Cochrane Library

Cochrane Database of Systematic Reviews

Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials: a meta-epidemiological study (Review)

Toews I, Anglemyer A, Nyirenda JLZ, Alsaid D, Balduzzi S, Grummich K, Schwingshackl L, Bero L

Toews I, Anglemyer A, Nyirenda JLZ, Alsaid D, Balduzzi S, Grummich K, Schwingshackl L, Bero L.
Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials: a meta-epidemiological study.
Cochrane Database of Systematic Reviews 2024, Issue 1. Art. No.: MR000034.
DOI: 10.1002/14651858.MR000034.pub3.

www.cochranelibrary.com

WILEY

TABLE OF CONTENTS

ABSTRACT 1
PLAIN LANGUAGE SUMMARY 2
SUMMARY OF FINDINGS 4
BACKGROUND 5
OBJECTIVES 6
METHODS 7
RESULTS 11
Figure 1. 12
Figure 2. 14
Figure 3. 15
Figure 4. 18
Figure 5. 20
Figure 6. 26
DISCUSSION 27
AUTHORS' CONCLUSIONS 29
ACKNOWLEDGEMENTS 30
REFERENCES 31
CHARACTERISTICS OF STUDIES 41
APPENDICES 124
WHAT'S NEW 130
HISTORY 130
CONTRIBUTIONS OF AUTHORS 130
DECLARATIONS OF INTEREST 131
SOURCES OF SUPPORT 131
DIFFERENCES BETWEEN PROTOCOL AND REVIEW 131
INDEX TERMS 132

[Methodology Review]

Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials: a meta-epidemiological study

Ingrid Toews[1], Andrew Anglemyer[2], John LZ Nyirenda[1], Dima Alsaid[1], Sara Balduzzi[3], Kathrin Grummich[1], Lukas Schwingshackl[1], Lisa Bero[4]

[1]Institute for Evidence in Medicine (for Cochrane Germany Foundation), Medical Center – University of Freiburg, Freiburg, Germany. [2]Department of Preventive and Social Medicine, Dunedin School of Medicine, University of Otago, Dunedin, New Zealand. [3]Biometrics Department, Antoni van Leeuwenhoek - Netherlands Cancer Institute, Amsterdam, Netherlands. [4]Charles Perkins Centre and School of Pharmacy, Faculty of Medicine and Health, The University of Sydney, Camperdown, Sydney, Australia

Contact: Ingrid Toews, ingrid.toews@uniklinik-freiburg.de.

Editorial group: Cochrane Methodology Review Group.
Publication status and date: New search for studies and content updated (no change to conclusions), published in Issue 1, 2024.

Citation: Toews I, Anglemyer A, Nyirenda JLZ, Alsaid D, Balduzzi S, Grummich K, Schwingshackl L, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials: a meta-epidemiological study. *Cochrane Database of Systematic Reviews* 2024, Issue 1. Art. No.: MR000034. DOI: 10.1002/14651858.MR000034.pub3.

ABSTRACT

Background

Researchers and decision-makers often use evidence from randomised controlled trials (RCTs) to determine the efficacy or effectiveness of a treatment or intervention. Studies with observational designs are often used to measure the effectiveness of an intervention in 'real world' scenarios. Numerous study designs and their modifications (including both randomised and observational designs) 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. An up-to-date systematic analysis is needed to identify differences in effect estimates from RCTs and observational studies. This updated review summarises the results of methodological reviews that compared the effect estimates of observational studies with RCTs from evidence syntheses that addressed the same health research question.

Objectives

To assess and compare synthesised effect estimates by study type, contrasting RCTs with observational studies.

To explore factors that might explain differences in synthesised effect estimates from RCTs versus observational studies (e.g. heterogeneity, type of observational study design, type of intervention, and use of propensity score adjustment).

To identify gaps in the existing research comparing effect estimates across different study types.

Search methods

We searched MEDLINE, the Cochrane Database of Systematic Reviews, Web of Science databases, and Epistemonikos to May 2022. We checked references, conducted citation searches, and contacted review authors to identify additional reviews.

Selection criteria

We included systematic methodological reviews that compared quantitative effect estimates measuring the efficacy or effectiveness of interventions tested in RCTs versus in observational studies. The included reviews compared RCTs to observational studies (including retrospective and prospective cohort, case-control and cross-sectional designs). Reviews were not eligible if they compared RCTs with studies that had used some form of concurrent allocation.

Data collection and analysis

Using results from observational studies as the reference group, we examined the relative summary effect estimates (risk ratios (RRs), odds ratios (ORs), hazard ratios (HRs), mean differences (MDs), and standardised mean differences (SMDs)) to evaluate whether there was a relatively larger or smaller effect in the ratio of odds ratios (ROR) or ratio of risk ratios (RRR), ratio of hazard ratios (RHR), and difference in (standardised) mean differences (D(S)MD).

If an included review did not provide an estimate comparing results from RCTs with observational studies, we generated one by pooling the estimates for observational studies and RCTs, respectively. Across all reviews, we synthesised these ratios to produce a pooled ratio of ratios comparing effect estimates from RCTs with those from observational studies. In overviews of reviews, we estimated the ROR or RRR for each overview using observational studies as the reference category.

We appraised the risk of bias in the included reviews (using nine criteria in total). To receive an overall low risk of bias rating, an included review needed: explicit criteria for study selection, a complete sample of studies, and to have controlled for study methodological differences and study heterogeneity. We assessed reviews/overviews not meeting these four criteria as having an overall high risk of bias.

We assessed the certainty of the evidence, consisting of multiple evidence syntheses, with the GRADE approach.

Main results

We included 39 systematic reviews and eight overviews of reviews, for a total of 47. Thirty-four of these contributed data to our primary analysis. Based on the available data, we found that the reviews/overviews included 2869 RCTs involving 3,882,115 participants, and 3924 observational studies with 19,499,970 participants.

We rated 11 reviews/overviews as having an overall low risk of bias, and 36 as having an unclear or high risk of bias. Our main concerns with the included reviews/overviews were that some did not assess the quality of their included studies, and some failed to account appropriately for differences between study designs – for example, they conducted aggregate analyses of all observational studies rather than separate analyses of cohort and case-control studies.

When pooling RORs and RRRs, the ratio of ratios indicated no difference or a very small difference between the effect estimates from RCTs versus from observational studies (ratio of ratios 1.08, 95% confidence interval (CI) 1.01 to 1.15). We rated the certainty of the evidence as low. Twenty-three of 34 reviews reported effect estimates of RCTs and observational studies that were on average in agreement.

In a number of subgroup analyses, small differences in the effect estimates were detected:

- pharmaceutical interventions only (ratio of ratios 1.12, 95% CI 1.04 to 1.21);
- RCTs and observational studies with substantial or high heterogeneity; that is, $I^2 \geq 50\%$ (ratio of ratios 1.11, 95% CI 1.04 to 1.18);
- no use (ratio of ratios 1.07, 95% CI 1.03 to 1.11) or unclear use (ratio of ratios 1.13, 95% CI 1.03 to 1.25) of propensity score adjustment in observational studies; and
- observational studies without further specification of the study design (ratio of ratios 1.06, 95% CI 0.96 to 1.18).

We detected no clear difference in other subgroup analyses.

Authors' conclusions

We found no difference or a very small difference between effect estimates from RCTs and observational studies. These findings are largely consistent with findings from recently published research. Factors other than study design need to be considered when exploring reasons for a lack of agreement between results of RCTs and observational studies, such as differences in the population, intervention, comparator, and outcomes investigated in the respective 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 is needed of how these factors might yield estimates reflective of true effectiveness.

PLAIN LANGUAGE SUMMARY

How similar are estimates of treatment effectiveness derived from randomised controlled trials and observational studies?

Key messages

- On average, the effect estimates of randomised controlled trials (RCTs) and observational studies differ only very slightly. Effect estimates are statistical constructs that describe the size of an intervention effect in terms of the difference between the outcomes of two groups of people in a clinical trial or study.
- We need more research with careful consideration of factors that might impact on the similarities and differences in effect estimates between different study types.

What are RCTs and observational studies, and why do their effect estimates potentially differ?

Randomised controlled trials (RCTs) are a type of healthcare experiment where participants are allocated at random to one of two (or more) treatment groups. One group is given an experimental treatment (also known as an 'intervention'); the other is the 'control' group, which is not given the intervention. RCTs test how effective and safe an experimental treatment is under ideal conditions.

Observational studies try to measure the effectiveness of an intervention in non-experimental, 'real world' scenarios. Case-control (or retrospective) studies and cohort studies are two common types of observational study. Case-control studies compare a group of people with a particular condition/disease to a group who do not have it but are otherwise similar. Cohort studies follow a group of people with a common characteristic over time to find out how many reach a certain health outcome of interest.

Sometimes, the results of RCTs and observational studies addressing the same question may have different results. These types of study differ in how they are conducted and their susceptibility to systematic error.

What did we want to find out?

We wanted to assess the impact of study type (RCT versus observational studies) on the summary effect estimate and to explore methodological aspects that might explain any differences.

What did we do?

We searched databases for reviews that systematically compared the effect estimates reported in RCTs and observational studies that addressed the same health research question. We looked for reviews that included any healthcare outcomes, without restrictions on the language of publication. We searched for reviews/overviews published between 01 January 1990 and 12 May 2022. We then compared the results of the reviews, and summarised the evidence. We rated our confidence in this evidence, based on factors such as the methods used in the reviews and their size, and the consistency of findings across reviews.

What did we find?

We identified 47 relevant reviews; 34 contributed data to our main analysis. The reviews compared the effect estimates of RCTs to those of cohort studies, case-control studies, or both. The reviews addressed a variety of health-related topics. They were conducted in countries around the world, but most were done in the USA. Twelve reviews did not report any information on funding. In 8 reviews, the authors reported receiving no funding. In 23 reviews, the authors reported receiving public funding, such as governmental funding or funding from universities or foundations. Two reviews were funded by the European Union and two reviews reported receiving industry funding. Most funded reviews reported multiple sources of funding.

Main results

- We found that the effect estimates of RCTs and observational studies may differ very little to not at all.
- There may be small differences when we compare effect estimates of studies investigating only medicines (as opposed to other healthcare treatments, such as surgery or physical therapy).

We also found little difference in the effect estimates that were based on data from:
- meta-analysis of RCTs and observational studies that showed substantial statistical heterogeneity; that is, variability in the intervention effects being evaluated in the different studies;
- observational studies that either did not use or were unclear about how they used methods to account for population characteristics that can have an impact on the effectiveness of an intervention (propensity score adjustment);
- observational studies that did not give sufficient information about their study design.

What are the limitations of the evidence?

We have little confidence in the evidence because the included reviews might be at risk for systematic errors because of how they were conducted. Moreover, the reviews were about different types of people and interventions, meaning that the individual findings amongst the reviews differed considerably.

How up to date is this review?

The evidence is current to May 2022.

SUMMARY OF FINDINGS

Summary of findings 1. Summary of findings

Patient or population: systematic reviews and overviews of systematic reviews

Setting: any setting

Intervention: effect estimates as reported by randomised controlled trials

Comparison: effect estimates as reported by observational studies

Outcomes	Anticipated absolute effects (95% CI)*		Relative effect (95% CI)[a]	No. of reviews	Certainty in the evidence (GRADE)	Comments
	Risk with observational studies	Risk with RCTs				
Difference between effect estimates from RCTs and observational studies [for RCTs vs OBSs]	Not estimable		**1.08 (1.01 to 1.15)**	34 reviews	⨁⨁◯◯ Low[b]	The effect estimate indicates that, on average, there might be little difference between the effect estimates obtained from RCTs and observational studies.

CI: confidence interval; **RCT:** randomised controlled trial

[a]The relative effect reports on a ratio of ratios comparing the effect estimates of RCTs versus observational studies.
[b]Rated down by two levels: one level for serious risk of bias, and one level for serious inconsistency with unexplained statistical heterogeneity of $I^2 = 69\%$.

BACKGROUND

Individuals, healthcare professionals, organisations, institutions, and healthcare decision-makers need trustworthy, reliable, accurate, specific, and up-to-date information about the effectiveness of healthcare interventions. This has been clearly demonstrated during the COVID-19 pandemic, where the need for clinical and public health evidence has been key to decision-making. Different study types are used to evaluate intervention effectiveness and inform healthcare decision-making. Evidence from randomised controlled trials (RCTs) is often used to determine the efficacy or effectiveness of a treatment or intervention under ideal or very strictly defined conditions. Studies of observational design are used to evaluate the effectiveness of an intervention in non-experimental, 'real world' scenarios at the population level. Both study types have advantages and disadvantages in connection to what questions they might be able to answer with a high degree of certainty, their ease and timeliness of planning and conducting, and their cost. For example, due to their design-specific features and the administrative regulations, observational studies are often considered easier and cheaper to plan and conduct than RCTs but the findings they produce are usually less accurate than those from RCTs. This creates a tension for decision-makers, who need accurate evidence comparing the effectiveness of different interventions in a timely manner.

Comparative effectiveness research is “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 comparative effectiveness research is to assist consumers, clinicians, purchasers, and policymakers to make informed decisions that will improve healthcare at both the individual and population levels" (Institute of Medicine 2009). Comparative effectiveness research has also been called "comparative clinical effectiveness research" and "patient-centred 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. Theoretically, RCTs and observational study designs are both suitable to produce information on the clinical effectiveness of interventions.

The debate about the validity of observational studies and RCTs for estimating the effectiveness of interventions has continued for decades and is ongoing (Reeves 2013). In order to understand the use of findings from the two study types for decision-making, systematic reviews that compare the effect sizes or biases in RCTs and observational studies are undertaken (see, for example, Allain 2017; Ankarfeldt 2017; Hoshino 2021a; Hoshino 2021b; Jainaud 2021; Kimachi 2021; Mathes 2021; Schwingshackl 2021; Bröckelmann 2022). Some reviews focus on the comparison of studies with specific design features or in specific health-related research fields (Hemkens 2016; Schwingshackl 2021). Other reviews investigate the association of specific design features, such as randomisation or inadequate randomisation, with selection bias and exaggerated treatment effects (Deeks 2003; Ioannidis 2001; Odgaard-Jensen 2011). However, there is still considerable uncertainty and gaps in knowledge about the validity of observational studies versus RCTs for estimating the effectiveness of interventions. Consequently, an up-to-date systematic analysis of study design features, risk of bias, and effect size for all types of studies used for comparative effectiveness research is needed to identify specific differences in design types and potential biases.

Description of the methods being investigated

According to the principles of clinical epidemiology, the effectiveness of an intervention should be investigated with rigorous methods, and RCTs are best suited to do so because of their capacity to account for unknown confounding factors through randomisation. Evidence from observational studies is considered to be at a higher risk of bias, mainly due to unmeasured confounding and poor adjustment for known confounding factors. As noted in the *Cochrane Handbook for Systematic Reviews of Interventions*, potential biases for all non-randomised study types are likely to be greater than for RCTs (Higgins 2021). Nonetheless, when RCTs are not available or possible, or their quality and size are suboptimal, observational studies may be used to give an indication of the effectiveness of an intervention (Cuello-Garcia 2022). In addition, healthcare decision-makers and healthcare funders might increasingly be compelled to rely on observational study results because that type of research (and its effect estimates) are faster, cheaper, easier, and more convenient to obtain (Bundesamt für Justiz 2020). And, as discussed further below, the methods for observational studies are evolving to better control for confounding.

Numerous study types and modifications of existing types, both RCTs and observational studies, are used for comparative effectiveness research. In RCTs, rigorous methods are employed to ensure an unbiased investigation of intervention effectiveness. Randomisation is a key element, where participants are allocated randomly to one of two or more intervention and control groups. This reduces selection bias. Blinding is employed to prevent bias in outcome assessment, with participants, researchers, and outcome assessors often unaware of group assignments. Standardisation of procedures and protocols across intervention and control groups helps maintain consistency in how the participants are treated. Additionally, intention-to-treat analysis is commonly used, including all randomised participants in the analysis regardless of adherence to the assigned intervention, to preserve the benefits of randomisation and enhance the validity of study findings. While RCTs are widely regarded as the gold standard in clinical research, critics argue that the strict inclusion criteria and controlled clinical environment may limit the generalisability of findings to diverse patient populations or real-world settings. Some contend that the emphasis on internal validity may come at the expense of external validity, potentially reducing the applicability of trial results in broader clinical practice. Additionally, ethical concerns may arise regarding the random assignment of participants to different interventions, especially when effective treatments are known, raising questions about equipoise. Despite these criticisms, RCTs remain indispensable for establishing causal relationships between interventions and outcomes, providing essential evidence for evidence-based medicine.

Observational study designs in health research typically comprise prospective and retrospective cohort studies as well as case-control studies. In prospective and retrospective cohort studies within health research, comprehensive methods are used to investigate the effect of interventions or exposures on outcomes over time. Participants, initially free from the outcome of interest, are categorised based on intervention/exposure status, and their

subsequent health outcomes are tracked, recorded, and analysed. However, critics highlight the susceptibility to confounding factors, emphasising the challenge of accounting for all potential confounders that may influence the observed effect. Additionally, loss to follow-up over extended periods can pose challenges to the validity of results. Despite these critiques, cohort studies provide valuable insights into long-term health effects and risk factors, contributing significantly to epidemiological evidence. In cohort studies, propensity score approaches, including matching on the propensity score, stratification on the propensity score, inverse probability of treatment weighting using the propensity score, and covariate adjustment using the propensity score, were shown to be useful for estimating causal effects (Desai 2019). Such approaches increase the reliability of findings generated by cohort studies and might subsequently form an increasingly solid basis for decision-making in healthcare.

In case-control studies within health research, researchers employ a retrospective approach to compare individuals with a specific outcome (cases) to those without the outcome (controls). Participants are selected based on their outcome status, and the exposure history is then assessed retrospectively. Matching or statistical adjustment is often utilised to control for potential confounding variables, enhancing the comparability between cases and controls. The selection of an appropriate control group is crucial to ensure the validity of the study findings. Data analysis typically involves calculating odds ratios to estimate the strength of association between exposures and outcomes. Despite their susceptibility to recall bias and the challenge of establishing temporal relationships, case-control studies are valuable for investigating rare diseases or outcomes and identifying potential risk factors.

There is meta-research that compares effect estimates from RCTs with those from observational studies for the same or similar PICO (Population; Intervention; Comparison; Outcome) question. One of the main aims of such meta-research is to investigate similarities and differences in effect estimates, as well as factors that might be associated with such similarities and differences. Such meta-research usually employs the methods of systematic reviews. These include conducting a systematic search for RCTs and observational studies addressing the same or similar PICO questions with clear eligibility criteria. After study selection, a comparison of the design-specific effect estimates per outcome of interest and a statistical evaluation of the effect estimates are conducted. As a result, a ratio of effect estimates from the different study types (ratio of ratios) is presented. Some meta-researchers also use visual inspections of the design-specific effect estimates and describe the proportions of effect estimates from RCTs and observational studies that: (a) point in similar directions; (b) lie in opposite directions; or (c) have (non-)overlapping 95% confidence intervals (CIs). Comparisons of effect estimates from different study types are also addressed in overviews of reviews. The approach is largely similar to that of systematic reviews in the same research field, except that the unit of analysis is not the individual studies but rather the meta-analyses of the constitutive individual studies. Such meta-research feeds into discussions about what study findings are needed and used for health-related decision-making as well as decision-making about priorisation in health research planning.

How these methods might work

In order to better understand how and to what extent evidence from RCTs and observational studies can be integrated and used in decision-making, this review investigates similarities and differences between effect estimates of RCTs and observational studies. To do this, we analyse methodological reviews that have been designed to compare effect estimates from RCTs and observational studies. Such reviews usually calculate ratios of effect estimates, such as ratios of risk ratios (RRRs), ratios of odds ratios (RORs) or ratio of hazard ratios (RHRs). Besides the formal statistical comparison of effect estimates, review authors may also conduct a so-called 'eyeball test' or visual inspection of meta-analyses (Moneer 2022). In such tests and inspections, proportions of effect estimates and 95% CIs from RCTs and observational studies that lie in similar directions, overlap, or are in opposite directions or do not overlap are quantified. With multiple approaches available to examine differences and similarities between different study types, there is no gold standard in this field of research.

Why it is important to do this review

This is an update of a Cochrane review first published in 2014 (Anglemyer 2014), which included 15 methodological reviews. Anglemyer and colleagues found that, on average, the effect estimates of observational studies and RCTs do not differ. They argued that a clearer understanding of factors impacting on effect estimates by study type is needed.

