrado-Godia E, Diez-Tejedor E, et al. Statin therapy and outcome after ischemic stroke: systematic review and meta-analysis of observational studies and randomized trials. *Stroke* 2013;**44**(2):448-56.

Nixdorf 2010 *{published data only}*

Nixdorf D, Moana-Filho E, Law A, McGuire LA, Hodges JS, John MT. Frequency of persistent tooth pain after root canal therapy: a systematic review and meta-analysis. *Journal of Endodotics* 2010;**36**(2):224-30.

Ottenbacker 1992 *{published data only}*

Ottenbacher K. Impact of random assignment on study outcome: an empirical examination. *Controlled Clinical Trials* 1992;**13**:50-61.

Papanastassiou 2012 *{published data only}*

Papanastassiou I, Phillips F, van Meirhaeghe J, Berenson J, Andersson G, Chung G, et al. Comparing effects of kyphoplasty, vertebroplasty, and non- surgical management in a systematic review of randomized and non-randomized controlled studies. *European Spine Journal* 2012;**21**:1826-43.

Phillips 1999 *{published data only}*

Phillips AN, Grabar S, Tassie JM, Costagliola D, Lundgren JD, Egger M. Use of observational databases to evaluate the effectiveness of antiretroviral therapy for HIV infection: comparison of cohort studies with randomized trials. *AIDS* 1999;**13**(15):2075-82.

Pratt 2012 *{published data only}*

Pratt N, Roughead E, Salter A, Ryan P. Choice of observational study design impacts on measurement of antipsychotic risks in the elderly: a systematic review. *BMC Medical Research Methodology* 2012;**12**(72):1-19.

Pyorala 1995 *{published data only}*

Pyorala S, Huttunen N, Uhari M. A review and meta-analysis of hormonal treatment of cryptorchidism. *Journal of Clinical Endocrinology and Metabolism* 1995;**80**(9):2795-9.

Schmoor 2008 *{published data only}*

Schmoor C, Caputo A, Schumacher M. Evidence from nonrandomized studies: a case study on the estimation of causal effects. *American Journal of Epidemiology* 2008;**167**(9):1120-9.

Scott 2007 *{published data only}*

Scott P, Kingsley G, Smith C, Choy EH, Scott DL. Non-steroidal anti-inflammatory drugs and myocardial infarctions: comparative systematic review of evidence from observational studies and randomised controlled trials. *Annals of the Rheumatic Diseases* 2007;**66**(10):1296-304.

Shah 2005 *{published data only}*

Shah B, Laupacis A, Hux J, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology* 2005;**58**(6):550-9.

Shepherd 2006 *{published data only}*

Shepherd J, Bagnall A, Colquitt J. 'Sometimes similar, sometimes different': a systematic review of meta-analyses of randomised and non-randomised policy intervention studies. 14th Cochrane Colloquium,. Dublin, Ireland, 23-26 October, 2006.

Steinberg 1994 *{published data only}*

Steinberg K, Smith J, Thacker S, Stroup DF. Breast cancer risk and duration of estrogen use: the role of study design in meta-analysis. *Epidemiology* 1994;**5**(4):415-21.

Stukel 2007 *{published data only}*

Stukel T, Fisher E, Wennberg D, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA* 2007;**297**(3):278-85.

Ward 1992 *{published data only}*

Ward LC, Fielding J, Dunn J, Kelly KA. The selection of cases for randomised trials:a registry survey of concurrent trial and non-trial patients. *British Journal of Cancer* 1992;**66**(5):943-50.

Watson 1994 *{published data only}*

Watson A, Vandekerckhove P, Lilford R, Vail A, Brosens I, Hughes E. A meta-analysis of the therapeutic role of oil soluble contrast media at hysterosalpingography: a surprising result?. *Fertility and Sterility* 1994;**61**(3):470-7.

Williams 1981 *{published data only}*

Williams PT, Fortmann SP, Farquhar JW, Varady A, Mellen S. A comparison of statistical methods for evaluating risk factor changes in community-based studies: an example from the

Stanford Three-Community Study. *Journal of Chronic Diseases* 1981;**34**(11):565-71.

Wilson 2001 *{published data only}*

Wilson D, Lipsey M. The role of method in treatment effectiveness research: evidence from meta-analysis. *Psychological Methods* 2001;**6**(4):413-29.

Additional references

Altman 2003

Altman D, Bland J. Interaction revisited: the difference between two estimates. *BMJ* 2003;**326**:219.

Dorsey 2010

Dorsey ER, de Roulet J, Thompson JP, Reminick JL, Thai A, White-Stellato Z, et al. Funding of US Biomedical Research, 2003-2008. *JAMA* 2010;**303**(2):137-43.

Higgins 2011

Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Chichester: John Wiley & Sons Ltd, 2011.

Institute of Medicine 2009

Institute of Medicine. Initial National Priorities for Comparitive Effectiveness Research. Institute of Medicine, Washington DC 2009.

Kamerow 2011

Kamerow D. PCORI: odd name, important job, potential trouble. *BMJ* 2011;**342**:d2635.

Kunz 1998

Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;**317**(7167):1185-90.

Kunz 2002

Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database of Systematic Reviews* 2002, Issue 4. [DOI: 10.1002/14651858.MR000012]

Lundh 2012

Lundh A, Sismondo S, Lexchin J, Busuioc OA, Bero L. Industry sponsorship and research outcome. *Cochrane Database of Systematic Reviews* 2012, Issue 12. [DOI: 10.1002/14651858.MR000033.pub2]

Montori 2004

Montori VM, Wilczynski NL, Morgan D, Haynes RB, Hedges Team. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. *BMJ* 2005 Jan 8;**330**(7482):68.

Odgaard-Jensen 2011

Odgaard-Jensen J, Timmer A, Kunz R, Akl EA, Schünemann H, Briel M, et al. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database of Systematic Reviews* 2011, Issue 4. [DOI: 10.1002/14651858.MR000012.pub3]

PCORI 2012

Patient Centered Outcomes Research Institute (PCORI). PCORI Methodology Standards. http://www.pcori.org/assets/PCORI-Methodology-Standards.pdf December 14, 2012.

Reeves 2013

Reeves B, Higgins J, Ramsay C, Shea B, Tugwall P, Wells G. An introduction to methodological issues when including non-randomised studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* 2013;**4**:1-11.

Sacks 1982

Sacks H, Chalmers T, Smith HJ. Randomized versus historical controls for clinical trials. *American Jourrnal of Medicine* 1982;**72**(2):233-40.