When conducting the first version of this review (Anglemyer 2014), there were no systematic reviews of comparisons of all study designs currently being used for comparative effectiveness research. Previously, reviews that compared RCTs with observational studies most often limited the comparison to cohort studies, or did not specify the types of observational designs included. In addition, the methodology for observational studies has continuously evolved. Since 2014, more methodological reviews comparing the effect estimates from RCTs and observational studies have been published. We have included those reviews which were eligible in this updated review. The findings of this review help prioritise the types of context-specific study designs that should be used to minimise bias and generate reliable evidence on intervention effectiveness. This review expands our knowledge about how evidence from RCTs and observational studies can be integrated.

OBJECTIVES

To assess and compare synthesised effect estimates by study type, contrasting RCTs with observational studies.

To explore factors that might explain differences in synthesised effect estimates from RCTs versus observational studies (e.g. heterogeneity, type of observational study design, type of intervention, and use of propensity score adjustment).

To identify gaps in the existing research comparing effect estimates across different study types.

METHODS

Criteria for considering studies for this review

Types of studies

We included two types of reviews:

- methodological systematic reviews that included RCTs and observational studies in which the review's main objective was to compare effect estimates by study type;
- methodological overviews of systematic reviews that investigated differences in effect estimates by study type in systematic reviews including RCTs and observational studies.

Specifically, we included designated comparisons of effect estimates from RCTs with any type of observational study.

We examined only systematic reviews or overviews of systematic reviews that were designed as methodological reviews. We defined a methodological review as a review designed and conducted with the aim of comparing effect estimates of studies that vary by a particular methodological factor (in this case, study design). In other words, included reviews/overviews could not aim solely to compare the clinical effectiveness of an intervention to no intervention or a comparator.

Within the eligible methodological overviews and systematic reviews, the relevant randomised study types were limited to studies described as head-to-head RCTs, cluster-RCTs, adaptive designs, practice/pragmatic/explanatory trials, PBE-CPI "practice-based evidence for clinical practice improvement," and natural experiments.

We limited comparisons within the reviews to those looking at quantitative effect estimates measuring the effectiveness or safety of interventions in RCTs with those in observational studies. Our focus was on reviews of the effectiveness or safety of health-related interventions. We did not exclude reviews/overviews based on their publication status.

In addition, for the purpose of this review and for better distinctions between the study types compared in this review, we differentiated the main characteristics of RCTs and observational studies by distinguishing investigator-initiated treatment from non-investigator-initiated treatment. In more detail, in RCTs, the decision about the experimental intervention and the control intervention is essentially determined by the researchers or investigators whilst formulating a research question. Accordingly, a limited choice of interventions is available to treat the study participants in the trial. The study participants are included based on their suitability for study participation and testing the efficacy, effectiveness, and safety of the intervention as compared to the control intervention (or no intervention). In cohort studies, on the other hand, the choice of intervention is essentially based on the individual needs of each patient and the assessment of their healthcare professionals. The study participants are included in observational studies on the basis of their treatment with a certain intervention. The study investigators of cohort studies are not in control of the intervention, but take on an observational role. Case-control studies typically include participants who already have a certain outcome (e.g. cardiovascular disease) with a treatment, intervention, or lifestyle already completed or ongoing. Thus, the data collection of the study takes place after the intervention or exposure and the study investigators have no influence on the intervention.

For this review, the only non-experimental studies we analysed 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 (Reeves 2013). For this review, we note that relevant observational study types could be described as prospective and retrospective cohort studies, case-control studies, observational or cross-sectional studies of registries and databases, including electronic medical records. Furthermore, there are observational studies employing so-called causal inference techniques (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 (Morshed 2009). Sometimes, the terms "observational study" and "non-randomised study" are used interchangeably.

We excluded comparisons of study types 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 also excluded reviews that:

- compared RCTs to non-randomised trials where the focus was on the design feature of randomisation and its impact on effect estimates. For example, we excluded reviews designed to compare studies with random allocation to those with non-random allocation or trials with adequate versus inadequate or unclear concealment of allocation;
- compared the results of meta-analyses with the results of single trials or single observational studies;
- compared previously published meta-analyses versus meta-analyses that were conducted by the review authors themselves;
- were unsystematic in building their study sample. For example, we excluded one review that was included in the previous version of this review because it searched for RCTs and observational studies by identifying these studies from a limited set of published systematic reviews (Concato 2000b). This search strategy might have missed relevant studies that could have been retrieved through a direct search for RCTs and observational studies in relevant databases;
- performed meta-analyses with both RCTs and observational studies with an incidental comparison of the results.

Types of data

We included systematic reviews or overviews of reviews that quantitatively compared the effect estimates of RCTs and observational studies with regard to the efficacy or effectiveness of alternative interventions to prevent or treat a clinical condition or to improve the delivery of care. Specifically, we included in the statistical analysis systematic reviews or overviews of reviews that reported pooled effect estimates from RCTs and observational studies or the ratio of their effect estimates (ratio of ratios).

Differences in effect 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 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 explained any differences in effect size estimates.

Types of methods

We included reviews comparing effect estimates between: (1) RCTs and observational studies; and (2) RCTs and different types of observational studies. Eligible comparisons included:

- RCTs/cluster-RCTs versus prospective/retrospective cohort studies;
- RCTs/cluster-RCTs versus case-control studies;
- RCTs/cluster-RCTs versus cross-sectional studies;
- RCTs/cluster-RCTs versus other observational design;
- RCTs/cluster-RCTs versus observational studies employing so-called causal inference analytical methods.

Types of outcome measures

We included all types of dichotomous and continuous outcome measures of effects, as reported in the methodological systematic reviews or overviews of systematic reviews (e.g. (ratio of) odds ratios (ORs), (ratio of) relative risks (or risk ratios (RRs), risk differences (RDs), (ratio of) hazard ratios (HRs), (differences in) mean differences (MDs), and standardised mean differences (SMDs)) and their accompanying measures of uncertainty. Where possible, we used pooled ratios of odds ratios (RORs) and ratios of risk ratios (RRRs) as the outcome measure.

Primary outcomes

Our primary outcome is the difference between effect estimates of RCTs and observational studies, measured as RORs or RRRs.

The RORs and RRRs measure how far the effect estimates differ between the two study types. The greater the difference between the effect estimates from RCTs and observational studies, the greater the ROR or RRR is above or below 1.0. The ROR or RRR does not give an indication of the direction of the effect; that is, whether it is a harmful or beneficial effect.

Secondary outcomes

We did not consider any secondary outcomes.

Search methods for identification of studies

We used systematic, sensitive searches in electronic databases and other sources to identify a comprehensive sample of reviews and overviews for our review. The searches were developed in consideration of the search strategy of the previously published version of this review (Anglemyer 2014).

Electronic searches

For this review update, an Information Specialist (KG) updated and modified the search strategy for an adequate balance of sensitivity and specificity. To identify relevant methodological reviews, we searched the electronic databases from 01 January 1990 to 12 May 2022. We limited the search date to this time span because methods for conducting health research are continuously evolving and improving. So, we wanted to capture an up-to-date overview of effect estimates as reported by relatively recent reviews and overviews.

We searched the following electronic databases:

- Cochrane Database of Systematic Reviews, in the Cochrane Library (Issue 4 of 12, April 2021), searched 21 April 2021;
- MEDLINE (Ovid) (R) ALL (1946 to 11 May 2022), searched 21 April 2021 and 12 May 2022;
- Epistemonikos (www.epistemonikos.org/) (2004 to 8 May 2021), searched 8 May 2021;
- Web of Science Core Collection (Science Citation Index Expanded (SCI-EXPANDED), Emerging Sources Citation Index (ESCI)), via Clarivate (1945 to 2021), searched 12 May 2021.

We selected these databases because of their comprehensive coverage of health-related literature and focus on systematic reviews. For our last update, we searched MEDLINE only because this was the most effective search approach. All but four records were retrieved by searches in MEDLINE.

The search strategy consisted of MeSH terms and relevant keywords. See Appendix 1 and Appendix 2 for all search strategies. We did not use search filters for methodological studies because they are unavailable (Neilson 2019).

The search strategy for MEDLINE and Epistemonikos was constructed using an AND combination with a search string, including search terms for methodological studies and meta-research, which were not used in the Web of Science and Cochrane Library searches. By developing and testing the search strategy, the AND combination with this search string in Web of Science decreased the number by more than 50%. Therefore, we decided to just combine terms for non-randomised and randomised studies with terms for assessing differences in effect estimates. The same applied to the search in the Cochrane Library.

Searching other resources

Reference lists

We conducted a formal citation screening up to 1 June 2022 for all included reviews/overviews, including backward and forward citation tracking (for the full list of included reviews/overviews, see Appendix 2). For this purpose, all references and all citing articles as noted in the Web of Science (all databases were searched) were retrieved and processed. All languages were included. Furthermore, we conducted forward citation tracking via Google Scholar on 5 May 2021, identifying literature that cited the original version of the review (Anglemyer 2014).

A closer review of the reviews/overviews included in the Anglemyer 2014 review revealed that published reviews that compared effect estimates of RCTs and observational studies referred to other published reviews with the same aim. The backward citation tracking in the previous review identified five additional relevant reviews. Our citation tracking identified three additional relevant records (Borkowska 2018; Tan 2017; Yanik 2013).

Correspondence

We contacted researchers known to be working in this field to ask about further relevant reviews.

Data collection and analysis

We conducted this review according to Cochrane standards (Higgins 2021), and considered the guidelines for reporting meta-epidemiological methodology research for the reporting of our research (Murad 2017). We also considered the methods outlined in the previous version of this review (Anglemyer 2014). We describe the differences between this update, the previous version of this review (Anglemyer 2014), and the protocol in the Differences between protocol and review section.

Selection of studies

After removing duplicate references, records were uploaded to Covidence for further processing (Covidence 2022). Four review authors (IT, JN, DAS, AA) independently selected potentially relevant records by scanning the titles, abstracts, and descriptor terms, and applying the inclusion criteria and a decision tree. We discarded irrelevant reports. We obtained the full-text articles of all potentially relevant or uncertain reports. The review authors (IT, JN, DAS, AA, LS) independently applied the inclusion criteria. We screened records against the eligibility criteria, including relevance based on study design, types of methods employed, and a comparison of effects based on different methodologies or designs. LB adjudicated any disagreements that could not be resolved through discussion.

Data extraction and management

For this review update, two review authors, working independently, extracted data with the help of the software RedCAP and MS Excel. We piloted the data extraction tool on five records, and used issues that emerged during piloting to revise the tool.

Two review authors (of IT, JN, DAS, AA, LS) independently double-coded and entered information from each selected review into standardised data extraction forms. We resolved any conflicts by re-reading the full publications and through discussion amongst the authors. Extracted information included the following.

- **Review details***:* citation, author conflicts of interest and study sponsorship source, start and end dates of search, methods of data retrieval and analysis, databases searched, location, eligibility criteria (inclusion and exclusion), study types compared, number of included studies and participants, methods and results of risk of bias assessment.
- **Comparison of methods details:** number of included studies and participants in each analysis, intervention and comparator analysed, effect estimates and statistical heterogeneity (I^2) from each study type within each publication, information on adjustment in observational studies. If no quantitative data were reported on effect estimates, we extracted narrative descriptions of differences and similarities of overall effect estimates by study type.

We did not retrieve any reviews that needed translation in order to extract data.

Assessment of risk of bias in included studies

We included systematic reviews of studies and overviews of systematic reviews. Therefore, the Cochrane tool for assessing the risk of bias for individual studies did 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 independently rated as 'yes', 'no', or 'unclear' by the authors (IT, JN, DAS, AA, LS). We resolved any disagreements by re-reading the full publication and through discussion.

In connection to investigating selective reporting, we did not specifically include dissemination bias. Dissemination bias was not assessed in this review.

We rated reviews that we coded as adequate (i.e. rated as 'yes') in four domains – (1) explicit criteria for study selection, (2) complete sample of studies, (3) controlled for methodological differences, and (4) controlled for heterogeneity – as having a low risk of bias overall. For reviews that were inadequate (i.e. rated as 'no' or 'unclear' in one or more of these domains, we assessed them as having a high risk of bias overall.

Measures of the effect of the methods

In general, outcome measures included relative risks or risk ratios (RR), odds ratios (OR), hazard ratios (HRs), standardised mean differences (SMDs), and mean differences (MDs). We also extracted ratios of odds ratios (RORs) and ratios of risk ratios (RRRs) from reviews that had calculated these.

Unit of analysis issues

We considered each included systematic review or overview of systematic reviews as the unit of analysis. Still, a systematic review can present results of different meta-analyses (that is, on different outcomes), and an overview of systematic reviews can present results of different systematic reviews.

When available, we extracted the overall summary measure (e.g. ROR) presented by the authors of each included systematic review or overview of systematic reviews. If this was not available, we conducted meta-analyses of the presented results of meta-analyses (in the case of systematic reviews) or meta-analyses of the presented results of systematic reviews (in the case of overviews of systematic reviews), aiming to obtain one summary measure per each included systematic review or overview of systematic reviews.

Dealing with missing data

This review is a secondary data analysis and did not incur the missing data issues seen in most systematic reviews. We contacted the review/overview authors if we considered that there may be data missing from their published work.

Assessment of heterogeneity

We synthesised data from multiple reviews to compare effects from RCTs with observational studies. We had a wide variety of outcomes and interventions synthesised, increasing the amount of heterogeneity between reviews. We assessed heterogeneity using the τ^2 statistic, X^2 statistic with a significance level of 0.10, and the I^2 statistic (Higgins 2005; Riley 2011). To estimate τ^2, we used the restricted maximum-likelihood method (Viechtbauer 2005). Together with the magnitude and direction of the effect, we interpreted an I^2 estimate between 30% and 50% as indicating moderate heterogeneity, 51% to 80% substantial heterogeneity, and 81% to 100% as a high level of heterogeneity. Furthermore, if an included review had already assessed the heterogeneity of its included studies, we reported the authors' original assessment of heterogeneity.

Assessment of reporting biases

We attempted to minimise the potential for publication bias through our comprehensive search strategy, which 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 compared effect estimates by study type, i.e. RCTs and observational studies. We analysed dichotomous and continuous outcomes separately.

By using observational studies as the reference group, we examined the estimates to see whether there was a relatively larger or smaller estimate coming from RCTs (estimate in RCTs > estimate in observational studies or estimate in RCTs < estimate in observational studies, respectively).

For this purpose, for binary outcomes, we obtained the ROR (calculated as: OR in the RCTs / OR in the observational studies) or the RRR (calculated as: RR in the RCTs / RR in the observational studies) (Altman 2003). We distinguished between "favourable" and "unfavourable" outcomes. Favourable outcomes are those one might want to observe (e.g. pain relief, improved function) and unfavourable outcomes are those one might not want to observe (e.g. mortality, wound infection). Considering the different direction of an estimate coming from either an RCT or an observational study, and therefore the different interpretation of a ROR (or RRR), we transformed favourable outcomes into unfavourable outcomes, by reversing the estimate provided, in order to have the possibility of pooling the outcomes together.

As explained in Unit of analysis issues, we aimed to obtain one summary measure per each included systematic review or overview of systematic reviews. In case a systematic review provided results on different outcomes (considering that the same primary studies might be included more than once in the same review because they provide information on different outcomes) or an overview provided results on different systematic reviews, we conducted multilevel meta-analyses, with either the systematic review or the overview of systematic reviews as the grouping variable. Multilevel meta-analyses were conducted through the R package mixmeta (Sera 2019). Once we obtained one summary measure, either RRR or ROR, per each included review or overview of reviews, we combined them in a random-effects meta-analysis using the inverse-variance method, reporting an overall effect estimate that pools both RRRs and RORs as ratio of ratios.

To avoid overlap of data between included reviews and overviews, we excluded reviews that were included in other included overviews. Additionally, we excluded those reviews stating that they used an inversion rule from the main analysis (Sterne 2018), but included them in a sensitivity analysis (see Sensitivity analysis).

We decided against pooling the comparisons for continuous outcomes, due to the diversity in the nature of the outcome; instead, we reported the results narratively.

In addition, we assessed in how many cases the effect estimates of RCTs were different from the effect estimates of observational studies beyond chance, and in how many cases the effect estimates of the two study types pointed in opposite directions.

Visual inspection of effect estimates

Besides the meta-analysis, we also visually inspected the effect estimates of RCTs and observational studies to examine the number and proportion of effect estimates that: (1) had overlapping 95% CIs and non-overlapping 95% CIs; (2) were in concordant or discordant direction; and (3) showed a statistically significant or non-significant effect estimate. We analysed these data with descriptive statistics and reported proportions in accordance with the 'Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline' (Campbell 2019).

Subgroup analysis and investigation of heterogeneity

Reducing bias in comparative effectiveness research is particularly important for studies comparing pharmacological interventions, given their implications for clinical care and healthcare purchasing. Thus, we planned, a priori, to conduct a subgroup analysis of reviews/overviews that compared and analysed studies that had performed pharmacological comparisons.

Additionally, we performed a subgroup analysis of the included methodological reviews according to the level of heterogeneity, subgrouping those with substantial and high heterogeneity (as measured in their respective meta-analyses) versus those with moderate or low heterogeneity.

Whenever possible, a priori planned subgroup analyses were conducted between:

- observational studies in which propensity score adjustment was used versus those in which propensity score adjustment was not used, to account for newer methods of study planning and analysis;
- RCTs versus cohort studies, to account for variances by study design;
- RCTs versus case-control studies, to account for variances by study design;
- overviews of reviews versus systematic reviews, to account for differences in the methodological design and included evidence in the data contributing to our analyses.

Sensitivity analysis

We added reviews that used a selective inversion approach, or “coining” of the intervention and control group, to the primary

statistical meta-analysis in a sensitivity analysis in order to test for the effect of their inclusion on the overall effect estimate.

We added reviews whose effect estimates were deemed favourable (e.g. survival), and were therefore transformed for the main meta-analysis, in a sensitivity analysis. In this analysis, we used their originally reported effect estimates as a basis for the analysis in order to test whether the transformation of favourable outcomes had an effect on the overall effect estimate of our analysis.

Assessment of the certainty of the evidence

We assessed the certainty of the evidence for the principal comparison with the GRADE approach and reported this assessment in the summary of findings table (Schünemann 2020). Evidence certainty can be graded as very low, low, moderate, or high. In GRADE, the default rating for a body of evidence from randomised trials is high certainty, whilst for observational study designs it is low certainty, reflecting the potential bias induced by the lack of randomisation (i.e. confounding and selection bias) (Schünemann 2020). Given that this review focuses not on randomised and non-randomised studies *directly*, but rather on the reviews and overviews comparing evidence from RCTs and observational studies, we decided to set the default rating for our principal comparison at high certainty (see Summary of findings 1). GRADE methodology for assessing overviews of systematic reviews of the effects of interventions is currently under development; we anticipate using it in future updates of this review.

RESULTS

Description of studies

See Characteristics of included studies, Characteristics of excluded studies, and Characteristics of ongoing studies.

Results of the search

Our update searches yielded 6979 unique references. We discarded 6880 references as they were clearly not relevant to the review. We selected 99 records for further review. Of these, we excluded 60 articles as they did not meet our inclusion criteria. We included 35 new reviews. Combined with 12 reviews from the Anglemyer 2014 version of this review, a total of 47 reviews (reported in 49 articles) met our inclusion criteria for this review. We also identified four ongoing studies in this review update (CRD42014013478; CRD42017059665; CRD42017079569; CRD42018062204). See Figure 1 for the study selection flow chart.

Figure 1. PRISMA 2020 flow diagram for updated systematic reviews which included searches of databases, registers and other sources

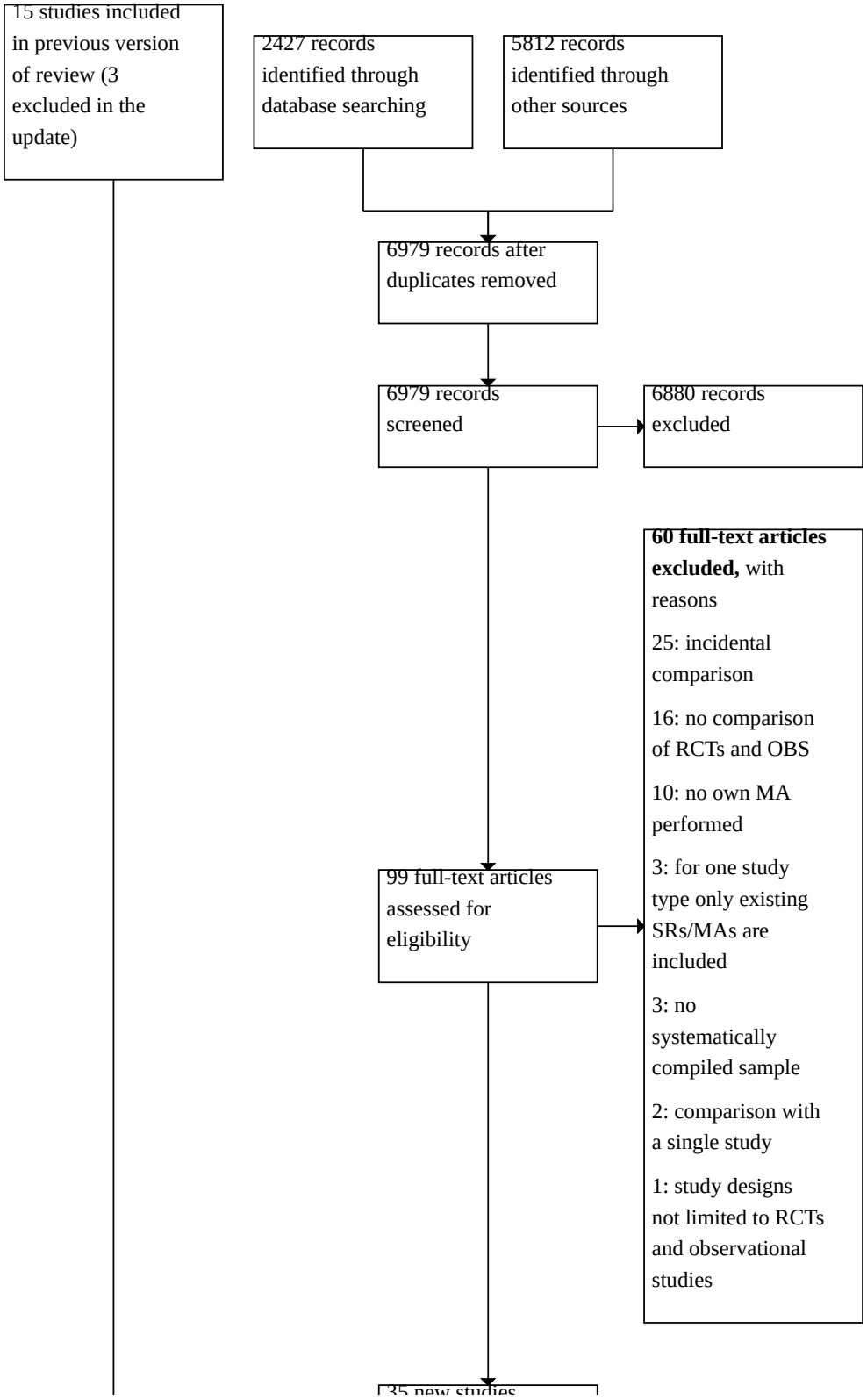
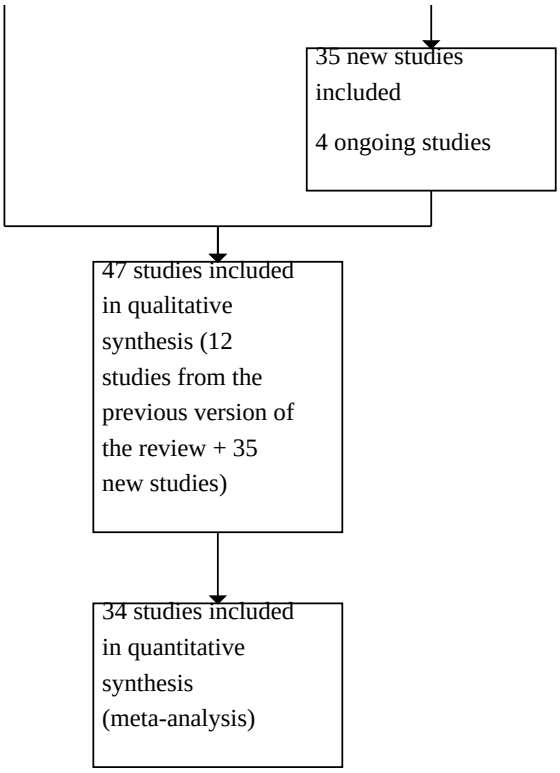


Figure 1. (Continued)



Included studies

The 47 included reviews/overviews were published between January 1998 and May 2022. Of the 35 newly included reviews/overviews, all compared effect estimates of RCTs with those of observational designs. Six reviews focused exclusively on pharmacological interventions (Gu 2020; Hong 2021; Moneer 2022; Morfaw 2021; Naudet 2011; Tzoulaki 2011), while 17 focused on pharmacological and other interventions, but reported data on effect estimates of pharmacological interventions that could be analysed separately (Allain 2017; Ankarfeldt 2017; Benson 2000; Beynon 2008; Dahabre 2012; Ioannidis 2001; Jainaud 2021; Kirson 2013; Kitsios 2015; Li 2016; Mathes 2021; Naudet 2011; Papanikolaou 2006; Safieddine 2021; Shen 2020; Tan 2017; Ziff 2015).

The included reviews and overviews included a total of 2869 RCTs (mean 68.31; range 2 to 727). The number of included RCTs was not reported in four reviews/overviews. The total number of participants included in the RCTs was 3,882,115. However, the total number of participants was unclear or not reported in 25 reviews. The total number of observational studies included in all reviews and overviews was 3924 (mean 60.34; range 5 to 986). The observational studies included a total of 19,499,970 participants.

Of the 47 included reviews/overviews, we included 34 in the quantitative meta-analysis: these reviews either had suitable data available, or we were able to obtain quantitative data from the authors, allowing us to calculate ratios of ratios. Of these 34 reviews, 16 were published in 2020, 2021, or 2022 and contributed 52.4% weight to the overall effect estimate; 18 were published between 2000 and 2019.

Three included reviews reported only continuous outcomes (Ankarfeldt 2017; Artus 2014; Naudet 2011). We did not include these in the meta-analysis but presented their data narratively.

We included three systematic reviews that investigated the effect of mammography screening on breast cancer (Demissie 1998; MacLehose 2000; Schmidt 2013). Of these, we included only the most recent review in the statistical analysis for this intervention-outcome combination (Schmidt 2013). However, we included evidence from MacLehose 2000 in the analysis for another outcome-intervention combination. We included four reviews that were based on similar, overlapping evidence for the outcome of major gastrointestinal bleeding (Gu 2020; Li 2016; Safieddine 2021; Shen 2020). Therefore, we included only the most recent review in our analysis for this outcome (Gu 2020). We included evidence for other outcomes from Safieddine 2021 and Shen 2020 in the quantitative analysis.

We included three reviews – Li 2016, Papanikolaou 2006, and Ziff 2015 – which are themselves included in three overviews of reviews: Papanikolaou 2006 in Golder 2011, Ziff 2015 in Bröckelmann 2022, and Li 2016 in Hong 2021. Therefore, we have described Li 2016, Papanikolaou 2006, and Ziff 2015 below, but we did not include them in the meta-analysis. Two reviews were reported as conference abstracts only and did not report sufficient data to be included in the qualitative analysis (Borkowska 2018; Yanik

2013). We did not include two reviews – Dahabre 2012 and Tzoulaki 2011 – in the primary quantitative analysis because their analyses employed a selective inversion approach which seems not feasible (Sterne 2018). Finally, we did not include the Allain 2017 and Tan 2017 reviews in the meta-analysis as we were unable to obtain sufficient data from them.

See Characteristics of included studies for more details of each included review.

Excluded studies

We excluded 60 reviews/overviews in this update, for the following reasons:

- 25 reviews were meta-analyses that comprised an incidental comparison of RCTs and observational studies, but were not designed for such a comparison;
- 16 reviews were methodological or statistical papers that did not compare the findings of RCTs and observational studies;
- 10 reviews did not conduct their own meta-analyses to compare effect estimates of the two study types but used meta-analyses already performed in published reviews;
- three reviews used a previously published meta-analysis to compare effect estimates from study types;
- three reviews did not describe a systematically compiled sample;
- two reviews contained only a single study available to report on the effect size for one study type;
- one review did not limit its comparison to RCTs and observational studies only.

The Anglemyer 2014 version of the review listed 42 excluded reviews. After re-screening its included reviews, we deemed three previously included reviews to be ineligible for inclusion (Concato 2000a; Oliver 2010; Shikata 2006). Concato 2000a and Oliver 2010 appear to have employed search strategies for studies that were not systematic. This means that other independent researchers would not have been able to replicate the searches and may not have retrieved the same sample of included studies. Shikata 2006 included only meta-analyses of RCTs that were previously conducted by other review authors: hence, Shikata and colleagues did not perform their own meta-analyses of RCTs for their review.

In total, this review update now contains 105 excluded reviews. Please see Characteristics of excluded studies for further details.

Risk of bias in included studies

Please see Figure 2 and Figure 3. Details of our assessments for risk of bias for each included review/overview are presented in the Characteristics of included studies.

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

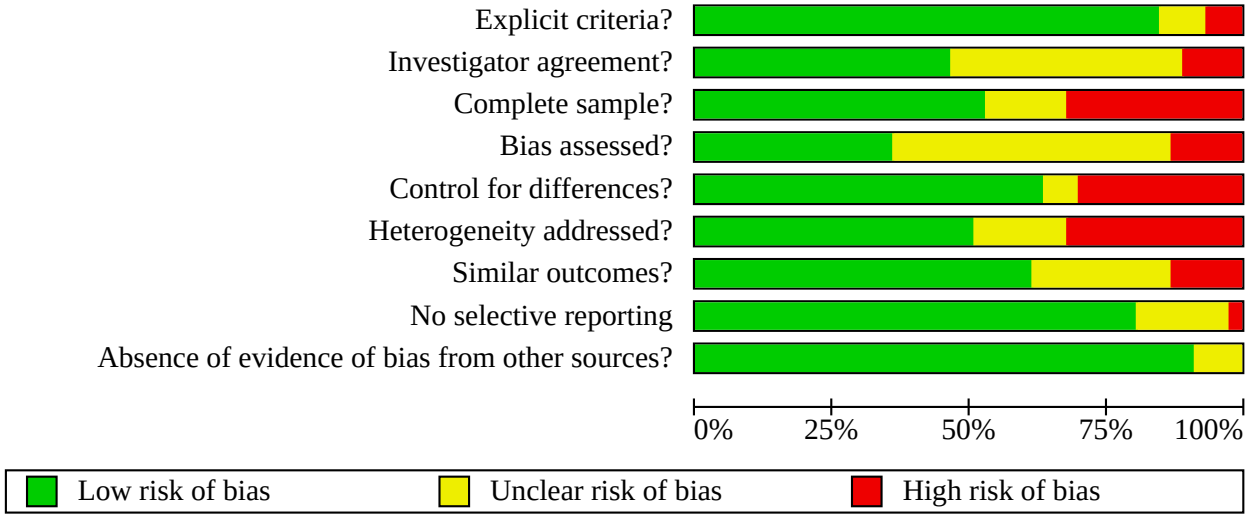


Figure 3. 'Risk of bias' summary: review authors' judgements about each risk of bias item for each included study.

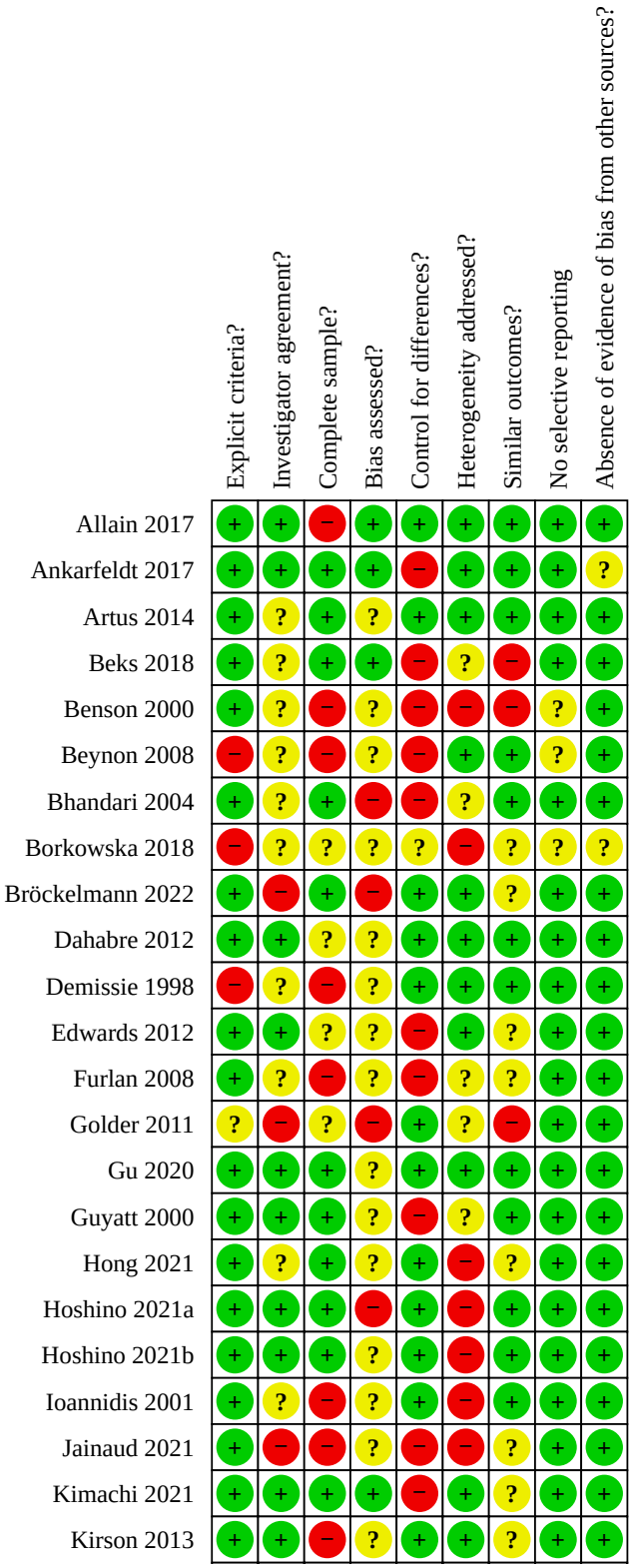
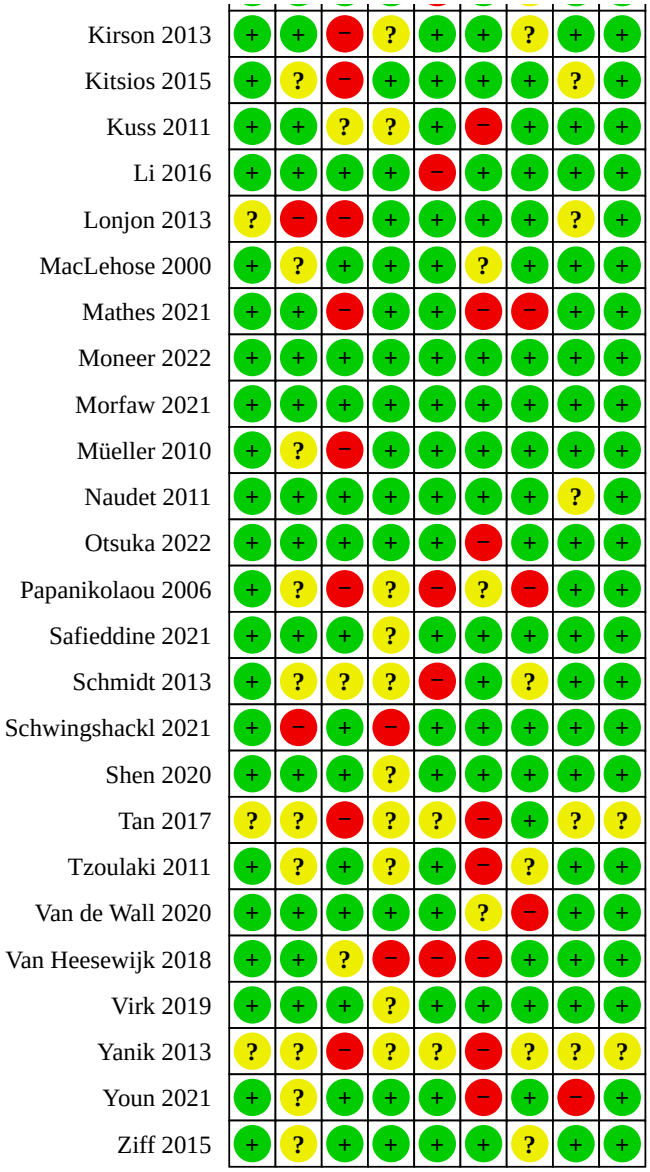


Figure 3. (Continued)



We judged 11 reviews to meet all four of our key risk of bias criteria (explicit criteria for study selection; complete sample of studies; controlled for methodological differences; controlled for heterogeneity), and assessed them as having an overall low risk of bias (Artus 2014; Bröckelmann 2022; Gu 2020; Moneer 2022; Morfaw 2021; Naudet 2011; Safieddine 2021; Schwingshackl 2021; Shen 2020; Virk 2019; Ziff 2015).

For three reviews, the criteria for selecting studies were not clearly reported (Beynon 2008; Borkowska 2018; Demissie 1998). For four included reviews, the eligibility criteria were reported rather vaguely and allowed for a wide inclusion of studies or reviews (Golder 2011; Lonjon 2013; Tan 2017; Yanik 2013).

In five reviews, at least one step of the study selection was conducted by a single reviewer (Bröckelmann 2022; Golder 2011; Jainaud 2021; Lonjon 2013; Schwingshackl 2021). In 20 reviews, the study selection process was not clearly described, and it remained unclear whether two or more independent reviewers were involved in the selection process (Artus 2014; Beks 2018; Benson 2000; Beynon 2008; Bhandari 2004; Borkowska 2018; Demissie 1998; Furlan 2008; Hong 2021; Ioannidis 2001; Kitsios 2015; MacLehose 2000; Müeller 2010; Papanikolaou 2006; Schmidt 2013; Tan 2017; Tzoulaki 2011; Yanik 2013; Youn 2021; Ziff 2015).

We suspected an incomplete sample in 15 reviews (Allain 2017; Benson 2000; Beynon 2008; Demissie 1998; Furlan 2008; Ioannidis 2001; Jainaud 2021; Kirson 2013; Kitsios 2015; Lonjon 2013; Mathes 2021; Müeller 2010; Papanikolaou 2006; Tan 2017; Yanik 2013), whilst for seven reviews, the completeness of the sample of included studies was unclear (Borkowska 2018; Dahabre 2012; Edwards 2012; Golder 2011; Kuss 2011; Schmidt 2013; Van Heesewijk 2018).

Of the included reviews, 22 reported no information about whether risk of bias was assessed (Artus 2014; Benson 2000; Beynon 2008;

Borkowska 2018; Dahabre 2012; Demissie 1998; Edwards 2012; Furlan 2008; Guyatt 2000; Hong 2021; Hoshino 2021b; Ioannidis 2001; Jainaud 2021; Kirson 2013; Kuss 2011; Papanikolaou 2006; Safieddine 2021; Schmidt 2013; Tan 2017; Tzoulaki 2011; Virk 2019; Yanik 2013). In three reviews, it was reported that study quality was assessed in some way, but the assessment was based on inappropriate or limited tools (Bhandari 2004; Golder 2011; Van Heesewijk 2018). In three reviews, two of which are overviews of reviews, it was reported that risk of bias was not assessed (Bröckelmann 2022; Hoshino 2021a; Schwingshackl 2021). In one review, the risk of bias was assessed for one study type, but there was no information about the assessment for the other study type (Gu 2020). In another review, the risk of bias for one study type was assessed adequately, but the risk of bias assessment of the other study type was based on superficial criteria (Shen 2020).

An appropriate control for differences in the compared RCTs and observational studies was not reported at all in three reviews (Borkowska 2018; Tan 2017; Yanik 2013). Fourteen reviews analysed RCTs and observational studies separately, but did not further account for methodological differences with the different study types (Ankarfeldt 2017; Beks 2018; Benson 2000; Beynon 2008; Bhandari 2004; Edwards 2012; Furlan 2008; Guyatt 2000; Jainaud 2021; Kimachi 2021; Li 2016; Papanikolaou 2006; Schmidt 2013; Van Heesewijk 2018).