Sampson 2009

Sampson M, McGowan J, Cogo E, Grimshaw J, Moher D, Lefebvre C. An evidence-based practice guideline for the peer review of electronic search strategies. *Journal of Clinical Epidemiology* 2009;**62**(9):944-52.

* Indicates the major publication for the study

CHARACTERISTICS OF STUDIES

Characteristics of included studies *[ordered by study ID]*

Benson 2000

Methods	Searched for all RCTs and observational studies that compared 2 or more treatments between 1985 and 1998
Data	136 reports about 19 disparate treatments and interventions
Comparisons	Combined magnitude of effects from RCTs vs combined magnitude of effects from observational studies for same treatment
Outcomes	17 of 19 analyses yielded no difference in magnitude of effects comparing methods

Benson 2000 *(Continued)*

Notes	Little evidence that estimates of treatment effects in observational studies are larger than effects from RCTs

Risk of bias

Item	Authors' judgement	Description
Explicit criteria?	Yes	Had four inclusion criteria for observational studies matched to RCTs
Investigator Agreement?	No	No mention of this
Complete sample?	No	They could have missed observational studies due to poor indexing
Bias assessed?	No	Not done
Control for differences?	No	Methodological differences noted, but not controlled for
Heterogeneity addressed?	No	Noted, but not controlled for
Similar outcomes?	Yes	The few exceptions where outcomes were not similar were noted
No selective reporting?	Unclear	Not discussed in detail
Absence of evidence of bias from other sources?	Yes	

Beynon 2008

Methods	Searched for RCTs and observational studies with all-cause mortality as the outcome for a sample of topics randomly selected from the medical literature
Data	114 RCTs and 71 observational studies on19 diverse topics identified
Comparisons	Ratio of relative risks (RRR) calculated comparing RCT vs observational studies for each outcome
Outcomes	16 of 19 analyses yielded no difference in RRRs comparing methods
Notes	Little evidence that estimates of treatment effects in observational studies are larger than effects from RCTs

Risk of bias

Item	Authors' judgement	Description
Explicit criteria?	Yes	Identified by outcome, then observational studies were matched to an RCT
Investigator Agreement?	No	No mention of this
Complete sample?	No	Topics selected at random
Bias assessed?	No	Not done
Control for differences?	No	Mentioned selection bias of observational studies but did not control for this

Beynon 2008 *(Continued)*

Heterogeneity addressed?	Yes	Controlled for heterogeneity
Similar outcomes?	Yes	All mortality
No selective reporting?	Unclear	Not discussed in detail
Absence of evidence of bias from other sources?	Yes	

Bhandari 2004

Methods	An analysis of all studies, observational studies and RCTs, published between 1962 and 2002 which compared internal fixation and arthroplasty in femoral neck fracture patients
Data	27 studies eligible for inclusion:14 RCTs and 13 observational studies
Comparisons	Pooled data across studies for each outcome and calculated relative risks
Outcomes	Observational studies underestimated the relative benefit of arthroplasty by 19.5% (the risk reduction for revision surgery with arthroplasty compared with internal fixations was 77% for RCTs and 62% for NRS)
Notes	Observational studies provide results that are dissimilar to results provided by RCTs specifically for arthroplasty vs internal fixation for revision rates and mortality in femoral neck fracture patients

Risk of bias

Item	**Authors' judgement**	**Description**
Explicit criteria?	Yes	4 explicit criteria on focused topics
Investigator Agreement?	Yes	Two reviewed
Complete sample?	Yes	Complete sample on focused topic
Bias assessed?	Yes	Yes, table 1
Control for differences?	No	Discussed, but not controlled for
Heterogeneity addressed?	No	No mention
Similar outcomes?	Yes	Part of selection criteria
No selective reporting?	Yes	Thorough search included seeking unpublished studies
Absence of evidence of bias from other sources?	Yes	

Concato 2000

Methods	Identified all meta-analyses published between 1991 and 1995 in five major journals

Concato 2000 *(Continued)*

Data	72 RCTs and 24 observational studies were identified, in addition to 6 meta-analyses of both study method types, which covered 5 clinical topic areas. A total of 1,871,681 study participants were included in all analyses.
Comparisons	Pooled data across studies for each outcome and calculated relative risks
Outcomes	Effectiveness of Bacille Calmette-Guerin vaccine and TB (no difference between study design); Mammography and mortality (no difference); cholesterol levels and death due to trauma (no difference); treatment of hypertension and stroke (no difference between study design); treatment of hypertension and coronary heart disease (no difference)
Notes	No noted difference in point estimates between observational study results and RCT study results.

Risk of bias

Item	Authors' judgement	Description
Explicit criteria?	Unclear	Studies were identified from published meta-analyses in 5 journals
Investigator Agreement?	Yes	2 reviewed the MA for inclusion
Complete sample?	Unclear	Depended on how the MA was done
Bias assessed?	No	Stated it was assessed, but not reported or controlled for except in a few cases
Control for differences?	No	Discussed, but not controlled for
Heterogeneity addressed?	No	No mention
Similar outcomes?	Unclear	For some comparisons not clear what outcomes were measured
No selective reporting?	Unclear	Depends on the included MA
Absence of evidence of bias from other sources?	Yes	

Edwards 2012

Methods	RCTs of breast cancer treatment published between 2003-2008 were identified and similar observational studies of the same topics were also identified.
Data	37 studies selected (26 observational studies and 12 RCTs) for inclusion. A total of 32,969 study participants were included in all analyses.
Comparisons	Pooled data across studies for each outcome and calculated relative risks
Outcomes	Nerve dissection versus preservation on sensory deficit (no difference between study designs); axillary lymph node dissection vs sentinel lymph node biopsy on death (no difference between designs); axillary lymph node dissection vs sentinel lymph node biopsy on local recurrence (observational studies may have shown larger effect than RCTs); axillary lymph node dissection vs sentinel lymph node biopsy on numbness (no difference between designs); mastectomy vs breast conserving therapy on death (no difference between designs); mastectomy vs breast conserving therapy on local recurrence (no difference between designs); pectoral minor dissection vs preservation on number of lymph nodes removed (no difference between designs)

Edwards 2012 *(Continued)*

Notes	RCT and observational study results were inconsistently different (3 out of 10 comparisons were different in the authors' presented analyses).