Heterogeneity in the populations or interventions was not assessed or reported in 15 reviews and therefore assessed as high risk of bias (Benson 2000; Borkowska 2018; Hong 2021; Hoshino 2021a; Hoshino 2021b; Ioannidis 2001; Jainaud 2021; Kuss 2011; Mathes 2021; Otsuka 2022; Tan 2017; Tzoulaki 2011; Van Heesewijk 2018; Yanik 2013; Youn 2021). In six reviews, heterogeneity was addressed but not systematically controlled for (Beks 2018; Furlan 2008; Golder 2011; Guyatt 2000; Papanikolaou 2006; Van de Wall 2020). Two reviews did not control for heterogeneity, but had such narrow eligibility criteria that the resulting risk of bias was assessed as unclear (Bhandari 2004; MacLehose 2000).

We judged that six reviews analysed data for outcomes that were insufficiently similar to warrant pooling in a meta-analysis (Beks 2018; Benson 2000; Golder 2011; Mathes 2021; Papanikolaou 2006; Van de Wall 2020). In one of these six reviews, the main outcome was only reported in 14 of 22 included studies (Beks 2018). We therefore assessed this review as having a high risk of bias based on the low similarity of outcomes across included studies. Twelve reviews did not report enough information on the selection of studies and eligible outcomes or definitions of outcomes to allow for an assessment of the similarity of the outcomes analysed (Borkowska 2018; Bröckelmann 2022; Edwards 2012; Furlan 2008; Hong 2021; Jainaud 2021; Kimachi 2021; Kirson 2013; Schmidt 2013; Tzoulaki 2011; Yanik 2013; Ziff 2015).

Selective reporting

We suspected selective reporting in one review (Youn 2021), which was missing details of subgroups' characteristics (such as participant numbers) and the level of heterogeneity of the meta-analysed data. We assessed eight reviews as having an unclear risk for selective reporting (Benson 2000; Beynon 2008; Borkowska 2018; Kitsios 2015; Lonjon 2013; Naudet 2011; Tan 2017; Yanik 2013).

Other potential sources of bias

We did not detect any other potential sources of bias in 43 of the included reviews/overviews. We rated four reviews as unclear: three reviews reported as conference abstracts reported too little information to assess the risk of other biases (Borkowska 2018; Tan 2017; Yanik 2013), and one review listed a number of biases that might affect their results, including plausible dissemination bias (Ankarfeldt 2017).

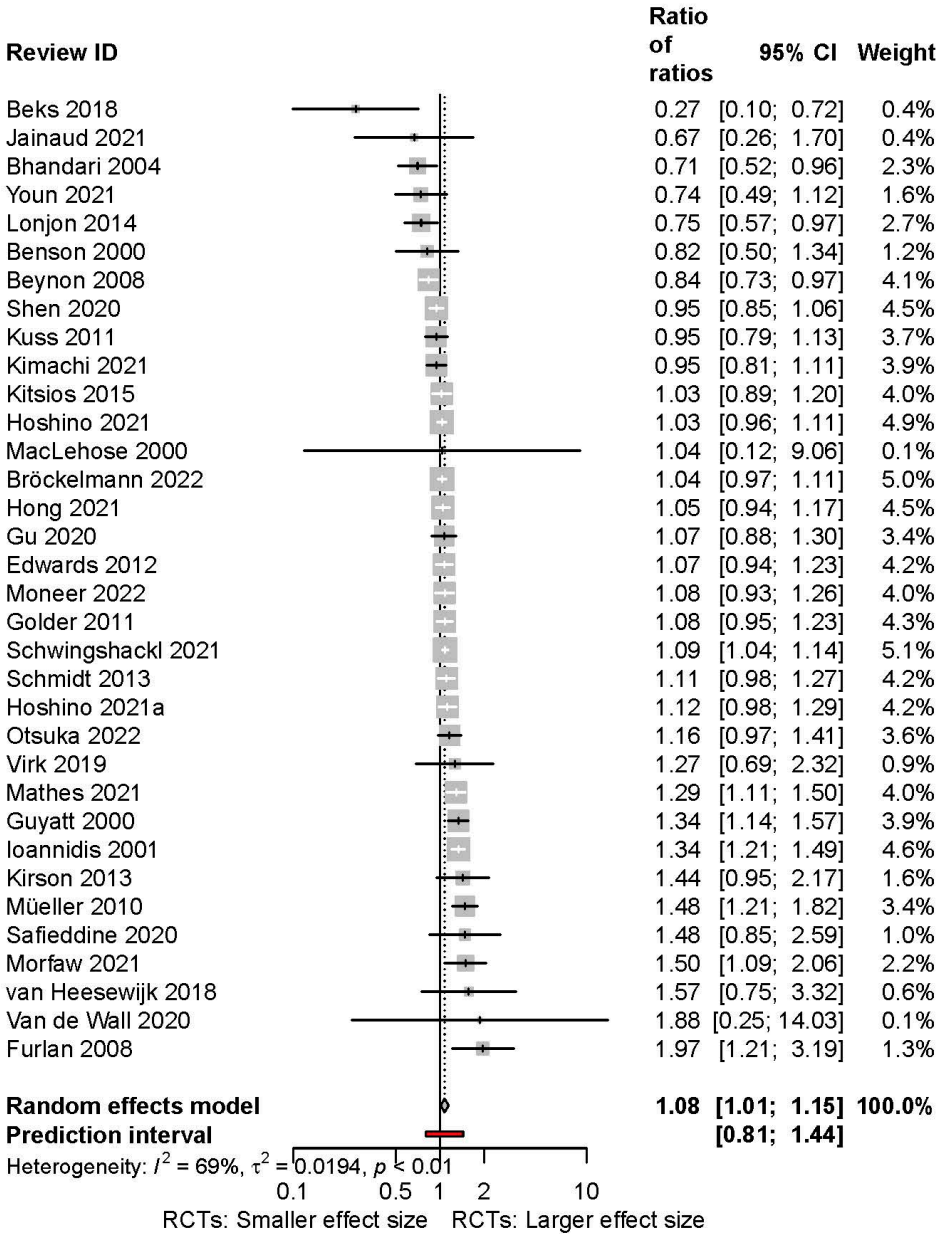
Effect of methods

Difference in effect estimates between RCTs and observational studies

Dichotomous outcomes

We included 34 reviews with 2138 RCTs and 2785 observational studies in the primary analysis. When pooling effect estimates reported as ROR and RRR together, the summary effect estimate indicated no difference or a very small difference between the effect estimates from RCTs versus those from observational studies (ratio of ratios 1.08, 95% CI 1.01 to 1.15; I^2 = 69%; 34 reviews; low-certainty evidence) (see Figure 4 and Summary of findings 1). Twenty-three of 34 reviews reported effect estimates of RCTs and observational studies that were on average in agreement (Benson 2000; Bröckelmann 2022; Edwards 2012; Furlan 2008; Golder 2011; Gu 2020; Hong 2021; Hoshino 2021a; Hoshino 2021b; Jainaud 2021; Kimachi 2021; Kirson 2013; Kitsios 2015; Kuss 2011; MacLehose 2000; Moneer 2022; Otsuka 2022; Safieddine 2021; Schmidt 2013; Shen 2020; Van de Wall 2020; Van Heesewijk 2018; Virk 2019; Youn 2021).

Figure 4. Forest plot for main analysis



Subgroup analyses

Pharmacological interventions versus non-pharmacological interventions

Six reviews focused exclusively on pharmacological interventions (Gu 2020; Hong 2021; Moneer 2022; Morfaw 2021; Naudet 2011; Tzoulaki 2011), while 17 focused on pharmacological and other interventions, but reported data on effect estimates of pharmacological interventions that could be analysed separately (Allain 2017; Ankarfeldt 2017; Benson 2000; Beynon 2008; Dahabre 2012; Ioannidis 2001; Jainaud 2021; Kirson 2013; Kitsios 2015; Li 2016; Mathes 2021; Naudet 2011; Papanikolaou 2006; Safieddine 2021; Shen 2020; Tan 2017; Ziff 2015). In 15 reviews, pharmacological interventions were investigated and the ratio of ratios indicated slight differences between effect estimates by study type (ratio of ratios 1.12, 95% CI 1.04 to 1.21; I^2 = 49%; 15 reviews). In bodies of evidence comparing the effect of non-pharmacological interventions, the ratio of ratios indicated agreement between the study types (ratio of ratios 1.06, 95% CI 0.93 to 1.20; I^2 = 79%; 21 reviews). In four reviews, the investigated interventions were unclear or varied (ratio of ratios 1.04, 95% CI 0.98 to 1.09; I^2 = 0%; 4 reviews).

Low to moderate versus substantial and high heterogeneity

We considered an I^2 estimate of between 30% and 50% to indicate moderate heterogeneity, 51% to 80% substantial heterogeneity, and 81% to 100% as a high level of heterogeneity. There seems to be general agreement between effect estimates from RCTs and observational studies that displayed low and moderate heterogeneity (ratio of ratios 1.06, 95% CI 0.96 to 1.17; I^2 = 75%; 18 reviews). There seems to be a small difference between effect estimates from RCTs and observational studies that had substantial and high heterogeneity (ratio of ratios 1.11, 95% CI 1.04 to 1.18; I^2 = 61%; 11 reviews). In the five reviews that did not report the statistical heterogeneity of meta-analysed RCTs and observational studies, the effect estimates from RCTs and observational studies were in overall agreement (ratio of ratios 0.75, 95% CI 0.42 to 1.34; I^2 = 59%; 5 reviews).

Propensity score adjustment used versus propensity score adjustment not used

In seven reviews that clearly reported a comparison of RCTs with observational studies that employed propensity score adjustment, on average the effect estimates from the two study types were in agreement (ratio of ratios 1.00, 95% CI 0.86 to 1.17; I^2 = 58%; 7 reviews). In 11 reviews of RCTs and observational studies that did not use propensity score adjustment methods, the effect estimates from RCTs and observational studies seemed to differ slightly (ratio of ratios 1.07, 95% CI 1.03 to 1.11; I^2 = 59%; 11 reviews). In 19 reviews, the use of propensity score adjustment in observational studies was not reported or unclear. The effect estimates from RCTs and observational studies in these reviews seemed to be in slight disagreement (ratio of ratios 1.13, 95% CI 1.03 to 1.25; I^2 = 73%; 19 reviews).

RCTs versus cohort studies or case-control studies

When looking at the comparison of bodies of evidence of RCTs and observational studies – which were either specifically labelled as "observational studies" or not further specified – their effect estimates were mostly in agreement (ratio of ratios 1.06, 95% CI 0.96 to 1.18; I^2 = 79%; 20 reviews). When looking at observational studies more specifically, the comparison of bodies of evidence from RCTs versus bodies of evidence from cohort studies show slight differences in their effect estimates (ratio of ratios 1.09, 95% CI 1.04 to 1.13; I^2 = 34%; 14 reviews). In observational studies clearly described as case-control studies, the effect estimate seems to be similar to that of RCTs (ratio of ratios 1.00, 95% CI 0.80 to 1.25; I^2 = 13%; 2 reviews).

Overviews of reviews versus systematic reviews

Effect estimates from RCTs and observational studies seemed similar in comparisons from systematic reviews (ratio of ratios 1.07, 95% CI 0.99 to 1.17; I^2 = 69%; 28 reviews), but not from overviews of systematic reviews (ratio of ratios 1.09, 95% CI 1.00 to 1.19; I^2 = 75%; 6 overviews).

Visual inspection of effect estimates

Dichotomous outcomes

For the dichotomous outcomes, we conducted a visual inspection at review level, which allowed for the investigation of similarities and differences in meta-analysed effect estimates from RCTs and observational studies per review. Effect estimates for individual dichotomous outcomes per study type in the included reviews are displayed in Figure 5.

Figure 5. Forest plot for the visual inspection of effect estimates for dichotomous outcomes from bodies of evidence from RCTs and observational studies RCT - randomised controlled trial, OBS - observational study, ICU - intensive care unit, m - male, f - female

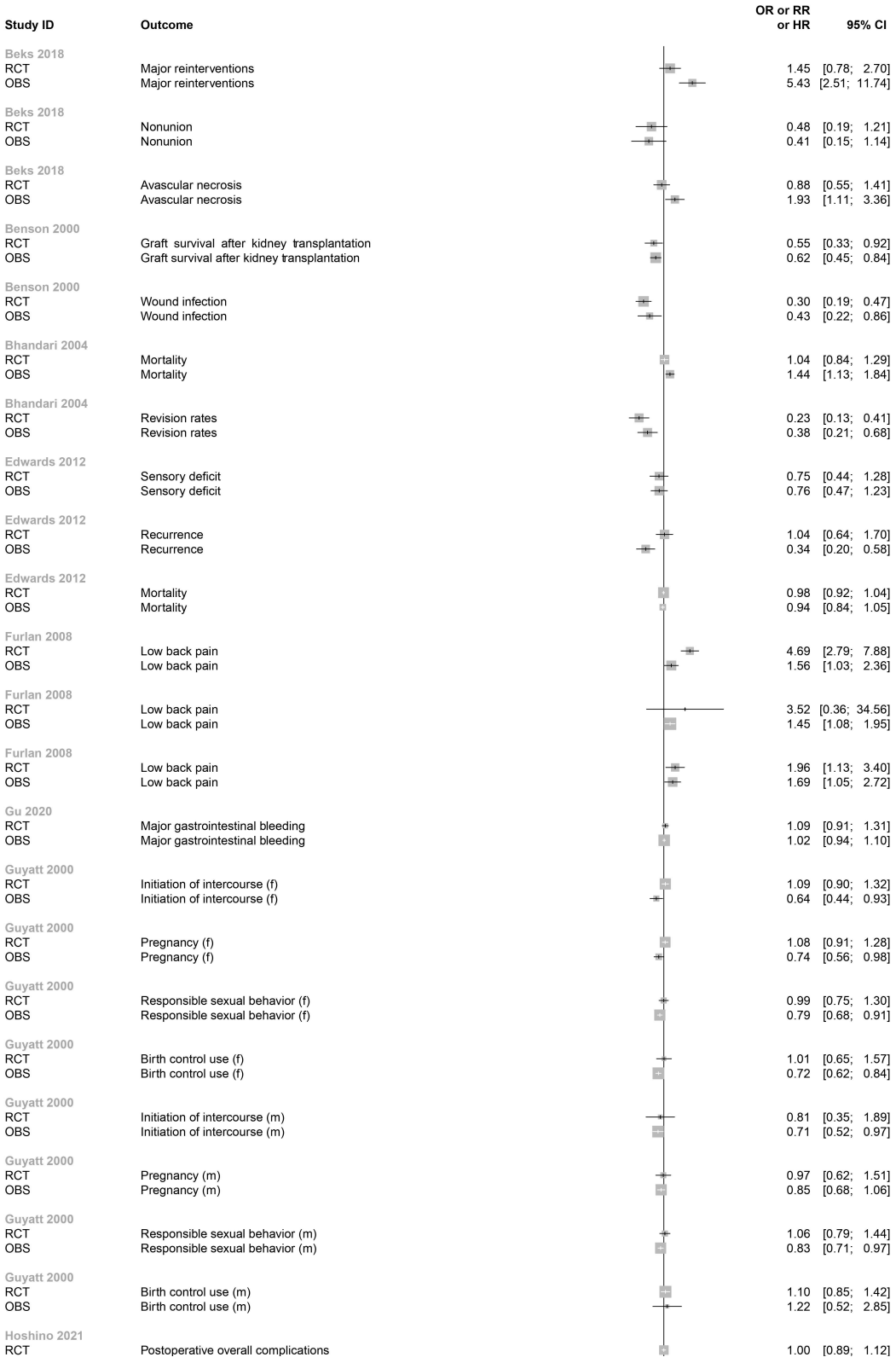


Figure 5. (Continued)

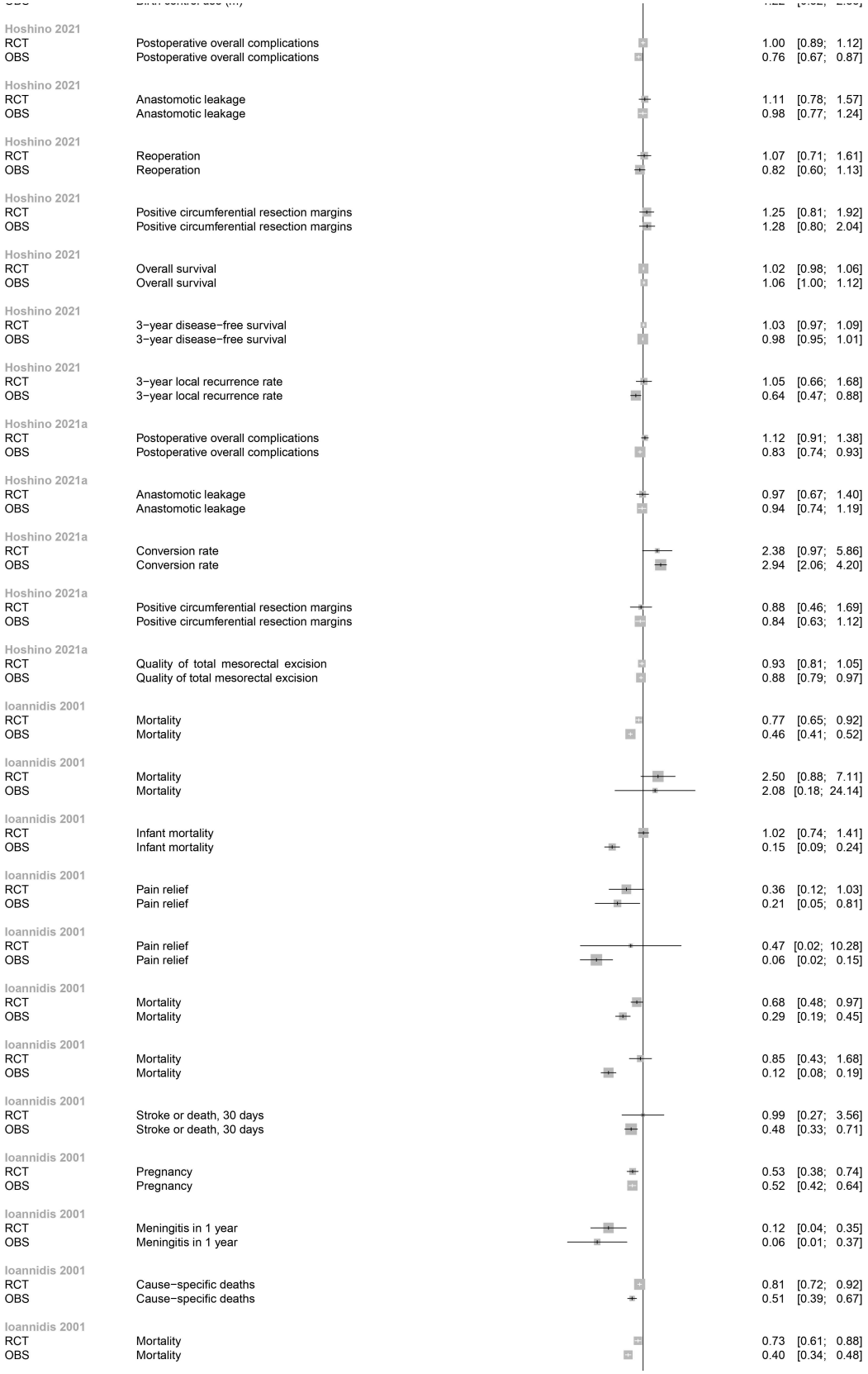


Figure 5. (Continued)