Risk of bias

Item	Authors' judgement	Description
Explicit criteria?	Yes	All studies had to meet clear, specific, inclusion criteria
Investigator Agreement?	Yes	2 reviewers assessed titles for inclusion
Complete sample?	Unclear	The selective search may have introduced bias by not selecting all available literature
Bias assessed?	No	This was not assessed
Control for differences?	No	Discussed, but not controlled for
Heterogeneity addressed?	Yes	The authors calculated the heterogeneity within each meta-analysis.
Similar outcomes?	Yes	The analyses were stratified by topic type
No selective reporting?	Unclear	RCTs were selected from a 5 year window
Absence of evidence of bias from other sources?	Yes	

Furlan 2008

Methods	Found comparative studies of low back pain published before May 2005. Studies of similar interventions were synthesized
Data	17 observational studies and 8 RCTs identified which covered 3 outcomes of interest
Comparisons	Observational studies were synthesized and compared to the synthesized estimates from RCTs, producing ORs for each outcome
Outcomes	For all 3 outcomes covering comparing study design, observational studies underestimated the effects when compared to RCTs
Notes	Across all studies and outcomes, there is only slight evidence that observational study estimates are different than RCT estimates

Risk of bias

Item	Authors' judgement	Description
Explicit criteria?	Yes	Observational studies identified according to specific criteria then matched to RCTs
Investigator Agreement?	No	No mention
Complete sample?	No	Selected interventions with the most observational studies

Furlan 2008 *(Continued)*

Bias assessed?	Yes	Assessed RoB plus other characteristics
Control for differences?	Yes	Subgrouped
Heterogeneity addressed?	Yes	Sensitivity analysis
Similar outcomes?	Unclear	Grouped by intervention not outcome
No selective reporting?	Yes	Thorough search included seeking unpublished studies
Absence of evidence of bias from other sources?	Yes	

Golder 2011

Methods	Meta-analysis of meta-analyses comparing estimates of harm derived from meta-analysis of RCTs to meta-analyses of observational studies
Data	58 meta-analyses identified
Comparisons	Effect estimates of meta-analyses of RCTs compared to effect estimates of meta-analyses of observational studies. drug and non-drug studies included in comparisons.
Outcomes	Pooled relative measures of adverse effect (odds ratio or risk ratio)
Notes	No evidence, on average, in risk estimate of adverse effect of interventions from meta-analyses of RCTs vs observational studies

Risk of bias

Item	**Authors' judgement**	**Description**
Explicit criteria?	Unclear	Studies were identified from published meta-analyses in 5 journals
Investigator Agreement?	Yes	Consensus
Complete sample?	Unclear	Depended on how the MA was done
Bias assessed?	Yes	Described in text
Control for differences?	No	Done descriptively
Heterogeneity addressed?	No	Done descriptively
Similar outcomes?	No	Only one outcome had multiple studies addressing it
No selective reporting?	Unclear	Depends on the included MA
Absence of evidence of bias from other sources?	Yes	

Ioannidis 2001

Methods	Identified meta-analyses that considered both RCTs and observational studies published before 2000
Data	45 topics identified from 240 RCTs and 168 observational studies
Comparisons	Effect estimates of meta-analyses of RCTs compared to effect estimates of meta-analyses of observational studies.
Outcomes	Observational studies tended to show larger treatment effect sizes, and in 7 outcomes of 45 studied, differences between RCTs and observational studies were significantly different
Notes	Differences between RCTs and observational studies are present (about 16% of the time)

Risk of bias

Item	Authors' judgement	Description
Explicit criteria?	Yes	Very explicit for meta-analyses identified and studies within the meta-analyses
Investigator Agreement?	Unclear	Says "we" but not explicit
Complete sample?	No	Could have missed identifying some MA that contained both observational studies and RCTs
Bias assessed?	No	Assessed some study characteristics but not RoB specifically
Control for differences?	Yes	Subgrouped
Heterogeneity addressed?	Yes	Subgrouped
Similar outcomes?	Yes	Grouped by outcomes
No selective reporting?	Yes	Did identify extent of trials that had been published after the included meta-analysis
Absence of evidence of bias from other sources?	Yes	

Kuss 2011

Methods	Performed a systematic review and meta-analysis that compared RCTs and propensity score analyses in similar populations
Data	10 topics identified from 51 RCTs and 28 observational studies that employed propensity scores
Comparisons	Effect estimates of meta-analyses of RCTs compared to effect estimates of meta-analyses of propensity score analyses
Outcomes	Propensity score analyses across all outcomes were no different than estimates from RCTs
Notes	Only a small bias, if any, may remain in propensity score analyses estimating the effects of off-pump versus on-pump surgery

Risk of bias

Kuss 2011 *(Continued)*

Item	Authors' judgement	Description
Explicit criteria?	Yes	The authors included all studies with propensity score analyses comparing off and on pump CABG
Investigator Agreement?	Yes	Two reviewers selected studies independently
Complete sample?	Unclear	It is possible that RCTs that were not previously identified in systematic reviews may have been missed
Bias assessed?	No	Bias not assessed
Control for differences?	Yes	Confounder data were extensively collected
Heterogeneity addressed?	No	Heterogeneity not addressed
Similar outcomes?	Unclear	All analyses were evaluating similar comparisons for disparate outcomes
No selective reporting?	Unclear	Their search was simple and used only MEDLINE for RCTs
Absence of evidence of bias from other sources?	Yes	

Lonjon 2013

Methods	Performed a systematic review and meta-analysis that compared RCTs and prospective observational studies using propensity scores addressing the same clinical questions
Data	31 clinical topics identified from 94 RCTs and 70 observational studies that employed propensity scores
Comparisons	Effect estimates of meta-analyses of RCTs compared to effect estimates of meta-analyses of propensity score analyses
Outcomes	Propensity score analyses across all outcomes were no different than estimates from RCTs
Notes	Prospective observational studies are reliable for providing evidence in the absence of RCTs

Risk of bias

Item	Authors' judgement	Description
Explicit criteria?	Unclear	31 different clinical questions were included, though it is unclear if these questions were conceived a priori
Investigator Agreement?	No	One reviewer extracted data and one reviewer selected studies based on clinical expertise
Complete sample?	No	Not all RCTs were selected for each research question--restricted to last 5 years
Bias assessed?	Yes	Performance, detection, and attrition biases were all assessed
Control for differences?	Yes	Sensitivity analyses performed
Heterogeneity addressed?	Yes	For all analyses, heterogeneity assessed using I^2 statistic

Lonjon 2013 *(Continued)*

Similar outcomes?	Yes	The authors' primary outcome was all-cause mortality
No selective reporting?	Unclear	As a result of not including all RCTs, selective reporting is possible
Absence of evidence of bias from other sources?	Yes	

Müeller 2010

Methods	Identified studies, including RCTs and observational studies that compared laparoscopic vs open cholecystectomy
Data	162 studies were identified, including 136 observational studies and 26 RCTs, covering 15 outcomes of interest
Comparisons	Effect estimates of RCTs were compared to estimates from observational studies
Outcomes	In 3 of 15 outcomes there were significant differences between results from observational studies and RCTs
Notes	Differences between RCTs and observational studies are present (about 20% of the time)

Risk of bias

Item	**Authors' judgement**	**Description**
Explicit criteria?	Yes	Identified RCTs and observational studies (cohorts) on a specific topic
Investigator Agreement?	No	No mention of this
Complete sample?	Yes	Complete sample on focused topic
Bias assessed?	Yes	Cochrane RoB criteria plus additional
Control for differences?	Yes	Sensitivity analysis
Heterogeneity addressed?	Yes	Sensitivity analysis
Similar outcomes?	Yes	Included studies with different outcomes, analyzed by outcome
No selective reporting?	Unclear	Their search was simplistic (NEDLINE)
Absence of evidence of bias from other sources?	Yes	

Naudet 2011

Methods	Identified published and unpublished studies from 1989 to 2009 that examined fluoxetine and venlafaxine as first line treatment for major depressive disorder
Data	12 observational studies and 109 RCTs were identified

Naudet 2011 *(Continued)*

Comparisons	Meta-regression estimates for outcomes of interest
Outcomes	The standardized treatment response in RCTs is greater by a magnitude of 4.59 compared to observational studies
Notes	Response to antidepressants is greater in RCTs than in observational studies

Risk of bias

Item	Authors' judgement	Description
Explicit criteria?	Yes	PICO specified
Investigator Agreement?	Yes	2 reviewed independently, consensus
Complete sample?	Yes	Searched for all studies on a specific topic, seems thorough
Bias assessed?	Yes	Different instruments for RCTs and observational studies
Control for differences?	Yes	Some RoB items included in meta-regression, also did sensitivity analysis
Heterogeneity addressed?	Yes	Meta-regression
Similar outcomes?	No	Converted to standardized scores
No selective reporting?	Yes	Limited evidence of publication bias based on funnel plots
Absence of evidence of bias from other sources?	Yes	

Oliver 2010

Methods	Identify systematic reviews that compareD results of policy interventions, stratifying estimates by observational study and RCT study design published between 1999 and 2004
Data	16 systematic reviews identified, with a median of 11.5 RCTs and 14.5 observational studies in each systematic review
Comparisons	Observational studies published in systematic reviews were pooled separately from RCTs published in the same systematic reviews.
Outcomes	Results stratified by study design were heterogeneous with no clear direction of magnitude
Notes	Overall, the authors found no evidence for clear systematic differences in terms of results between RCTs and observational studies.

Risk of bias

Item	Authors' judgement	Description
Explicit criteria?	Yes	Identified systematic reviews including observational studies and RCTs on a specific topic
Investigator Agreement?	Yes	All disagreements were settled by consensus or referral to third reviewer

Oliver 2010 *(Continued)*

Complete sample?	Yes	Searched for all studies on a specific topic,
Bias assessed?	Yes	Bias was discussed in detail
Control for differences?	Yes	Sensitivity analyses were detailed in the results
Heterogeneity addressed?	Yes	Heterogeneity was discussed in detail
Similar outcomes?	Yes	Various outcomes from policy interventions analyzed by intervention type
No selective reporting?	Unclear	Not discussed in detail
Absence of evidence of bias from other sources?	Yes	

Papanikolauo 2006

Methods	The authors compareD evidence from RCTs to observational studies that have explored the effects of interventions on the risk of harm. Harms of interest were identified from RCTs with more than 4000 patients. Observational studies of more than 4000 patients were also included for comparison
Data	15 harms of interest were identified and relative risks were extracted for 13 topics
Comparisons	Data from 25 observational studies were compared to results from RCTs. Relative risks for each outcome/harm were calculated for both study types
Outcomes	The estimated increase in RR differed by more than two-fold between observational studies and RCTs for 54% of the topics studied.
Notes	Observational studies usually under-estimated the absolute risk of harms.

Risk of bias

Item	Authors' judgement	Description
Explicit criteria?	Yes	Matched observational studies to published RCTs on particular topics
Investigator Agreement?	Yes	2 independently, consensus
Complete sample?	Unclear	Unclear whether they were able to match observational studies to all the RCTs
Bias assessed?	No	Not done
Control for differences?	No	Not done
Heterogeneity addressed?	Unclear	Did assess mathematical heterogeneity between reviews of RCT and observational studies
Similar outcomes?	Unclear	"Harms" broadly defined, could include multiple outcomes
No selective reporting?	No	Selection of observational studies could have missed some
Absence of evidence of bias from other sources?	Yes	

Shikata 2006

Methods	The authors identified all meta-analyses of RCTs and observational studies of digestive surgery published between 1966 and 2004.
Data	52 outcomes for 18 disparate topics were identified from 276 articles (96 RCTs and 180 observational studies)
Comparisons	Pooled odds ratios and relative risks were extracted for each outcome, using the same indicator that had been used in the meta-analysis of interest
Outcomes	Approximately 1/4 of all outcomes of interest yielded different results between observational studies and RCTs
Notes	

Risk of bias

Item	Authors' judgement	Description
Explicit criteria?	Unclear	MA were identified, if meta-analysis did not include observational studies, then searched for them separately
Investigator Agreement?	Yes	2 reviewed independently, then consensus
Complete sample?	Yes	Complete sample on focused topic
Bias assessed?	No	Not done
Control for differences?	No	Not done
Heterogeneity addressed?	No	Not done
Similar outcomes?	Yes	Grouped by outcomes, noted that measures were similar
No selective reporting?	Unclear	Search strategy comprehensive but odd (MA + OBS)
Absence of evidence of bias from other sources?	Yes	

CABG: **c**oronary artery bypass graft
NRS: non-randomized study
PICO: population, intervention, comparison and outcome
RCT: randomized controlled trial
RoB: risk of bias

Characteristics of excluded studies *[ordered by study ID]*

Study	Reason for exclusion
Ather 2011	An original meta-analysis with an incidental comparison of RCTs and observational studies.
Begg 1991	This is a statistical methods paper that did not have a systematic selection of studies for identified outcomes or interventions.