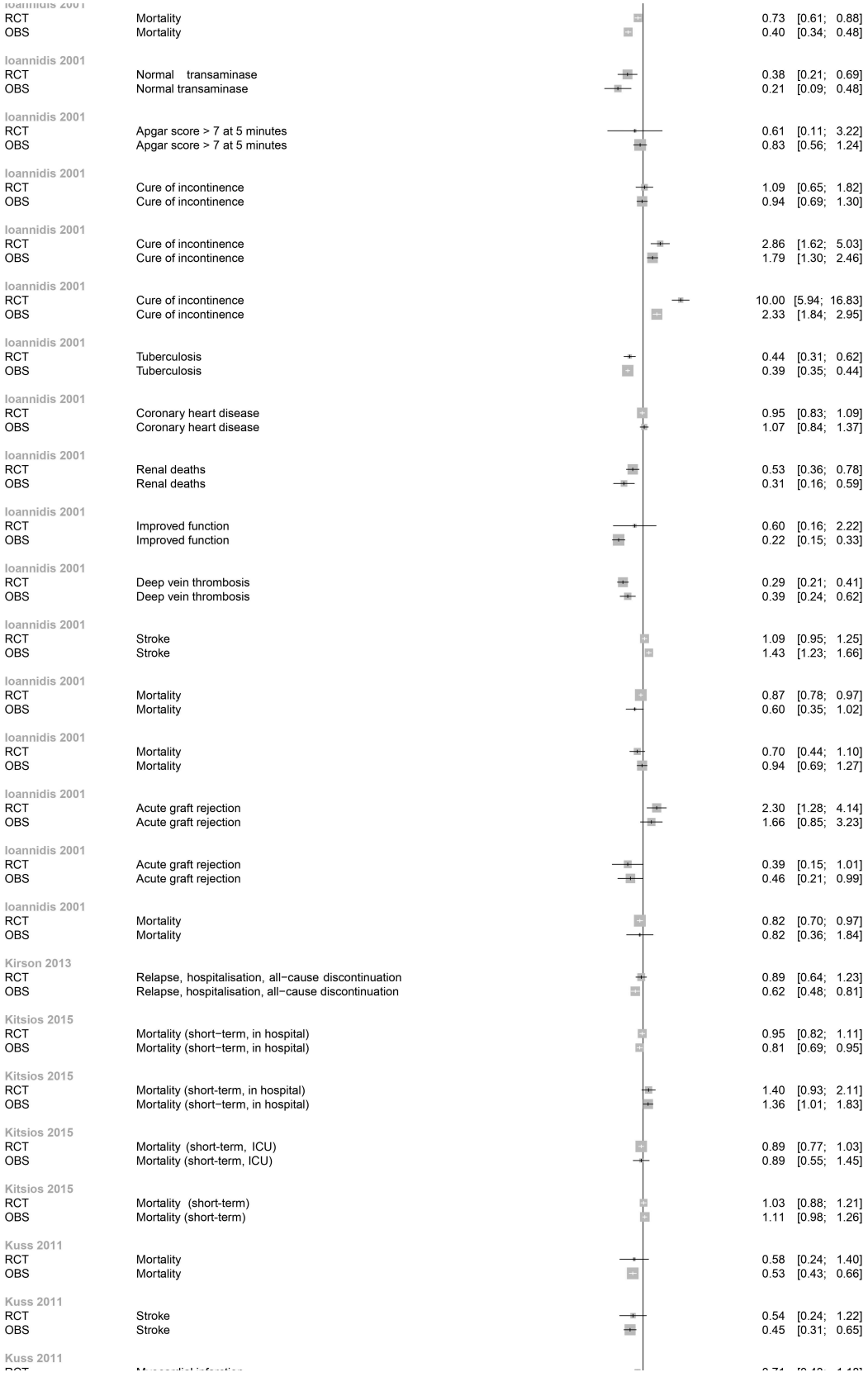


Figure 5. (Continued)

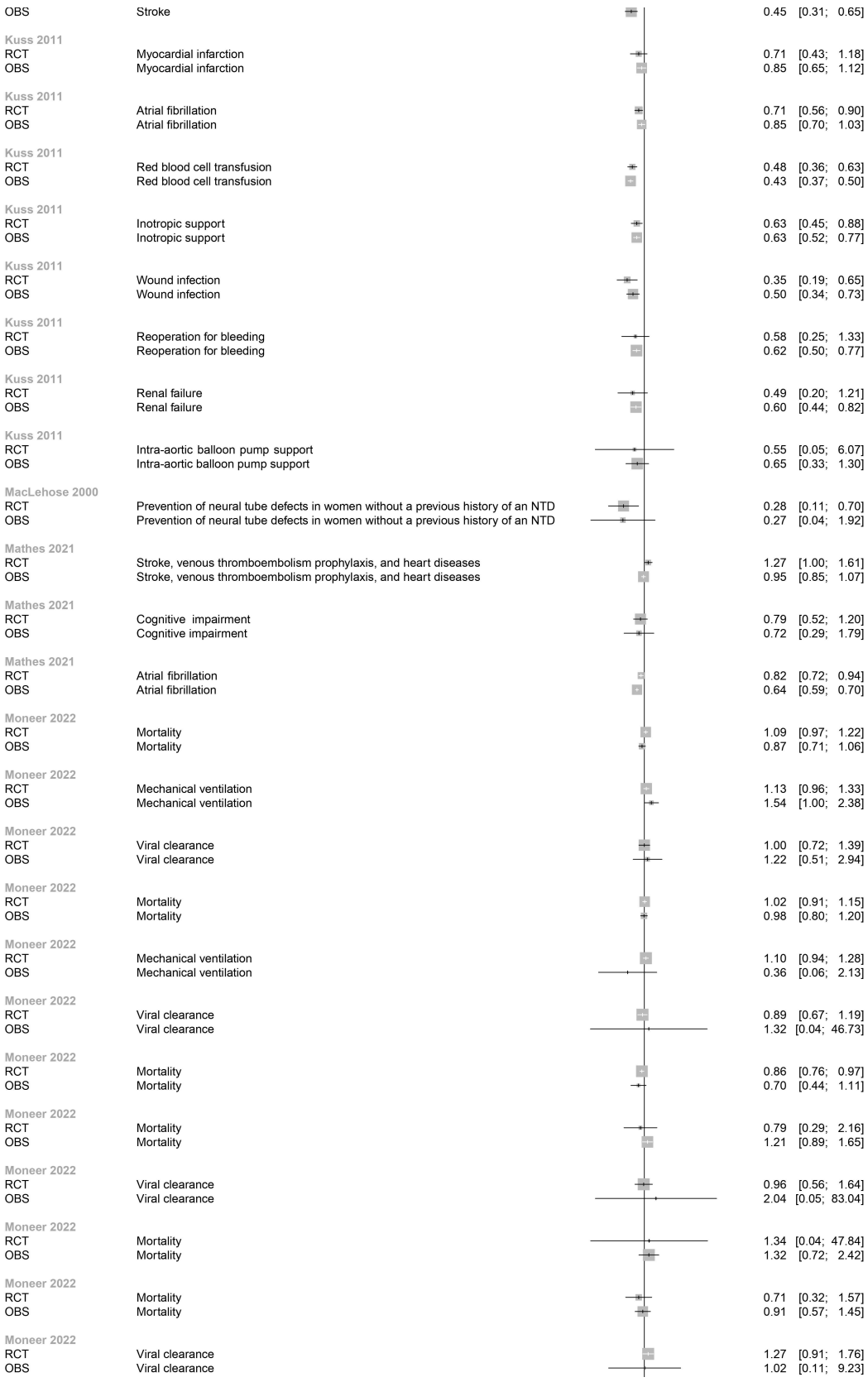


Figure 5. (Continued)

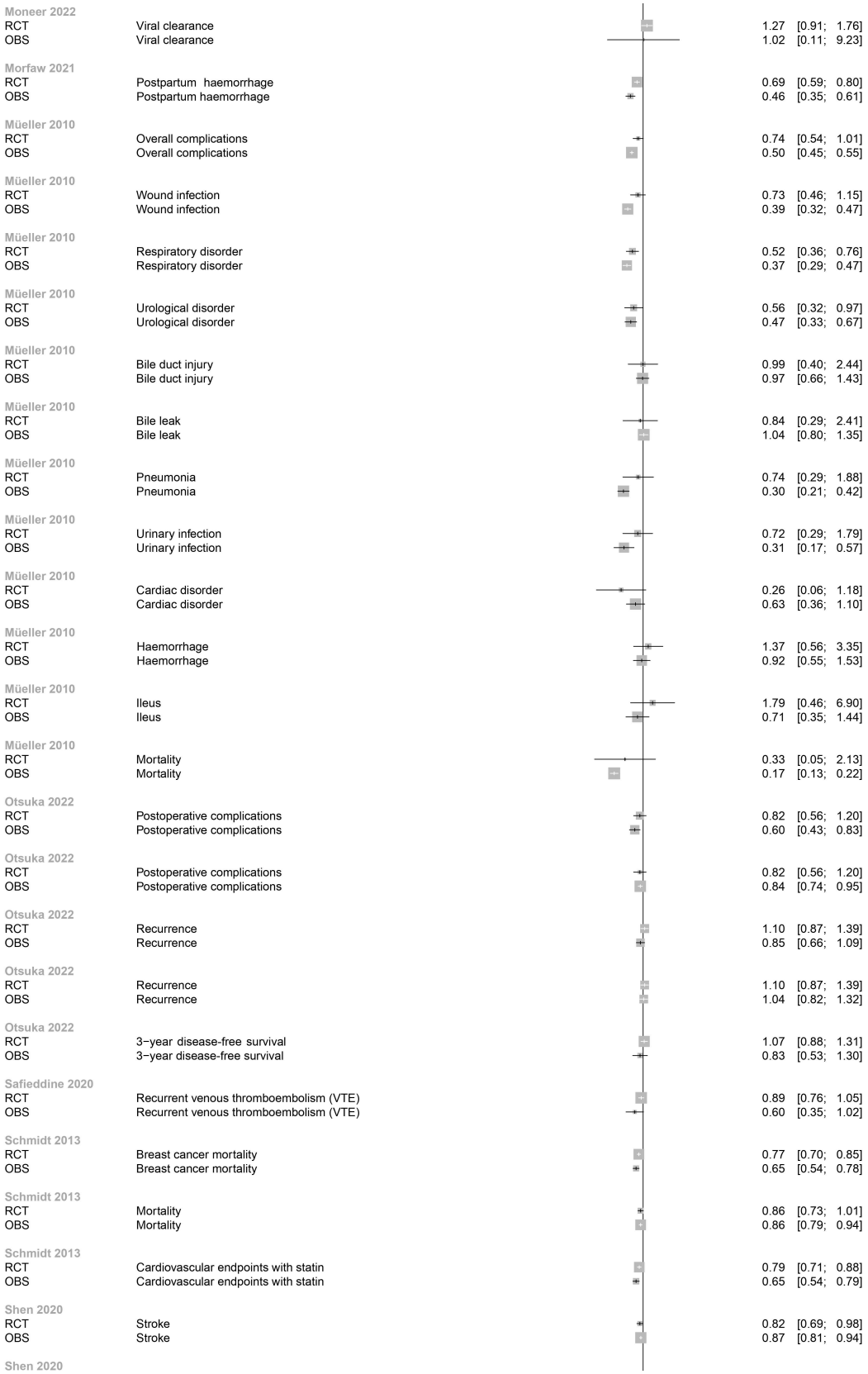
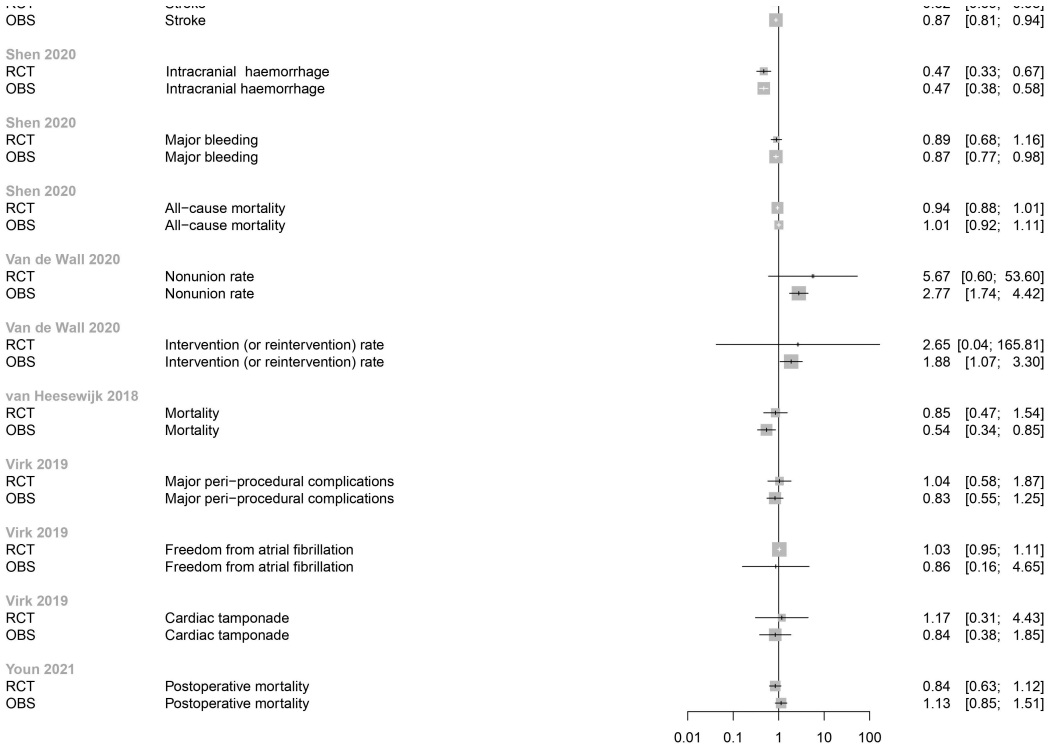


Figure 5. (Continued)

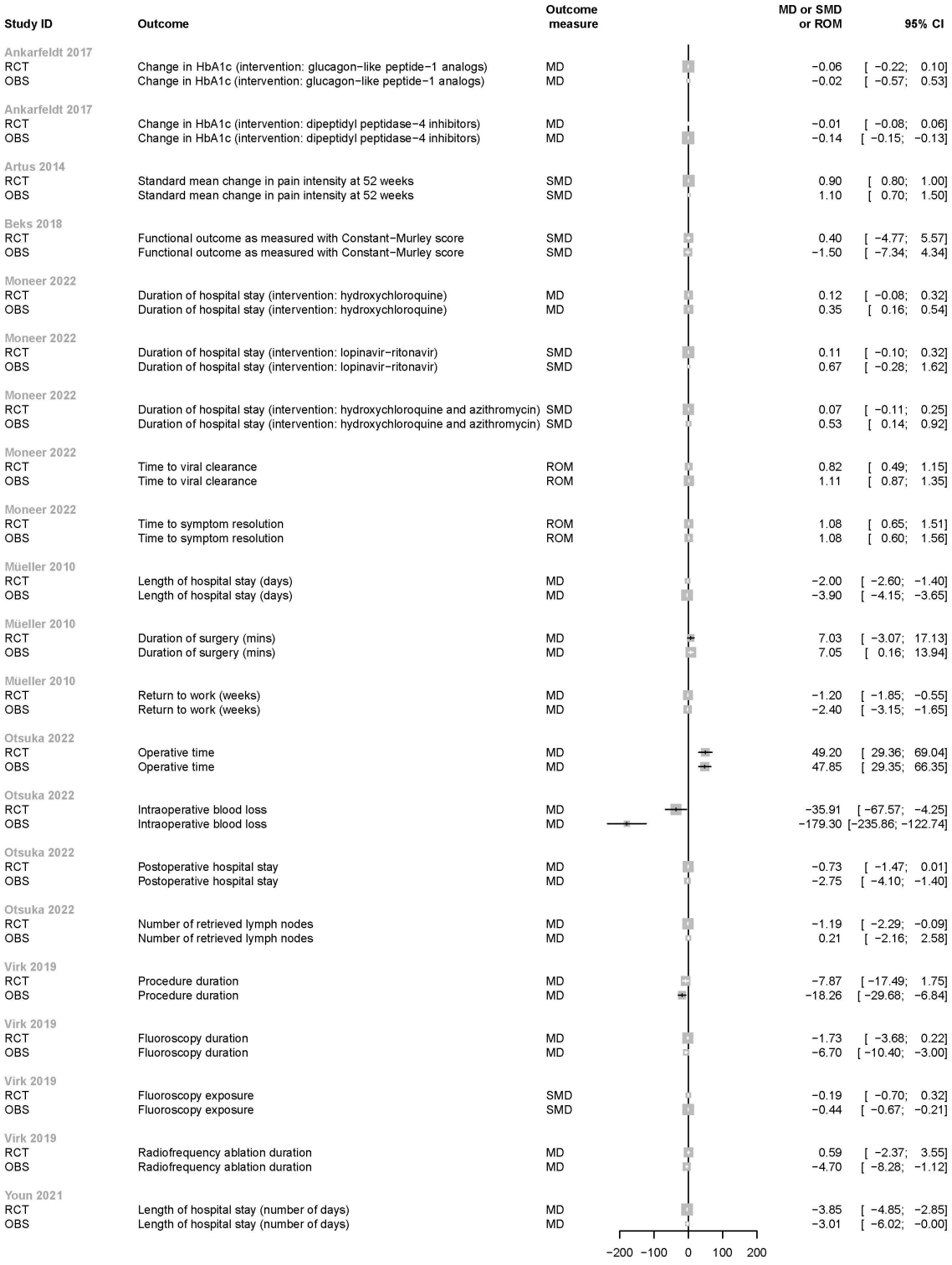


Twenty-eight reviews reported sufficient data to be considered in the visual inspection. Of these 28, 10 pooled effect estimate pairs were probably discordant in effect-estimate direction. All 28 pairs of effect estimates had overlapping 95% CIs. The 95% CIs of 10 effect estimates from observational studies did not include 1. Only half of these also had a corresponding RCT effect estimate whose 95% CI did not include 1. Overall, the 95% CIs of seven effect estimates from RCTs did not include 1.

Continuous outcomes

The effect measures and the outcomes reported as continuous effect measures in the included reviews were too different from each other to pool across reviews. Hence, we visually inspected the individual comparisons in these reviews. Effect estimates for individual continuous outcomes per study type in the included reviews are displayed in Figure 6. Meta-analyses of continuous outcomes as reported in the systematic reviews are reported in Appendix 3.

Figure 6. Forest plot for the visual inspection of effect estimates for continuous outcomes from bodies of evidence from RCTs and observational studies HbA1c - hemoglobin A1c; RCT - randomised controlled trial, OBS - observational study



For the majority of the 22 comparisons, the effect estimates from RCTs and observational studies were in agreement. For 21 comparisons, there were sufficient data to conduct a visual inspection. Seventeen pairs of effect estimates were in concordant direction. In the four comparisons where the effect estimates of RCTs and observational studies were in opposite directions, their

CIs overlapped. Overall, in 20 of 22 comparisons, the 95% CIs of RCTs and observational studies overlapped. In one instance, the 95% CIs did not overlap, but the effects were in a concordant direction, did not include 0 (no effect), and there was a larger effect in the observational studies. The CI of eight RCTs and 15 observational studies did not include 0 (no effect). In seven pairs of effect estimates, for both RCTs and observational studies, 0 (no effect) was not included in the 95% CI.

Sensitivity analyses

When adding findings from the two reviews that used a selective inversion approach to the primary analysis, the summary effect estimate attenuated very slightly to also include the null-effect (ratio of ratios 1.06, 95% CI 1.00 to 1.13; I^2 = 71%; 36 reviews).

A similar effect was observed when the primary analysis was repeated without reversal of favourable outcomes (ratio of ratios 1.05, 95% CI 1.00 to 1.11; I^2 = 65%; 34 reviews).

DISCUSSION

Summary of main results

Our results showed that, on average, there is little difference between the effect estimates obtained from RCTs and observational studies (ratio of ratios 1.08, 95% CI 1.01 to 1.15). In several subgroup analyses, we noted small differences between the effect estimates of study types. These differences were seen: in comparisons of pharmaceutical interventions only (ratio of ratios 1.12, 95% CI 1.04 to 1.21); in RCTs and observational studies with substantial or high heterogeneity (ratio of ratios 1.11, 95% CI 1.04 to 1.18); in comparisons with no use or unclear use of propensity score adjustment in observational studies (ratio of ratios 1.07, 95% CI 1.03 to 1.11; ratio of ratios 1.13, 95% CI 1.03 to 1.25, respectively); and where observational studies without further specification of the study design were compared to RCTs (ratio of ratios 1.06, 95% CI 0.96 to 1.18). However, due to the substantial clinical and statistical heterogeneity, there may be important differences between subgroups of reviews that we were unable to identify.

It is possible that the difference between effect estimates obtained from RCTs and observational studies has decreased in recent years due to researchers' improved understanding of how to handle adjustments in observational studies. On the same note, there seems to be a small difference between the two study types when observational studies that do not use propensity score adjustment methods are the comparator (ratio of ratios 1.07, 95% CI 1.03 to 1.11). In the present review, it was not always very clear which observational studies included adjusted effect estimates and which did not in the included reviews. Bhandari and colleagues 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 confounders. Mueller and colleagues 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). Our results suggest that although observational designs may be at higher risk of bias than RCTs, this does not consistently result in substantially different effect estimates.