Study	Reason for exclusion
Beyersmann 2008	This is a statistical methods paper that did not have a systematic selection of studies for identified outcomes or interventions.
Bosco 2010	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Britton 1998	The authors chose to include uncontrolled trials in their data collection.
Chambers 2010	This is a methods paper that did not have a systematic selection of studies for identified outcomes or interventions. There was no meta-analysis of observational data performed.
Coulam 1994	From this study it was not possible to separate out uncontrolled, quasi-, or pseudo-randomized studies from other studies.
Dahabreh 2012	Not a comprehensive or systematic search of RCT data. RCT data matched selectively to observational data.
Deeks 2002	This study was unique in that it created non-randomised studies through resampling of RCTs. This is a statistical methods paper that did not have a systematic selection of studies for identified outcomes or interventions.
Deeks 2003	The authors included quasi-experimental and quasi-randomized studies.
Diehl 1986	Not designed to specifically compare the effect sizes of RCT and observational studies.
Diez 2010	Not designed to specifically compare the effect sizes of RCT and observational studies, but to test new analytic methods that takes study design into account
Flossmann 2007	An original meta-analysis with an incidental comparison of RCTs and observational studies.
Hallstrom 2000	An original meta-analysis with an incidental comparison of RCTs and observational studies.
Henry 2001	Not designed to specifically compare the effect sizes of RCT and observational studies, but to qualitatively assess agreement between designs.
Hlatky 1988	Did not have a systematic selection of studies for identified outcomes or interventions.
Ioannidis 2005	This is a qualitative comparison of high cited RCTs and observational studies and their initially stronger effects that are often later contradicted.
Labrarere 2006	This is a methods paper that did not have a systematic selection of studies for identified outcomes or interventions.
LaTorre 2009	An original meta-analysis of harms outcomes among only observational studies.
Linde 2007	An incidental comparison of RCTs and observational studies; did not have a systematic selection of studies for identified outcomes or interventions.
Lipsey 1993	From this study it was not possible to separate out uncontrolled, quasi-, or pseudo-randomized studies from other studies.
Loke 2011	An original meta-analysis with an incidental comparison of RCTs and observational studies.
MacLehose 2000	The authors included quasi-experimental studies.
Mak 2009	An original meta-analysis with an incidental comparison of RCTs and observational studies.

Study	Reason for exclusion
McCarron 2010	This is a statistical methods paper that did not have a systematic selection of studies for identified outcomes or interventions; the authors re-analyzed previously published data.
McKee 1999	A commentary and/or descriptive analysis.
Moreira 2012	No meta-analysis; RCT data included quasi-experimental.
Ni Chroinin 2013	An original meta-analysis with an incidental comparison of RCTs and observational studies.
Nixdorf 2010	An original meta-analysis with an incidental comparison of RCTs and observational studies.
Ottenbacker 1992	A commentary and/or descriptive analysis.
Papanastassiou 2012	An original meta-analysis with an incidental comparison of RCTs and observational studies.
Phillips 1999	This study had no systematic selection of meta-analyses; only included three large prospective studies that were the focus of the analysis.
Pratt 2012	No meta-analysis performed.
Pyorala 1995	An original meta-analysis with an incidental comparison of RCTs and observational studies.
Schmoor 2008	This study had no systematic selection of meta-analyses; only an embedded prospective study within an RCT that was the focus of the analysis.
Scott 2007	An original meta-analysis with an incidental comparison of RCTs and observational studies.
Shah 2005	No meta-analysis, only a quantitative comparison of results between observational studies with different designs.
Shepherd 2006	A commentary and/or descriptive analysis.
Steinberg 1994	An analysis of previously published meta-analyses that aimed to compare effects between sources of controls within observational study designs.
Stukel 2007	A primary analysis; this is a statistical methods paper that did not have a systematic selection of studies for identified outcomes or interventions; no RCT data.
Ward 1992	This is a statistical methods paper that did not have a systematic selection of studies for identified outcomes or interventions; not a review of meta-analyses.
Watson 1994	An original meta-analysis with an incidental comparison of RCTs and observational studies; the authors include non-randomized as observational studies.
Williams 1981	This is a statistical methods paper that did not have a systematic selection of studies for identified outcomes or interventions; not a review of meta-analyses and no meta-analysis performed.
Wilson 2001	From this study it was not possible to separate out uncontrolled, quasi-, or pseudo-randomized studies from other studies.

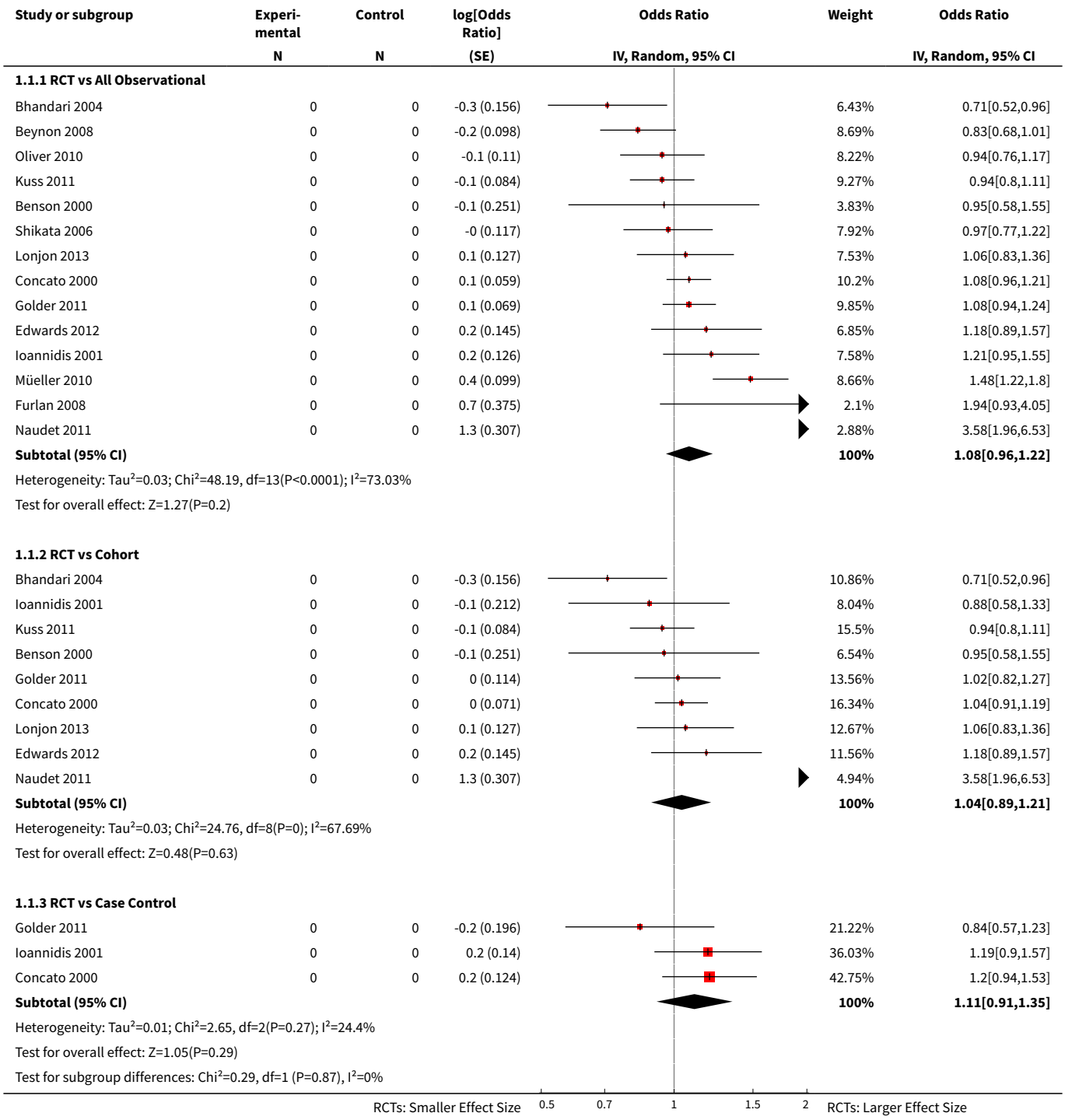
RCT: randomized controlled trial

DATA AND ANALYSES

Comparison 1. RCT vs Observational

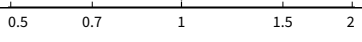
Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Summary Ratios of Ratios: RCTs vs Observational Studies	14		Odds Ratio (Random, 95% CI)	Subtotals only
1.1 RCT vs All Observational	14		Odds Ratio (Random, 95% CI)	1.08 [0.96, 1.22]
1.2 RCT vs Cohort	9		Odds Ratio (Random, 95% CI)	1.04 [0.89, 1.21]
1.3 RCT vs Case Control	3		Odds Ratio (Random, 95% CI)	1.11 [0.91, 1.35]
2 Summary Ratios of Ratios: RCTs vs Observational Studies (Heterogeneity Subgroups)	14		Odds Ratio (Random, 95% CI)	Subtotals only
2.1 Low Heterogeniety (I^2: 0% to 30%)	4		Odds Ratio (Random, 95% CI)	1.00 [0.72, 1.39]
2.2 Moderate Heterogeneity (I^2:31% to 60%)	8		Odds Ratio (Random, 95% CI)	1.11 [0.95, 1.30]
2.3 Significant Heterogeneity (I^2: 61% to 100%)	2		Odds Ratio (Random, 95% CI)	1.08 [0.87, 1.34]
3 Summary Ratios of Ratios: RCTs vs Observational Studies (Pharmacological Studies vs non-Pharmacological Studies)	13		Odds Ratio (Random, 95% CI)	Subtotals only
3.1 Pharmacological Studies	6		Odds Ratio (Random, 95% CI)	1.17 [0.95, 1.43]
3.2 Non-Pharmacological Studies	11		Odds Ratio (Random, 95% CI)	1.03 [0.87, 1.21]
4 Summary Ratios of Ratios: RCTs vs Observational Studies (Propensity Scores)	14		Odds Ratio (Random, 95% CI)	Subtotals only
4.1 RCTs vs Observational Studies (propensity score adjustment)	2		Odds Ratio (Random, 95% CI)	0.98 [0.85, 1.12]
4.2 RCTs vs Observational Studies (no propensity score adjustment)	12		Odds Ratio (Random, 95% CI)	1.10 [0.96, 1.27]

Analysis 1.1. Comparison 1 RCT vs Observational, Outcome 1 Summary Ratios of Ratios: RCTs vs Observational Studies.



Analysis 1.2. Comparison 1 RCT vs Observational, Outcome 2 Summary Ratios of Ratios: RCTs vs Observational Studies (Heterogeneity Subgroups).

Study or subgroup	Experimental N	Control N	log[Odds Ratio] (SE)	Odds Ratio IV, Random, 95% CI	Weight	Odds Ratio IV, Random, 95% CI
1.2.1 Low Heterogeniety (I2: 0% to 30%)						
Bhandari 2004	0	0	-0.3 (0.156)		24.53%	0.71[0.52,0.96]
Kuss 2011	0	0	-0.1 (0.084)		28.98%	0.94[0.8,1.11]
Benson 2000	0	0	-0.1 (0.251)		18.32%	0.95[0.58,1.55]
Müeller 2010	0	0	0.4 (0.099)		28.16%	1.48[1.22,1.8]
Subtotal (95% CI)					**100%**	**1[0.72,1.39]**
Heterogeneity: Tau²=0.09; Chi²=20.11, df=3(P=0); I²=85.08%						
Test for overall effect: Z=0.01(P=1)						
1.2.2 Moderate Heterogeneity (I2:31% to 60%)						
Beynon 2008	0	0	-0.2 (0.098)		15.46%	0.83[0.68,1.01]
Oliver 2010	0	0	-0.1 (0.11)		14.59%	0.94[0.76,1.17]
Lonjon 2013	0	0	0.1 (0.127)		13.35%	1.06[0.83,1.36]
Concato 2000	0	0	0.1 (0.059)		18.23%	1.08[0.96,1.21]
Golder 2011	0	0	0.1 (0.069)		17.58%	1.08[0.94,1.24]
Edwards 2012	0	0	0.2 (0.145)		12.1%	1.18[0.89,1.57]
Furlan 2008	0	0	0.7 (0.375)		3.66%	1.94[0.93,4.05]
Naudet 2011	0	0	1.3 (0.307)		5.02%	3.58[1.96,6.53]
Subtotal (95% CI)					**100%**	**1.11[0.95,1.3]**
Heterogeneity: Tau²=0.03; Chi²=26.39, df=7(P=0); I²=73.48%						
Test for overall effect: Z=1.34(P=0.18)						
1.2.3 Significant Heterogeneity (I2: 61% to 100%)						
Shikata 2006	0	0	-0 (0.117)		52.12%	0.97[0.77,1.22]
Ioannidis 2001	0	0	0.2 (0.126)		47.88%	1.21[0.95,1.55]
Subtotal (95% CI)					**100%**	**1.08[0.87,1.34]**
Heterogeneity: Tau²=0.01; Chi²=1.65, df=1(P=0.2); I²=39.34%						
Test for overall effect: Z=0.68(P=0.49)						
Test for subgroup differences: Chi²=0.34, df=1 (P=0.84), I²=0%						