We also found that the effect estimate differences between observational studies and RCTs were potentially influenced by the heterogeneity within meta-analyses. Prespecified subgroup analyses comparing effect estimates by heterogeneity indicated that there might be relevant differences between effect estimates of RCTs and observational studies where substantial or high heterogeneity was observed. Meta-analyses of RCTs and observational studies may be particularly influenced by heterogeneity, and researchers should, for example, consider the risk of bias in the included RCTs and observational studies or the specific types of outcomes investigated, when designing such comparisons.

In addition, analyses that take into account the risk of bias of the individual included studies could help to interpret differences in effect estimates. In this review, we found several reviews that conducted sensitivity analyses excluding only studies at high or low risk of bias. The tools to assess risk of bias in RCTs and observational studies varied across reviews. The risk of bias in the included reviews was generally high and only 11 out of 47 reviews met the key criteria for low risk of bias. In particular, around one-third of all included reviews either did not include a complete sample or there was not enough information provided to make a determination, and around 60% of the reviews did not assess the risk of bias of their included studies at all or with a validated tool. Furthermore, nearly half of the included reviews did not report heterogeneity within the bodies of evidence from RCTs or observational studies. More than half did not control for differences such as study design within observational studies; for example, researchers conducted aggregate analyses for observational studies instead of separating cohort studies from case-control studies or other designs.

Overall completeness and applicability of evidence

We have conducted a thorough literature search, including four electronic databases, backward and forward citation tracking, and contacting authors. For the searches, a sensitive literature search strategy was used. This should have ensured that a complete sample of relevant reviews was included in our review. We revised the search strategy for this review update (see Differences between protocol and review). However, we believe that these changes were faithful to the review's original purpose. Moreover, the revised searches led to a set of 47 included reviews and a set of conclusions in line with the conclusions of the previous version (Anglemyer 2014). Hence, a potential effect of the revised search strategy on the conclusions of the review appears to be very unlikely.

The included reviews addressed a wide array of clinical topics and included various populations, interventions (pharmacological and non-pharmacological), comparators, and outcomes. This supports the external validity of the findings of this review.

Some of the included reviews discussed underlying reasons for the differences in effect estimates. It became clear that there might be differences in how RCTs and observational studies are conducted in specific clinical areas, and that each area might have its own reasons for systematic differences between observational studies and RCTs. For example, selection bias was reportedly suspected in RCTs in several reviews (Li 2016; Lonjon 2013; Müeller 2010; Schmidt 2013). Artus and colleagues suspected that the "care effect" – a result of care throughout the trial – and the "protocol effect" – the effect of adhering strictly to treatment protocols – impacted on the estimated effects of RCTs (Artus 2014). Hence, the evidence for differences and similarities between effect estimates from different study types must also account for the particular topic area which is addressed by the review and studies.

Another possible source of differences in effect estimates between study types is that review authors may have classified comparisons in RCTs and observational studies as similar, even if they partially differed. Analysis methods such as the target-trial approach might currently be best suited for analysing data from observational studies. Emulations of RCTs with observational data indicate that the selection of the comparator intervention, its administration, and the population receiving the comparator intervention, in particular, might impact on the generation of findings from observational data that are comparable to those of RCTs (Franklin 2021). However, the observational studies in the included reviews did not use the target-trial analysis method. A focus on comparisons of RCTs with observational studies using the target-trial approach might yield findings that differ from the findings of our main analysis.

Our findings report on an overall comparison of meta-analysed effect estimates from RCTs and observational studies. These meta-analysed effect estimates were again meta-analysed in our primary analysis. Heterogeneity might likely be present in the different levels of this comparison. First, the pooled effect estimates from individual RCTs and observational studies might be heterogeneous within the individual reviews. Second, the meta-analysed summary effect estimates of RCTs and observational studies across all included reviews might be affected by heterogeneity to different degrees. We accounted for this in subgroup analyses and found that the effect estimates of studies with reportedly low and moderate to high heterogeneity did not clearly differ from one another. Third, we found unexplained statistical heterogeneity in the summary effect estimate of our main analysis (I^2 = 69%) that could not be explored further. This might hint at high statistical heterogeneity in the two underlying levels of data. This is supported by our finding that 10 out of 28 comparisons were probably discordant in direction, still with overlapping 95% CIs, upon visual inspection. All levels at which heterogeneity might occur should be carefully considered when interpreting the findings of this review. Exploratory 95% prediction intervals for our primary analysis ranged from 0.81 to 1.44, and indicate that the difference between the effect estimates of the two study types might be more pronounced in either direction.

Besides the potential impact of heterogeneity on our findings, the potential but unknown risk of bias in the underlying data informing our analysis should be considered when interpreting the findings. We did not formally assess the risk of bias at the level of each study included in each included review. We only extracted information from the review on whether a risk of bias assessment of the individual studies was conducted. In the included reviews, a wide array of approaches was used to assess the risk of bias in RCTs and observational studies, with only a few review authors reporting that they had used validated and standardised tools, such as the Cochrane risk of bias tool for RCTs or the Cochrane Risk Of Bias In Non-randomized Studies of Interventions/Exposure (ROBINS-I) for observational studies. A more careful consideration of the potential biases in each study design, and especially observational studies, would have been useful to highlight the limitations of the underlying data, and aid a more accurate interpretation of our findings. However, these methods were beyond the scope of this review.

Gaps in the reporting of the findings of the reviews limited our ability to categorise the included observational study designs more specifically as cohort studies, case-control studies, or another study design. Often, the observational study designs also included non-randomised studies.

Certainty of the evidence

We assessed the certainty of the evidence for our main analysis (see Summary of findings 1) and found low-certainty evidence. We down rated the evidence certainty for serious concerns about risk of bias in the included evidence, as well as serious concerns about inconsistency in the included comparisons with unexplained statistical heterogeneity of 69%.

The included reviews varied considerably in terms of the reported populations, interventions, and outcomes. Because there was an insufficient number of reviews with similar characteristics, we were unable to conduct subgroup analyses that might explain heterogeneity. In addition, the reporting of the included reviews lacked important details, such as the specific study designs included and compared, information about the use of any type of adjustment or propensity score adjustment, and information about statistical heterogeneity.

Potential biases in the review process

We reduced the likelihood of bias in our review process by: imposing no language restrictions on our search; and by having two review authors independently screen titles, abstracts, and full texts for eligibility and a third review author adjudicate any conflicts. Additionally, two review authors independently conducted data extraction and risk of bias assessment. Nevertheless, we acknowledge the potential for the introduction of an unknown risk of bias in our methods as we collected a myriad of data from 43 reviews. Two review authors (LS, SB) were co-authors of included reviews (Bröckelmann 2022; Schwingshackl 2021). To minimise the risk of bias, the data of these two reviews was extracted by researchers (IT, MT) other than the two co-authors.

The risk of bias tool we used was designed to cover relevant aspects within the included reviews that could potentially introduce a risk of bias to our findings. Still, the tool was only used in one other review and is not validated by other means (Odgaard-Jensen 2011). Hence, in our assessment, we may inadvertently have missed relevant aspects that might have introduced a risk of bias to our findings. For example, the risk of bias tool assesses homogeneity within RCTs and observational studies in a rather broad and generalised manner. A thorough matching of topics, including the population, intervention, comparator, and outcomes, in RCTs and observational studies was done with varying levels of detail in the included reviews. This might have led to a comparison of heterogeneous bodies of evidence.

The findings of this review might further be at risk of bias because heterogeneity amongst study types was not assessed and reported in all included systematic reviews or overviews. Strong heterogeneity amongst one study type might lead to imprecise summary effect estimates. Such effect estimates with wide 95% CIs are subsequently more likely to overlap with 95% CIs from meta-analysis of the other study type. Hence, a clear difference between the study types will not be demonstrated.

Finally, the feasibility of RORs and RRRs as effect measures has been discussed in the scientific literature. Their feasibility was found to be dependent on the direction of the comparison (Franklin 2017). Our analysis did not account for the particular interventions and

comparators in the included reviews and might therefore be at risk of bias. In order to account for this, we also added an individual comparison of effect estimates and CIs at review level.

Agreements and disagreements with other studies or reviews

Our primary analysis shows no to very small differences between the effect estimates of RCTs and observational studies (ratio of ratios 1.08, 95% CI 1.01 to 1.15). This largely reflects what was found in the previous version of this review (Anglemyer 2014), and in other reviews investigating the differences and similarities in effect estimates of different study types.

Golder 2011 – and consequently, Papanikolaou 2006 – were the only reviews that focused exclusively on adverse effects. Golder and colleagues' findings do not support the notion that observational studies are more likely to detect harm than RCTs, 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 that Golder 2011 studied; in this context, one might expect shorter-term trials to be as likely to detect harm as longer-term observational studies. Golder and colleagues reported that larger studies reported more precise effect estimates, which might be due to better study quality or because of more homogenous populations within larger studies.

Hong and colleagues report that their overview of reviews shows variation in the consistency of effect estimates as reported by observational studies and RCTs (Hong 2021). Jainaud and colleagues conducted an overview of umbrella reviews of RCTs and observational studies that assessed risk and protective factors (Jainaud 2021). Overall, the authors found that effect estimates showed different directions between RCTs and observational studies in 37.1% of the comparisons they studied. In 43.5% of the comparisons, they found effect estimates differing beyond chance in the two study types. In another overview of systematic reviews (Kimachi 2021), the authors found that, on average, effect estimates were consistent between the study types.

In an evaluation of differences in the effect estimates of RCTs and observational studies in the research field of nutrition, Schwingshackl and colleagues analysed 97 diet-disease outcome pairs (Schwingshackl 2021). They found that the RRR comparing RCTs and cohort studies showed, on average, a small difference in the effect estimates of the two study types (RRR 1.09, 95% CI 1.04 to 1.14). Still, the authors found that clinical heterogeneity explained most of the differences between the effect estimates. In comparisons where the intake and exposure were similar in both study types, differences in effect estimates seemed to attenuate.

In a `meta-epidemiological` study that used similar methods as Schwingshackl 2021, Bröckelmann 2022 found that pooled ratios of ratios were similar in RCTs and observational studies, with considerable statistical heterogeneity. Also, in continuous outcomes, the difference between observational studies and RCTs was small.

We excluded reviews (and overviews) that were unsystematic in building the study sample for their analysis. This means that we excluded reviews that did report a reconstructable method of searching and selecting studies. We did this in order to limit the evidence base for this overview to systematic comparisons of study types and to exclude reviews that illustrate selected topics where findings of RCTs and observational studies were clearly consistent or inconsistent. Such striking examples of (in-)consistency would have introduced a risk of bias to our evidence base.

AUTHORS' CONCLUSIONS

Implication for systematic reviews and evaluations of healthcare

Our findings show that it is important for review authors to consider not only study type (e.g. observational studies or randomised controlled trials (RCTs)), but also the level of methodological, clinical, and statistical heterogeneity within the meta-analyses of each study type.

A better understanding of how study types and their characteristics influence or bear on effect estimates can potentially yield effect estimates more reflective of true effectiveness. For example, the duration of follow-up is often longer in observational studies than it is in trials. This might be a reason why more events might be observed in observational studies.

When assessing the validity of effect estimates, it is more important to critically appraise the specific methods of a study, assess its risk of bias and the certainty of the body of evidence than to simply assess the validity of an effect estimate based on whether the underlying body of evidence is from RCTs or observational studies. Within the GRADE approach, the integration of evidence from observational studies might be warranted in specific situations (Cuello-Garcia 2022). When considering evidence from non-randomised studies, the ROBINS-I tool ('Risk Of Bias In Non-randomised Studies of Interventions') should be used to assess these studies' risk of bias (Sterne 2016). The tool allows for a thorough assessment of potential sources of bias in observational studies before the intervention, at the time of the intervention, and after the intervention.

Overall, it is important to consider the most appropriate or feasible design for a particular research question. All these considerations, not just the study type, must be taken into account by researchers conducting comparative effectiveness research.

Implication for methodological research

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 most common bias in an observational study compared to an RCT, and methods for accounting for confounding in meta-analyses of observational studies should be applied and evaluated (Reeves 2013).

Researchers have employed different methods for analysing and comparing bodies of evidence from RCTs and observational studies (Sterne 2018). In order to understand how and why effect estimates differ between RCTs and observational studies, developing a consensus about the methods for such research would be beneficial. In this review, we have developed and used methods (including statistical analysis methods, a risk of bias assessment tool, and an approach to assessing evidence certainty) that may serve as a basis for further methodological development for systematically comparing effect estimates.

Additionally, the validity and feasibility of newly emerging study designs within RCTs and observational studies should be considered in future assessments. For example, studies based on routinely collected data or real-world data, or RCTs with nested cohort studies could represent feasible alternatives to classical RCTs or cohort studies. Future assessments of the design features of such studies, possible design adaptations as well as their validity and feasibility, and the comparability of their effect estimates with effect estimates from RCTs are relevant for understanding which study designs are credible alternatives to RCTs.

ACKNOWLEDGEMENTS

We appreciate important discussions with Joerg J. Meerpohl.

Maria Petropoulou critically commented on a review version and prepared data for the visual inspection.

Hacsi T. Horvath conducted the searches and screened all references for the previous version of this review.

Julia Stadelmaier extracted data from ongoing reviews. Jessica Beyerbach supported the data extraction of one overview of reviews; these data were re-extracted in the course of the editorial process of the review.

Markus Toews extracted data from two included reviews and assessed their risk of bias.

Editorial contributions:

Cochrane Methodology Review Group supported the authors in the development of this systematic review. The following people conducted the editorial process for this article:

Sign-off Editor (final editorial decision): Mike Clarke, Queen's University Belfast, UK; Managing Editor (selected peer reviewers, collated peer-reviewer comments, provided editorial guidance to authors, edited the article): Marwah Anas El-Wegoud, Cochrane Central Editorial Service; Editorial Assistant (conducted editorial policy checks and supported editorial team): Lisa Wydrzynski, Cochrane Central Editorial Service; Copy Editor (copy editing and production): Faith Armitage, Cochrane Central Production Service.

Peer-reviewers (provided comments and recommended an editorial decision): Carlos Cuello, McMaster University, GRADE Working Group; Gregg M. Gascon, PhD, CHDA, OhioHealth; and Robin Featherstone, Cochrane Central Editorial Service (search review). Two additional peer reviewers provided peer review but chose not to be publicly acknowledged.

REFERENCES

References to studies included in this review

Allain 2017 *{published data only}*

Allain N, Leven C, Falissard B, Allain JS, Batail JM, Polard E, et al. Manic switches induced by antidepressants: an umbrella review comparing randomized controlled trials and observational studies. *Acta Psychiatrica Scandinavica* 2017;**135**(2):106-16. [DOI: 10.1111/acps.12672]

Ankarfeldt 2017 *{published data only}*

Ankarfeldt MZ, Adalsteinsson E, Groenwold RH, Ali MS, Klungel OH. A systematic literature review on the efficacy-effectiveness gap: comparison of randomized controlled trials and observational studies of glucose-lowering drugs. *Clinical Epidemiology* 2017;**23**(9):41-51. [DOI: 10.2147/CLEP.S121991]

Artus 2014 *{published data only}*

Artus M, van der Windt D, Jordan KP, Croft PR. The clinical course of low back pain: a meta-analysis comparing outcomes in randomised clinical trials (RCTs) and observational studies. *BMC Musculoskeletal Disorders* 2014;**7**(15):68. [DOI: 10.1186/1471-2474-15-68]

Beks 2018 *{published data only}*

Beks RB, Ochen Y, Frima H, Smeeing DP, Van der Meijden O, Timmers T, et al. Operative versus nonoperative treatment of proximal humeral fractures: a systematic review, meta-analysis, and comparison of observational studies and randomized controlled trials. *Journal of Shoulder and Elbow Surgery* 2018;**27**(8):1526-34. [DOI: 10.1016/j.jse.2018.03.009]

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, Savovic J, Harris R, Altman D, Sterne J, Deeks J. The quantification of bias in randomised and non-randomised studies: the BRANDO NRS Study [Poster]. In: 16th Cochrane Colloquium. Freiburg im Breisgau, Germany, 3-7 October 2008.

Bhandari 2004 *{published data only}*

Bhandari M, Tornetta P 3rd, 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.

Borkowska 2018 *{published data only}*

Borkowska K, Taieb V, Germain N, Mejri A, Jakubowska A, Aballea S. Comparison of patient-reported outcomes in randomised controlled trials and in observational studies for overactive bladder: a systematic literature review. *Value in Health* 2018;**21**(Suppl 3):S481.

Bröckelmann 2022 *{published data only}*

Bröckelmann N, Balduzzi S, Harms L, Beyerbach J, Petropoulou M, Kubiak C, et al. Evaluating agreement between bodies of evidence from randomized controlled trials and cohort studies in medical research: a meta-epidemiological study. *BMC Medicine* 2022;**20**(1):174. [DOI: 10.1186/s12916-022-02369-2]

Dahabre 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.

Demissie 1998 *{published data only}*

Demissie K, Mills OF, Rhoads GG. Empirical comparison of the results of randomized controlled trials and case-control studies in evaluating the effectiveness of screening mammography. *Journal of Clinical Epidemiology* 1998;**51**(2):81-91. [DOI: 10.1016/s0895-4356(97)00243-6]

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.

Gu 2020 *{published data only}*

Gu ZC, Wei AH, Zhang C, Wang XH, Zhang L, Shen L, et al. Risk of major gastrointestinal bleeding with new vs conventional oral anticoagulants: a systematic review and meta-analysis. *Clinical Gastroenterology and Hepatology* 2020;**18**(4):792–9. [DOI: 10.1016/j.cgh.2019.05.056]

Guyatt 2000 *{published data only}*

Guyatt GH, DiCenso A, Farewell V, Willan A, Griffith L. Randomized trials versus observational studies in adolescent pregnancy prevention. *Journal of Clinical Epidemiology* 2000;**53**(2):167-74. [DOI: 10.1016/s0895-4356(99)00160-2]

Hong 2021 *{published data only}*

* Hong YD, Jansen JP, Guerino J, Berger ML, Crown W, Goettsch WG, et al. Comparative effectiveness and safety of pharmaceuticals assessed in observational studies compared with randomized controlled trials. *BMC Medicine* 2021;**19**(1):307.

Hong YD, Mullins CD, Orsini L, Guerino J, Willke R, Berger M, et al. Comparative effectiveness and safety of pharmaceuticals

assessed in observational studies compared with randomized controlled trials. *Value in Health* 2020;**23**:S521.

Hoshino 2021a *{published data only}*

Hoshino N, Fukui Y, Hida K, Obama K. Similarities and differences between study designs in short- and long-term outcomes of laparoscopic versus open low anterior resection for rectal cancer: a systematic review and meta-analysis of randomized, case-matched, and cohort studies. *Annals of Gastroenterological Surgery* 2021;**5**(2):183-93. [DOI: 10.1002/ags3.12409]

Hoshino 2021b *{published data only}*

Hoshino N, Sakamoto T, Hida K, Takahashi Y, Okada H, Obama K, et al. Difference in surgical outcomes of rectal cancer by study design: meta-analyses of randomized clinical trials, case-matched studies, and cohort studies. *BJS Open* 2021;**5**(2):zraa067. [DOI: 10.1093/bjsopen/zraa067]

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.

Jainaud 2021 *{published data only}*

Janiaud P, Agarwal A, Tzoulaki I, Theodoratou E, Tsilidis KK, Evangelou E, et al. Validity of observational evidence on putative risk and protective factors: appraisal of 3744 meta-analyses on 57 topics. *BMC Medicine* 2021;**19**(1):1-17. [DOI: 10.1186/s12916-021-02020-6]

Kimachi 2021 *{published data only}*

Kimachi M, Onishi A, Tajika A, Kimachi K, Furukawa TA. Systematic differences in effect estimates between observational studies and randomized control trials in meta-analyses in nephrology. *Scientific Reports* 2021;**11**(1):10. [DOI: 10.1038/s41598-021-85519-5]

Kirson 2013 *{published data only}*

Kirson NY, Weiden PJ, Yermakov S, Huang W, Samuelson T, Offord SJ, et al. Efficacy and effectiveness of depot versus oral antipsychotics in schizophrenia: synthesizing results across different research designs. *Journal of Clinical Psychiatry* 2013;**74**(6):568-75. [DOI: 10.4088/JCP.12r08167]

Kitsios 2015 *{published data only}*

Kitsios GD, Dahabreh IJ, Callahan S, Paulus JK, Campagna AC, Dargin JM. Can we trust observational studies using propensity scores in the critical care literature? A systematic comparison with randomized clinical trials. *Critical Care Medicine* 2015;**43**(9):1870-9. [10.1097/CCM.0000000000001135]

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.