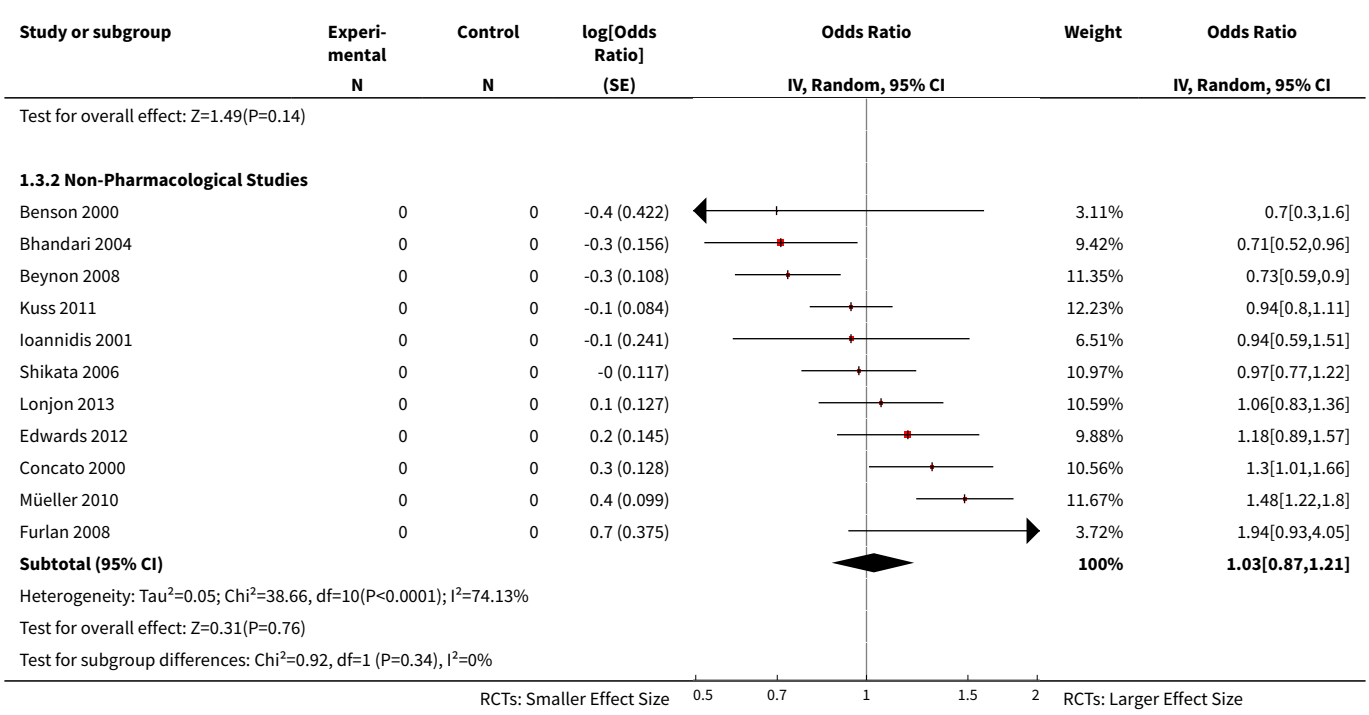


RCTs: Smaller Effect Size — RCTs: Larger Effect Size

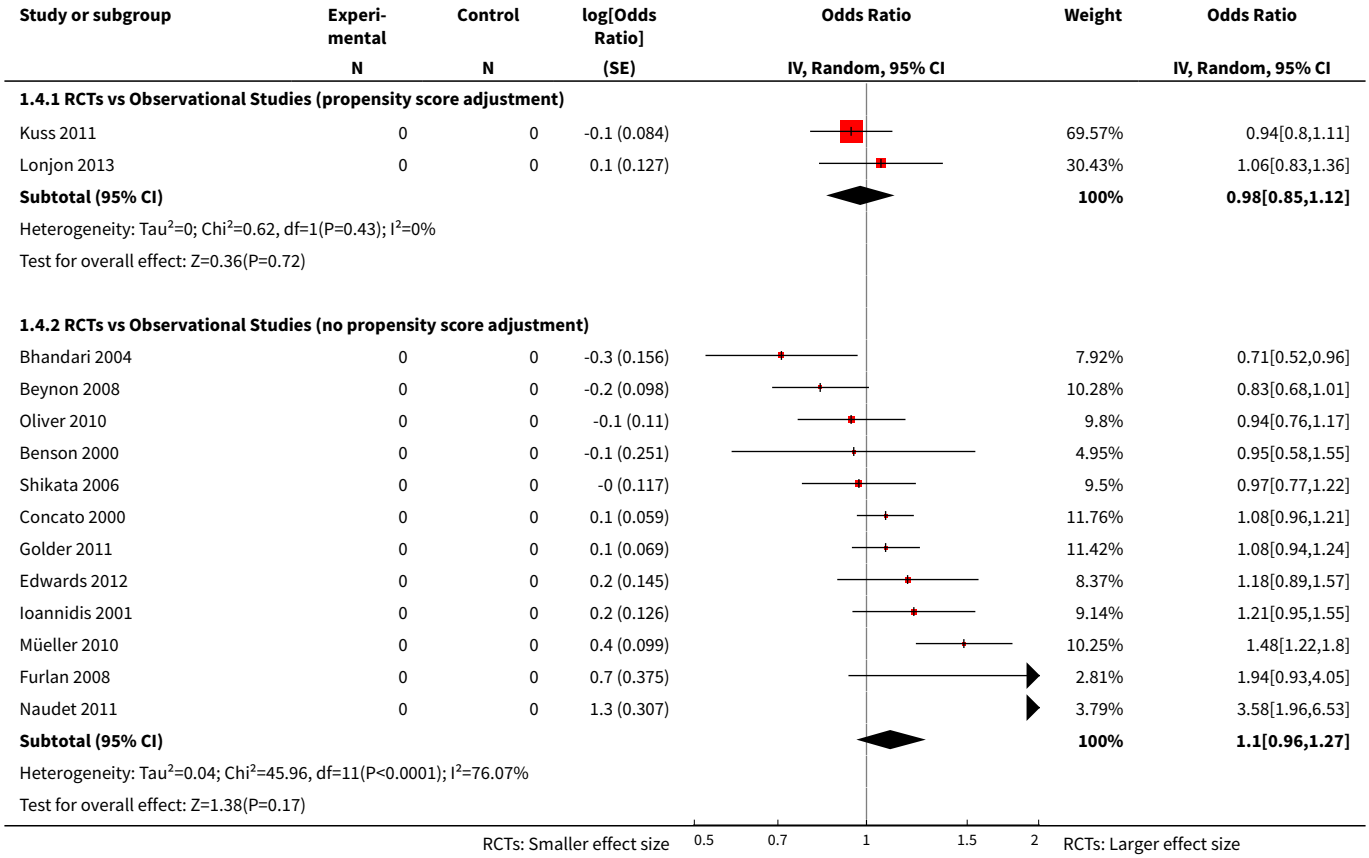
Analysis 1.3. Comparison 1 RCT vs Observational, Outcome 3 Summary Ratios of Ratios: RCTs vs Observational Studies (Pharmacological Studies vs non-Pharmacological Studies).