Li 2016 *{published data only}*

CRD42015025940. Comparison of treatment effect estimates of non-vitamin K antagonist oral anticoagulants versus warfarin in non-valvular atrial fibrillation between observational studies using propensity score methods and randomized controlled trials: a systematic review. www.crd.york.ac.uk/prospero/display_record.php?RecordID=25940 (first registered 03 September 2015).

Li GW, Holbrook A, Jin YL, Zhang YH, Levine MA, Mbuagbaw L, et al. Comparison of treatment effect estimates of non-vitamin K antagonist oral anticoagulants versus warfarin between observational studies using propensity score methods and randomized controlled trials. *European Journal of Epidemiology* 2016;**316**(6):451-61. [DOI: 10.1007/s10654-016-0178-y]

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.

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.

Mathes 2021 *{published data only}*

Mathes T, Rombey T, Kuss O, Pieper D. No inexplicable disagreements between real-world data-based non-randomized controlled studies and randomized controlled trials were found. *Journal of Clinical Epidemiology* 2021;**133**:1-13. [DOI: 10.1016/j.jclinepi.2020.12.019]

Moneer 2022 *{published data only}*

Moneer O, Daly G, Skydel JJ, Nyhan K, Lurie P, Ross JS, et al. Agreement of treatment effects from observational studies and randomized controlled trials evaluating hydroxychloroquine, lopinavir-ritonavir, or dexamethasone for COVID-19: meta-epidemiological study. *BMJ* 2022;**377**:e069400. [DOI: 10.1136/bmj-2021-069400]

Morfaw 2021 *{published data only}*

Morfaw F, Miregwa B, Bi A, Mbuagbaw L, Anderson LN, Thabane L. Comparing and combining evidence of treatment effects in randomized and nonrandomized studies on the use of misoprostol to prevent postpartum hemorrhage. *Journal of Evidence-Based Medicine* 2021;**14**(3):198-207. [DOI: 10.1111/jebm.12440]

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.

Otsuka 2022 *{published data only}*

Otsuka R, Hayashi H, Uesato M, Hayano K, Murakami K, Kano M, et al. Comparison of estimated treatment effects between randomized controlled trials, `case-matched`, and cohort studies on laparoscopic versus open distal gastrectomy for advanced gastric cancer: a systematic review and `meta-analysis`. *Langenbecks Archives of Surgery* 2022;**3**:1-17. [DOI: 10.1007/s00423-022-02454-3]

Papanikolaou 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.

Safieddine 2021 *{published data only}*

Safieddine M, Chapelle C, Ollier E, Ferdynus C, Bertoletti L, Mismetti P, et al. Compared to randomized studies, observational studies may overestimate the effectiveness of DOACs: a meta-epidemiological approach. *Journal of Clinical Epidemiology* 2021;**130**:49-58. [DOI: https://doi.org/10.1016/j.jclinepi.2020.10.013]

Schmidt 2013 *{published data only}*

Schmidt AF, Rovers MM, Klungel OH, Hoes AW, Knol MJ, Nielen M, et al. Differences in interaction and subgroup-specific effects were observed between randomized and nonrandomized studies in three empirical examples. *Journal of Clinical Epidemiology* 2013;**66**(6):599-607. [DOI: 10.1016/j.jclinepi.2012.08.008]

Schwingshackl 2021 *{published data only}*

Schwingshackl L, Balduzzi S, Beyerbach J, Bröckelmann N, Werner SS, Zähringer J, et al. Evaluating agreement between bodies of evidence from randomised controlled trials and cohort studies in nutrition research: meta-epidemiological study. *BMJ* 2021;**374**:n1864. [DOI: 10.1136/bmj.n1864]

Shen 2020 *{published data only}*

Shen N-N, Wu Y, Wang N, Kong L-C, Zhang C, Wang J-L, et al. Direct oral anticoagulants vs. vitamin-K antagonists in the elderly with atrial fibrillation: a systematic review comparing benefits and harms between observational studies and randomized controlled trials. *Frontiers in Cardiovascular Medicine* 2020;**10**(7):132. [DOI: 10.3389/fcvm.2020.00132]

Tan 2017 *{published data only}*

Tan TH, Soon YY, Cheo T, Tey JCS. Inferring the effects of concurrent chemotherapy with or without adjuvant chemotherapy versus radiotherapy alone in stage II-IVB nasopharyngeal carcinoma: comparison of results from meta-analysis of observational studies and randomized trials. *Journal of Clinical Oncology* 2017;**35**(15 Suppl):e17573. [DOI: 10.1200/JCO.2017.35.15_suppl.e17573]

Tzoulaki 2011 *{published data only}*

Tzoulaki I, Siontis KK, Ioannidis JP. Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. *BMJ* 2011;**343**:1-11. [DOI: 10.1136/bmj.d6829]

Van de Wall 2020 *{published data only}*

Van de Wall BJ, Ochen Y, Beeres FJ, Babst R, Link BC, Heng M, et al. Conservative vs. operative treatment for humeral shaft fractures: a meta-analysis and systematic review of randomized clinical trials and observational studies. *Journal of Shoulder and Elbow Surgery* 2020;**29**(7):1493-504. [DOI: 10.1016/j.jse.2020.01.072]

Van Heesewijk 2018 *{published data only}*

Van Heesewijk AE, Rush ML, Schmidt B, Kirpalani H, DeMauro SB. Agreement between study designs: a systematic review comparing observational studies and randomized trials of surgical treatments for necrotizing enterocolitis. *Journal of Maternal-fetal & Neonatal Medicine* 2020;**33**(12):1-9. [DOI: 10.1080/14767058.2018.1533948]

Virk 2019 *{published data only}*

Virk SA, Ariyaratnam J, Bennett RG, Kumar S. Updated systematic review and meta-analysis of the impact of contact force sensing on the safety and efficacy of atrial fibrillation ablation: discrepancy between observational studies and randomized control trial data. *EP Europace* 2019;**21**(2):239-49. [DOI: 10.1093/europace/euy266]

Yanik 2013 *{published data only}*

Yanik E, Brookhart A, Sturmer T, Poole C. Comparison and meta-analysis of results from randomized and non-randomized studies in epidemiologic research. *American Journal of Epidemiology* 2013;**177**(Suppl 11):S135. [DOI: 10.1093/aje/kwt103]

Youn 2021 *{published data only}*

Youn S, Wong SA, Chrystoja C, Tomlinson G, Wijeysundera HC, Bell CM, et al. Bias estimation in study design: a meta-epidemiological analysis of transcatheter versus surgical aortic valve replacement. *BMC Surgery* 2021;**21**(285):1-11. [DOI: 10.1186/s12893-021-01278-0]

Ziff 2015 *{published data only}*

Ziff OJ, Lane DA, Samra M, Griffith M, Kirchhof P, Lip GY, et al. Safety and efficacy of digoxin: systematic review and meta-analysis of observational and controlled trial data. *BMJ* 2015;**351**(h4451):1-9. [DOI: 10.1136/bmj.h4451]

References to studies excluded from this review

Abraham 2010 *{published data only}*

Abraham NS, Byrne CJ, Young JM, Solomon MJ. Meta-analysis of well-designed nonrandomized comparative studies of surgical procedures is as good as randomized controlled trials. *Journal of Clinical Epidemiology* 2010;**63**(3):238-45.

Ahren 2014 *{published data only}*

Ahren B, Mathieu C, Bader G, Schweizer A, Foley JE. Efficacy of vildagliptin versus sulfonylureas as add-on therapy to metformin: comparison of results from randomised controlled and observational studies. *Diabetologia* 2014;**57**(7):1304-7.

Algra 2012 *{published data only}*

Algra AM, Rothwell PM. Effects of regular aspirin on long-term cancer incidence and metastasis: a systematic comparison of evidence from observational studies versus randomised trials. *Lancet Oncology* 2012;**13**:518-27.

Anderson 2019 *{published data only}*

Anderson R D, Ariyarathna N, Lee G, Virk S, Trivic I, Campbell T, et al. Catheter Ablation versus Medical Therapy for Treatment of Ventricular Tachycardia associated with Structural Heart Disease: Systematic Review and Meta-analysis of Randomised Controlled Trials and Comparison with Observational Studies. *Heart rhythm* 2019;**16**(10):1484-91.

Arditi 2016 *{published data only}*

Arditi C, Burnand B, Peytremann-Bridevaux I. Adding non-randomised studies to a Cochrane review brings complementary information for healthcare stakeholders: an augmented systematic review and meta-analysis. *BMC Health Services Research* 2016;**16**:598. [DOI: 10.1186/s12913-016-1816-5]

Atar 2015 *{published data only}*

Atar D, Ong S, Lansberg P J. Expanding the Evidence Base: Comparing Randomized Controlled Trials and Observational Studies of Statins. *American journal of therapeutics* 2015;**22**(5):e141-50.

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.

Beyerbach 2022 *{published and unpublished data}***10.1093/advances/nmab095**

Beyerbach J, Stadelmaier J, Hoffmann G, Balduzzi S, Brockelmann N, Schwingshackl L. Evaluating Concordance of Bodies of Evidence from Randomized Controlled Trials, Dietary Intake, and Biomarkers of Intake in Cohort Studies: A Meta-Epidemiological Study. *Advances in Nutrition* 2022;**13**(1):48-65.

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.

Collins 2012 *{published data only}*

Collins GS, Le Manach Y. Comparing treatment effects between propensity scores and randomized controlled trials: improving conduct and reporting. *European heart journal* 2012;**33**(15):1867-9.

Concato 2000a *{published data only}*

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

Concato 2000b *{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.

Coscia 2019 *{published data only}*

Coscia C, Jaureguizar A, Quezada C A, Muriel A, Monreal M, Villen T, et al. Comparison of All-Cause Mortality Following VTE Treatment Between Propensity Score-Adjusted Observational Studies and Matched Randomized Controlled Trials Meta-Epidemiologic Study. *Chest* 2019;**155**(4):689-98.

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.

CRD42017058116 *{published data only}*

Efficacy and safety of tranexamic acid in arthroplasty: Bayesian network mixed comparison meta-analysis of randomized controlled trials and cohort studies. https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=58116 (last retrieved 19 October 2021).

Efficacy and safety of tranexamic acid in arthroplasty: Bayesian network mixed comparison meta-analysis of randomized controlled trials and cohort studies. https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=58116.

CRD42018104452 *{published data only}*

Retention in Opioid Agonist Treatment: A Systematic Review and Meta-analysis comparing Observational Studies and Randomized Controlled Trials. https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=104452 (last retrieved on 19 October 2021).

CRD42019130585 *{published data only}*

The effectiveness of proton pump inhibitors versus histamine-2-receptor antagonists for stress ulcer prophylaxis in critically ill patients: comparison of randomized trials with cohort studies. https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=130585 (last retrieved on 19 October 2021).

da Silva 2017 *{published data only}*

da Silva SG, Ricardo LI, Evenson KR, Hallal PC. Leisure-Time Physical Activity in Pregnancy and Maternal-Child Health: A Systematic Review and Meta-Analysis of Randomized Controlled Trials and Cohort Studies. *Sports medicine* 2017;**47**(2):295-317.

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. In: 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.

El-Hayek 2014 *{published data only}*

El-Hayek G, Benjo A, Uretsky S, Al-Mallah M, Cohen R, Bamira D, et al. Meta-analysis of coronary computed tomography angiography versus standard of care strategy for the evaluation of low risk chest pain: are randomized controlled trials and cohort studies showing the same evidence? *International Journal of Cardiology* 2014;**177**(1):238-45.

Ewald 2020 *{published data only}*

Ewald H, Ioannidis JPA, Ladanie A, Mc Cord K, Bucher HC, Hemkens LG. Nonrandomized studies using causal-modeling may give different answers than RCTs: a meta-epidemiological study. *Journal of Clinical Epidemiology* 2020;**118**:29-41. [DOI: 10.1016/j.jclinepi.2019.10.012]

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.

Fukuta 2017 *{published data only}*

Fukuta H, Goto T, Wakami K, Ohte N. Effect of renin-angiotensin system inhibitors on mortality in heart failure with preserved ejection fraction: a meta-analysis of observational cohort and randomized controlled studies. *Heart failure reviews* 2017;**22**(6):775-82.

Furlan 2008a *{published data only}*

Furlan AD, Tomlinson G, R Jadad A, Bombardier C. Methodological quality and homogeneity influenced agreement between randomized trials and nonrandomized studies of the same intervention for back pain. *Journal of Clinical Epidemiology* 2008;**61**(3):209-31.

Gray 2017 *{published data only}*

Gray E, Norris S, Schmitz S, O'Leary A. Do disparities between populations in randomized controlled trials and the real world lead to differences in outcomes? *Journal of Comparative Effectiveness Research* 2017;**6**(1):65-82. [DOI: 10.2217/cer-2016-0042]

Gyawali 2020 *{published data only}*

Gyawali B, D'Andrea E, Franklin J M, Kesselheim A S. Response Rates and Durations of Response for Biomarker-Based Cancer Drugs in Nonrandomized Versus Randomized Trials. *Journal of the National Comprehensive Cancer Network* 2020;**18**(1):36-43.

Hallstrom 2000 *{published data only}*

Hallstrom A, Anderson Jl, 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. In: 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.

Hundscheid 2021 *{published data only}*

Hundscheid T, Jansen E J S, Onland W, Kooi E M W, Andriessen P, de Boode W P. Conservative Management of Patent Ductus Arteriosus in Preterm Infants-A Systematic Review and Meta-Analyses Assessing Differences in Outcome Measures Between Randomized Controlled Trials and Cohort Studies. *Frontiers in pediatrics* 2021;**9**:15.

Ioannidis 2005 *{published data only}*

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

Jee 2016 *{published data only}*

Jee YM, Bak JS, Weinlander E, Anderson PA. Comparing Nonrandomized Observational Studies With Randomized

Controlled Trials in Cervical Disc Arthroplasty A Meta-analysis. *Spine* 2016;**41**(5):419-28.

Khan 2019 *{published data only}*

Khan SU, Lone AN, Asad ZU, Rahman H, Khan MS, Saleem MA, et al. Meta-Analysis of Efficacy and Safety of Proton Pump Inhibitors with Dual Antiplatelet Therapy for Coronary Artery Disease. *Cardiovascular Revascularization Medicine* 2019;**20**(12):1125-33. [DOI: 10.1016/j.carrev.2019.02.002]

Kilcher 2018 *{published data only}*

Kilcher G, Hummel N, Didden EM, Egger M, Reichenbach S. Rheumatoid arthritis patients treated in trial and real world settings: comparison of randomized trials with registries. *Rheumatology* 2018;**57**(2):354-69. [DOI: 10.1093/rheumatology/kex394]

Kim 2020 *{published data only}*

Kim JY, Steingroever J, Lee KH, Oh J, Choi MJ, Lee J et al. Clinical Interventions and All-Cause Mortality of Patients with Chronic Kidney Disease: An Umbrella Systematic Review of Meta-Analyses. *Journal of Clinical Medicine* 2020;**9**(2):394. [DOI: 10.3390/jcm9020394]

Kirk 2004 *{published data only}*

Kirk O. Effects of highly active antiretroviral therapy among HIV-infected patients - Results from randomised and observational studies. *Danish medical bulletin* 2004;**51**(1):63-81.

Kishimoto 2021 *{published data only}*

Kishimoto T, Hagi K, Kurokawa S, Kane J M, Correll C U. Long-acting injectable versus oral antipsychotics for the maintenance treatment of schizophrenia: a systematic review and comparative meta-analysis of randomised, cohort, and pre-post studies. *Lancet Psychiatry* 2021;**8**(5):387-404.

Kitsios 2015a *{published data only}*

Kitsios Georgios D. Propensity score studies are unlikely to underestimate treatment effects in critical care medicine: a critical reanalysis. *Journal of Clinical Epidemiology* 2015;**68**(4):467-9.

Krogh 2021 *{published data only}*

Krogh J, Hjorthoj C, Kristensen SL, Selmer C, Haugaard SB. The effect of sodium-glucose transport protein 2 inhibitors on mortality and heart failure in randomized trials versus observational studies. *Diabetic Medicine* 2021;**38**(9):e14600. [DOI: 10.1111/dme.14600]

Kunz 1998 *{published data only}*

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

Kuss 2020 *{published data only}*

Kuss O, Miller M. Unknown confounders did not bias the treatment effect when improving balance of known confounders in randomized trials. *Journal of Clinical Epidemiology* 2020;**126**:9-16.

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.

Lai 2010 *{published data only}*

Lai JK, Lucas RM, Clements MS, Roddam AW, Banks E. Hip fracture risk in relation to vitamin D supplementation and serum 25-hydroxyvitamin D levels: a systematic review and meta-analysis of randomised controlled trials and observational studies. *BMC Public Health* 2010;**10**(1):331.

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.

Leichsenring 2008 *{published data only}*

Leichsenring F, Rabung S. Effectiveness of long-term psychodynamic psychotherapy - A meta-analysis. *Journal of the American Medical Association* 2008;**300**(13):1551-65.

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.

Liu 2017 *{published data only}*

Liu HY, Han Y, Chen XS, Bai L, Guo SP, Li L, et al. Comparison of efficacy of treatments for early syphilis: A systematic review and network meta-analysis of randomized controlled trials and observational studies. *PLoS One* 2017;**12**(6):15.

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.

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. *British Medical Journal* 1999;**319**:312-5.

Mehyar 2021 *{published data only}*

Mehyar L, Sandler J, Thiruvenkatachari B. Does observational study on the effectiveness of the Twin Blocks overestimate or underestimate the results? A comparative analysis of retrospective samples versus randomized controlled trial. *Journal of the World federation of orthodontists* 2021;**10**(2):43-8.

Melloni 2015 *{published data only}*

Melloni C, Washam JB, Jones WS, Halim SA, Hasselblad V, Mayer SB, et al. Conflicting Results Between Randomized Trials and Observational Studies on the Impact of Proton Pump Inhibitors on Cardiovascular Events When Coadministered With Dual Antiplatelet Therapy Systematic Review. *Circulation* 2015;**8**(1):47-U91.

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.

Morgan 2014 *{published data only}*

Morgan JL, Reed MW, Wyld L. Primary endocrine therapy as a treatment for older women with operable breast cancer - a comparison of randomised controlled trial and cohort study findings. *European journal of surgical oncology* 2014;**40**(6):676-84.

Moyer 2002 *{published data only}*

Moyer A, Finney JW. Randomized versus nonrandomized studies of alcohol treatment: Participants, methodological features and posttreatment functioning. *Journal of studies on alcohol* 2002;**63**(5):542-50.

Mugavero 2011 *{published data only}*

Mugavero MJ, May M, Ribaudo HJ, Gulick RM, Riddler SA, Haubrich R et al. Comparative Effectiveness of Initial Antiretroviral Therapy Regimens: ACTG 5095 and 5142 Clinical Trials Relative to ART-CC Cohort Study. *Journal of acquired immune deficiency syndromes* 2011;**58**(3):253-60.

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.

Nigwekar 2009 *{published data only}*

Nigwekar SU, Kandula P, Hix JK, Thakar CV. Off-Pump Coronary Artery Bypass Surgery and Acute Kidney Injury: A Meta-analysis of Randomized and Observational Studies. *American journal of kidney diseases* 2009;**54**(3):413-23.

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.

NN 2014 *{published data only}*

NA. Mammographic breast cancer screening. *Rev Prescrire* 2014;**34**(373):842-846.

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.

Ottenbacker 1992 *{published data only}*

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

Papageorgiou 2015 *{published data only}*

Papageorgiou SN, Xavier GM, Cobourne MT. Basic study design influences the results of orthodontic clinical investigations. *Journal of Clinical Epidemiology* 2015;**68**(12):1512-22.

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.

Pasala 2016 *{published data only}*

Pasala T, Gajulapalli RD, Bolen S, Bajaj NS, Gandhi S, Tandar A et al. Transradial access mitigates bleeding benefit offered by bivalirudin over heparin in patients undergoing percutaneous coronary intervention: Insights from meta-analysis of randomized and observational studies. *International Journal of Cardiology* 2016;**221**:601-8.

Peinemann 2013 *{published data only}*

Peinemann F, Tushabe DA, Kleijnen J. Using Multiple Types of Studies in Systematic Reviews of Health Care Interventions - A Systematic Review. *PloS One* 2013;**8**(12):e85035.