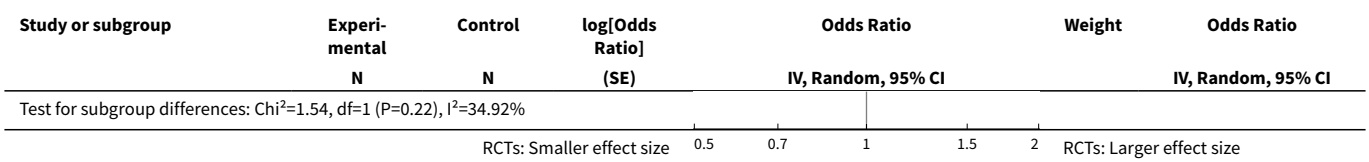
Study or subgroup	Experimental N	Control N	log[Odds Ratio] (SE)	Odds Ratio IV, Random, 95% CI	Weight	Odds Ratio IV, Random, 95% CI
1.3.1 Pharmacological Studies						
Beynon 2008	0	0	-0.2 (0.095)		20.95%	0.83[0.69,1]
Concato 2000	0	0	0 (0.05)		24.04%	1.04[0.94,1.15]
Golder 2011	0	0	0.1 (0.069)		22.87%	1.08[0.94,1.24]
Benson 2000	0	0	0.1 (0.315)		7.55%	1.12[0.61,2.08]
Ioannidis 2001	0	0	0.3 (0.148)		16.76%	1.41[1.06,1.88]
Naudet 2011	0	0	1.3 (0.307)		7.83%	3.58[1.96,6.53]
Subtotal (95% CI)					**100%**	**1.17[0.95,1.43]**
Heterogeneity: Tau²=0.04; Chi²=26.32, df=5(P<0.0001); I²=81%						

RCTs: Smaller Effect Size — RCTs: Larger Effect Size



Analysis 1.4. Comparison 1 RCT vs Observational, Outcome 4 Summary Ratios of Ratios: RCTs vs Observational Studies (Propensity Scores).





APPENDICES

Appendix 1. PubMed strategy, which was modified as appropriate for use in the other databases

Search	Terms
#4	(((#1) AND #2) AND #3)
#3	compara*[tiab] OR comparison*[tiab] OR contrast*[tiab] OR similar*[tiab] OR consistent*[tiab] OR inconsistent*[tiab] OR dissimilar*[tiab] OR differen*[tiab] OR concordan*[tiab] OR discordan*[tiab] OR heterogene*[tiab] OR "Research Design"[mh]
#2	"Observation"[mh] OR "Cohort Studies"[mh] OR "Longitudinal Studies"[mh] OR "Retrospective Studies"[mh] OR "Prospective Studies"[mh] OR observational[tiab] OR cohort*[tiab] OR crosssectional[tiab] OR crossectional[tiab] OR cross-sectional[tiab] OR cross sectional[tiab] OR longitudinal[tiab] OR causal inference*[tw] OR causality[tw] OR “instrumental variable”[tw] OR “structural model”[tw] OR practice-based[tw] OR propensity score*[tw] OR natural experiment*[tw] OR case-control[tw] OR before-after[tw] OR pre-post[tw] OR case-cohort[tw] OR case-crossover[tw] OR serial[tiab] OR nonexperimental[tiab] OR non-experimental[tiab] OR “nonrandomized”[tiab] OR “non-randomised”[tiab] OR “non-randomised”[tiab] OR “nonrandomised”[tiab] OR “study designs”[tiab] OR “newcastle ottawa”[tiab] OR overestimat*[tiab] OR over-estimat*[tiab] OR bias[tiab] OR "are needed"[tiab] OR (evidence[tiab] AND quality[tiab])
#1	Cochrane Database Syst Rev [TA] OR search[tiab] OR meta-analysis[PT] OR MEDLINE[tiab] OR PubMed[tiab] OR (systematic*[tiab] AND review*[tiab]) OR review[ti]

CONTRIBUTIONS OF AUTHORS

All authors contributed to drafting of the review. LB conceived the idea for the study. THH conducted all searches and reviewed the final manuscript. LB and AA screened titles, wrote the final manuscript, and revised the manuscript in response to peer review comments. AA conducted all analyses.

DECLARATIONS OF INTEREST

None to declare.

SOURCES OF SUPPORT

Internal sources

- Clinical and Translational Sciences Institute (CTSI), University of California, San Francisco (UCSF), USA.

External sources

- No sources of support supplied

DIFFERENCES BETWEEN PROTOCOL AND REVIEW

We were unable to conduct subgroup analyses by topic area of the research, or differences in interventions and conditions, as proposed, because these parameters were too diverse to permit grouping of studies. For the same reasons, we were unable to explore the impact of confounding by indication.

INDEX TERMS

Medical Subject Headings (MeSH)

*Observational Studies as Topic; *Randomized Controlled Trials as Topic; Meta-Analysis as Topic; Outcome Assessment, Health Care [*methods]

MeSH check words

Humans