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.

Podmore 2021 *{published data only}*

Podmore B, Qizilbash N, Lacetera A, Ubillos I, Andresen K, Martín AR et al. Tocilizumab and mortality in hospitalised patients with covid-19. A systematic review comparing randomised trials with observational studies. medRxiv 2021.

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.

Rivera-Caravaca 2018 *{published data only}*

Rivera-Caravaca JM, Esteve-Pastor MA, Marin F, Valdes M, Vicente V, Roldan V et al. A Propensity Score Matched Comparison of Clinical Outcomes in Atrial Fibrillation Patients Taking Vitamin K Antagonists: Comparing the "Real-World" vs Clinical Trials. *Mayo Clinic proceedings* 2018;**93**(8):1065-73.

Rompen 2021 *{published data only}*

Rompen IF, Knobe M, Link B-C, Beeres FJP, Baumgaertner R, Diwersi N et al. Cement augmentation for trochanteric femur fractures: A meta-analysis of randomized clinical trials and observational studies. *PLoS One* 2021;**16**(6):e0251894.

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. In: 14th Cochrane Colloquium,. Dublin, Ireland, 23-26 October, 2006.

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.

Sison 2018 *{published data only}*

Sison J, Vega RMR, Dayi H, Bader G, Brunel P. Efficacy and effectiveness of valsartan/amlodipine and valsartan/amlodipine/hydrochlorothiazide in hypertension: randomized controlled versus observational studies. *Current medical research and opinion* 2018;**34**(3):501-15.

Smeeing 2017 *{published data only}*

Smeeing DPJ, van der Ven DJC, Hietbrink F, Timmers TK, van Heijl M, Kruyt MC et al. Surgical Versus Nonsurgical Treatment for Midshaft Clavicle Fractures in Patients Aged 16 Years and Older: A Systematic Review, Meta-analysis, and Comparison of Randomized Controlled Trials and Observational Studies. *The American journal of sports medicine* 2017;**45**(8):1937-45.

Soni 2019 *{published data only}*

Soni PD, Hartman HE, Dess RT, Abugharib A, Allen SG, Feng FY et al. Comparison of Population-Based Observational Studies With Randomized Trials in Oncology. *Journal of Clinical Oncology* 2019;**37**(14):1209-16.

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.

Stuart 2017 *{published data only}*

Stuart B L, Grebel L E N, Butler C C, Hood K, Verheij T J M, Little P. Comparison between treatment effects in a randomised controlled trial and an observational study using propensity scores in primary care. *British Journal of General Practice* 2017;**67**(662):E643-9.

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. *Journal of the American Medical Association* 2007;**297**(3):278-85.

Sun 2007 *{published data only}*

Sun P, Buesching D, Obenchain R. Treatment effect estimation: A comparison between randomized controlled trials and observational study with three different causal effect methods-propensity score, marginal structure model and doubly robust methods. *Value in health* 2007;**10**(3):A17.

Syn 2020 *{published data only}*

Syn NL, Kabir T, Koh YX, Tan HL, Wang LZ, Chin BZ et al. Survival Advantage of Laparoscopic Versus Open Resection For Colorectal Liver Metastases A Meta-analysis of Individual Patient Data From Randomized Trials and Propensity-score Matched Studies. *Annals of Surgery* 2020;**272**(2):253-65.

Theodoratou 2014 *{published data only}*

Theodoratou E, Tzoulaki I, Zgaga L, Ioannidis JPA. Vitamin D and multiple health outcomes: umbrella review of systematic reviews and meta-analyses of observational studies and randomised trials. *British Medical Journal* 2014;**348**:g2035. [DOI: 10.1136/bmj.g2035]

Trikalinos 2012 *{published data only}*

Trikalinos TA, Moorthy D, Chung M, Yu WW, Lee J, Lichtenstein AH et al. Concordance of randomized and nonrandomized studies was unrelated to translational patterns of two nutrient-disease associations. *Journal of Clinical Epidemiology* 2012;**65**(1):16-29.

Vigil-De Gracia 2020 *{published data only}*

Vigil-De Gracia P, Ludmir J. Conservative management of early-onset severe preeclampsia: comparison between randomized and observational studies a systematic review. *The journal of maternal-fetal and neonatal medicine* 2020;**10**:1-8.

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.

Wilkes 2010 *{published data only}*

Wilkes MM, Navickis RJ, Chan WW, Lewiecki EM. Bisphosphonates and osteoporotic fractures: a cross-design synthesis of results among compliant/persistent postmenopausal women in clinical practice versus randomized controlled trials. *Osteoporos Interational* 2010;**21**(4):679-88.

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.

Yank 2009 *{published data only}*

Yank V, Logan A C, Vaughan Tuohy C, Bravata D M, Staudenmayer K, Eisenhut R, et al. Comparison of thromboembolic event rates in randomized controlled trials and observational studies of recombinant factor VIIA for off-label indications. *Annals of Internal Medicine* 2009;**154**(8):529-40.

Zhang 2014a *{published data only}*

Zhang ZH, Ni HY, Xu X. Do the observational studies using propensity score analysis agree with randomized controlled trials in the area of sepsis? *Journal of Critical Care* 2014;**29**(5):7. [DOI: 10.1016/j.jcrc.2014.05.023]

Zhang 2014b *{published data only}*

Zhang Z, Ni H, Xu X. Observational studies using propensity score analysis underestimated the effect sizes in critical care medicine. *Journal of Clinical Epidemiology* 2014;**67**(8):932-9.

References to ongoing studies

CRD42014013478 *{published data only}*

CRD42014013478. Comparison of the associations of specific foods on body weight between RCTs and observational epidemiologic studies. www.crd.york.ac.uk/prospero/display_record.php?RecordID=13478 (first registered 02 September 2014).

CRD42017059665 *{published data only}*

CRD42017059665. Comparison of cohort and controlled studies of suicide risk categorisation: a meta-analysis. www.crd.york.ac.uk/prospero/display_record.php?RecordID=59665 (first registered 18 March 2017).

CRD42017079569 *{published data only}*

CRD42017079569. Examining the treatment effect of antibiotic use to treat acute respiratory tract infections in primary care: a systematic review comparing the outcomes in randomised control trials (RCTs) and observational studies. www.crd.york.ac.uk/prospero/display_record.php?RecordID=79569 (first registered 25 October 2017).

CRD42018062204 *{published data only}*

CRD42018062204. Comparison of treatment effects in randomised vs. non-randomised studies and the role of analytical methods to control for confounding: a meta-epidemiological study. www.crd.york.ac.uk/prospero/display_record.php?RecordID=62204 (first registered 18 October 2018).

Additional references

Altman 2003

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

Bundesamt für Justiz 2020

Bundesamt für Justiz. Regulation on the procedural principles for the assessment of examination and treatment methods in statutory health insurance-accredited medical care and in hospitals; 2020 Rule Number: BGBl. I S. 1379 [Verordnung über die Verfahrensgrundsätze der Bewertung von Untersuchungs- und Behandlungsmethoden in der vertragsärztlichen Versorgung und im Krankenhaus (Methodenbewertungsverfahrensverordnung - MBVerfV); Methodenbewertungsverfahrensverordnung vom 23. Juni 2020 (BGBl. I S. 1379)]. Gesetze im Internet 2020.

Campbell 2019

Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ* 2019;**368**:l6890. [DOI: 10.1136/bmj.l6890]

Covidence 2022 [Computer program]

Covidence systematic review software. Melbourne, Australia: Veritas Health Innovation, 2022. Available at www.covidence.org.

Cuello-Garcia 2022

Cuello-Garcia CA, Santesso N, Morgan RL, Verbeek J, Thayer K, Ansari MT, et al. GRADE guidance 24 optimizing the integration of randomized and non-randomized studies of interventions in evidence syntheses and health guidelines. *Journal of Clinical Epidemiology* 2022;**142**:200-8. [DOI: 10.1016/j.jclinepi.2021.11.026]

Desai 2019

Desai, RJ, Franklin, JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. *BMJ* 2019;**367**:l5657. [DOI: 10.1136/bmj.l5657]

Franklin 2017

Franklin JM, Dejene S, Huybrechts KF, Wang SV, Kulldorff M, Rothman KJ. A bias in the evaluation of bias comparing randomized trials with nonexperimental studies. *Epidemiological Methods* 2017;**6**(1):20160018. [DOI: 10.1515/em-2016-0018]

Franklin 2021

Franklin JM, Patorno E, Desai RJ, Glynn RJ, Martin D, Quinto K, et al. Emulating randomized clinical trial with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE Initiative. *Circulation* 2021;**143**(10):1002-13. [DOI: 10.1161/CIRCULATIONAHA.120.051718]

Hemkens 2016

Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. *BMJ* 2016;**352**:i493. [DOI: 10.1136/bmj.i493]

Higgins 2005

Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;**327**(7414):557-60.

Higgins 2021

Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021). Cochrane, 2021. Available from www.training.cochrane.org/handbook.

Institute of Medicine 2009

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

Kamerow 2011

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

Morshed 2009

Morshed S, Tornetta P 3rd, Bhandari M. Analysis of observational studies: a guide to understanding statistical methods. *Journal of Bone and Joint Surgery. American Volume* 2009;**91**(Suppl 3):50-60. [DOI: 10.2106/JBJS.H.01577. PMID: 19411500]

Murad 2017

Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. *BMJ Evidence-Based Medicine* 2017;**22**:139-42. [DOI: 10.1136/ebmed-2017-110713]

Neilson 2019

Neilson C, Lê ML. A failed attempt at developing a search filter for systematic review methodology articles in Ovid Embase. *Journal of the Medical Library Association* 2019;**107**(2):203-9. [DOI: 10.5195/jmla.2019.519]

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. Art. No: MR000012. [DOI: 10.1002/14651858.MR000012.pub3]

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.

Riley 2011

Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011;**342**:d549.

Schünemann 2020

Schünemann HJ, Higgins JP, Vist GE, Glasziou P, Akl EA, Skoetz N, et al. Chapter 14: Completing 'Summary of findings' tables and grading the certainty of the evidence. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 6.1 (updated September 2020). Cochrane, 2020. Available from www.training.cochrane.org/handbook.

Sera 2019

Sera F, Armstrong B, Blangiardo M, Gasparrini A. An extended mixed-effects framework for meta-analysis. *Statistics in Medicine* 2019;**29**:5429-44.

Sterne 2016

Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;**12**(355):i4919. [DOI: 10.1136/bmj.i4919. PMID: 27733354; PMCID: PMC5062054]

Sterne 2018

Sterne, J. Commentary: does the selective inversion approach demonstrate bias in the results of studies using routinely collected data? *BMJ* 2018;**362**:k3259.

Viechtbauer 2005

Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random effects model. *Journal of Educational and Behavioral Statistics* 2005;**30**:261-93.

References to other published versions of this review

Anglemyer 2014

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]

* Indicates the major publication for the study

CHARACTERISTICS OF STUDIES

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

Allain 2017

Study characteristics	
Methods	Overview of reviews
Data	35 RCTs (n = 6244) and 22 prospective and retrospective cohort studies (n = 76,544) examining manic switches induced by antidepressants. The overview included studies published between 1990 and 2013, found in searches of MEDLINE, the Cochrane Library, Embase, and congress abstracts (not further specified). **Inclusion criteria:** "In this review, we considered RCTs and observational cohorts (longitudinal non-randomized and non-blinded studies) that provided data about rates of manic switch after antidepressant treatment. Only study reports in English, French, and Spanish were included. We focused our attention on antidepressant treatment (selective serotonin reuptake inhibitor (SSRI), serotonin-norepinephrine reuptake inhibitors (SNRI), imipraminics (IP), other classes of antidepressant, all classes). Where appropriate, in RCTs, we also extracted data about placebo arms." **Exclusion criteria:** "Patients treated with monoamine oxidase inhibitors were excluded from this review because these drugs are currently rarely used in clinical practice."
Comparisons	Pooled RR and regression coefficients for manic switches in RCTs and observational studies were extracted
Outcomes	1 outcome of relevance for this review: manic switches
Notes	Allain 2017 did not report sufficient data to be included in the primary meta-analysis. Reported results: RCTs underestimated the rate of the manic switch (RRR 0.53, 95% CI 0.32 to 0.87). Funding: "this work was supported by a local grant from Rennes CHU (CORECT: COmitée de la Recherche Clinique et Translationelle)"

Risk of bias		
Item	**Authors' judgement**	**Support for judgement**
Explicit criteria?	Yes	Quote: "In this review, we considered RCTs and observational cohorts (longitudinal non-randomized and non-blinded studies) that provided data about rates of manic switch after antidepressant treatment. Only study reports in English, French, and Spanish were included. We focused our attention on antidepressant treatment; where appropriate also placebo arms. Patients treated with monoamine oxidative inhibitors were excluded." Comment: clear inclusion and exclusion criteria were reported and applied.

Allain 2017 *(Continued)*

Investigator agreement?	Yes	Quote: "Study selection was performed in two steps. In a first step, meta-analyses, simple, and systematic reviews were identified in a blinded standardized manner by two reviewers (CL, FN). In a second step, all relevant references were extracted (NA, FN)." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	No	Quote: "Our method of extracting studies from meta-analyses and reviews about manic switch does not enable an exhaustive review as commonly performed in meta-analysis. This could have led to a biased sample. But this biased sample is a straightforward reflection of the literature. Our aim was indeed to understand the literature on rates of manic switch from a meta-epidemiological perspective."
Bias assessed?	Yes	Quote: "Each paper was assessed for methodological quality prior to inclusion in the review, using two appropriate standardized critical appraisal instruments (10), one for RCTs and one for observational studies." Comment: methodology and quality of included studies was assessed and reported in table 1.
Control for differences?	Yes	Quote: "Sensitivity analyses taking quality into account showed the robustness of our estimation for study type, antidepressant class, diagnosis, and age class." p109 Comment: RCTs and observational studies were analysed separately.
Heterogeneity addressed?	Yes	Quote: "To adjust our comparison of observational studies and randomized controlled trials on identified sources of heterogeneity, and to quantify the impact of certain variables on the manic switch rate (MSR), we performed a meta-regression." Comment: the review considered the differences in the ages of the participants and the designs of the studies. The reviewers also performed statistical heterogeneity analysis for the studies. Limiting inclusion criteria were applied. There were interaction analyses/ meta-regression analyses taking intervention and patient characteristics into account.
Similar outcomes?	Yes	Quote: "The primary outcome was manic switch prevalence in the different arms of each study." And: "In older studies (21), the psychomotor agitation associated with IP was liable to be labeled as a manic switch, while today the diagnostic criteria are more restrictive. Criteria have changed over time." Comment: outcome measures used for studies were similar. However, the definition of manic switches was reported to have changed over time.
No selective reporting	Yes	Comment: analyses reported in methods section are also reported in results section. Post hoc analyses are clearly labelled.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Ankarfeldt 2017

Study characteristics

Ankarfeldt 2017 *(Continued)*

Methods	Systematic review
Data	11 RCTs (n = 4181) and 7 observational studies (n = 101,603) identified through systematic searches in MEDLINE, Embase, Current Content, and Biosis, as well as the reference lists of identified studies published between 2000 and 2015. **Inclusion criteria:** Studies in "English language [which...] compared either glucagon-like peptide-1 analogs (GLP-1) with insulin or dipeptidyl peptidase-4 inhibitors (DPP-4i) with sulfonylurea, all with change in HbA1c as an outcome. The chosen comparator groups were to compare second-line (DPP-4i and sulfonylurea) and third-line (GLP-1 and insulin) treatments, respectively." **Exclusion criteria:** "Studies comparing DPP-4i with sulfonylurea during Ramadan in Muslim populations (three RCTs and six observational studies) because we did not want to compare across fasting and nonfasting studies and to exclude studies with fast-acting insulin (five RCTs) because we did not want to compare across fast-acting and basal insulins; none of the post hoc exclusion criteria are in conflict with the initial inclusion criteria and they only narrow the inclusion criteria further."
Comparisons	RCTs versus observational studies
Outcomes	MD in RCTs and observational studies in HbA1c
Notes	Ankarfeldt 2017 reported continuous outcomes, therefore, the study did not contribute to the primary meta-analysis. Reported results: "No differences were observed in [...] effect sizes across study designs."; Mean effect sizes were reported to range from −0.43 to 0.91 and from −0.80 to 1.13 in RCTs and observational studies, respectively, comparing GLP-1 with insulin, and from −0.13 to 2.70 and −0.20 to 0.30 in RCTs and observational studies, respectively, comparing DPP-4i and sulfonylurea. Funding: "The work leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115546, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and European Federation of Pharmaceutical Industries and Association (EFPIA) companies in kind contribution. In addition, as a special form of the IMI JU grant, University Medical Center Utrecht received a direct financial contribution from Novo Nordisk A/S to support work on this study. MZA and EA belong to EFPIA member companies in the IMI JU and costs related to their part in the research were carried by the respective company as in kind contribution under the IMI JU scheme."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "fulfilling the following inclusion criteria: published between 1 January, 2000 and 31 January, 2015 in English language and compared either glucagon-like peptide-1 analogs (GLP-1) with insulin or dipeptidyl peptidase-4 inhibitors (DPP-4i) with sulfonyl urea, all with change in HbA1c as an outcome." Comment: clear inclusion and exclusion criteria are reported.
Investigator agreement?	Yes	Quote: "The studies identified through the literature search were screened on title and abstract by two reviewers independently. [...] Full text was read by a single reviewer" Comment: study selection was conducted independently by two or more reviewers.

Ankarfeldt 2017 *(Continued)*

Complete sample?	Yes	Comment: a representative study sample was built from systematic searches and predefined inclusion and exclusion criteria.
Bias assessed?	Yes	Quote: "For the observational studies, additional information was extracted: confounding adjustment, analysis of initiator by having a “wash-out” period, selection bias related to clear and reasonable inclusion criteria or handling of missing data, and information bias related to the assessment of exposure and outcome. Comprehensive methods to assess quality of observational studies, such as, for example, ACROBATE NRSI,23 were not deemed necessary". And: "Generally, the data extraction protocol was based on the Cochrane Handbook". Comment: risk of bias partially assessed.
Control for differences?	No	Comment: no sensitivity analyses by study design reported. RCTs and observational studies were analysed separately. Observational study designs are not further differentiated by cohort studies or case-control studies.
Heterogeneity addressed?	Yes	Quote: "The study populations did not differ across RCTs and observational studies with regard to age, sex ratio, BMI, time since diagnosis of type 2 diabetes mellitus, and baseline HbA1c neither in the studies that compared GLP-1 with insulin nor in the studies that compared DPP-4i with sulfonylureas. Generally, this goes for both means and SDs. One exception is HbA1c among studies of GLP-1 and insulin, where the HbA1c distribution in the observational studies was more heterogeneous than in the RCTs."
Similar outcomes?	Yes	Quote: "[...] all with change in HbA1c as an outcome." Comment: blood-glucose lowering was used as an outcome in all studies.
No selective reporting	Yes	Quote: "Post hoc, it was decided to exclude studies comparing DPP-4i with sulfonylurea during Ramadan in Muslim populations (three RCTs and six observational studies) because we did not want to compare across fasting and non-fasting studies and to exclude studies with fast-acting insulin (five RCTs) because we did not want to compare across fast-acting and basal insulins. Studies investigating mixed insulin (combination of fast-acting and intermediate/long-acting) were included. Post hoc exclusion criteria were applied as we gained knowledge when working on the review. Importantly, none of the post hoc exclusion criteria are in conflict with the initial inclusion criteria and they only narrow the inclusion criteria further." Comment: no selective reporting suspected. Post hoc decisions transparently described.
Absence of evidence of bias from other sources?	Unclear	Comment: possible biases in this review are the following: - Selection bias may also have been a problem in the observational studies because inclusion criteria were only partly clear in the observational studies, and all observational studies either excluded participants with missing information or did not report how missing data were handled. - The limited number of studies in this review may also have affected the findings. - As to the number of available studies, publication bias may also have affected our results. - Characteristics of the study populations and other features of the studies may differ in ways not quantified in the data extraction.

Ankarfeldt 2017 *(Continued)*

	- It is possible that the observational studies were designed to be comparable with the RCTs with regard to, for example, the study population. - If the studies have had similar subgroup analyses across RCTs and observational studies, this could be used to investigate the potential efficacy-effectiveness gap even further.

Artus 2014

Study characteristics	
Methods	Systematic review
Data	70 RCTs (n = 11,363) and 19 observational studies (n = 13,097) investigating the clinical course of low back pain. The review included RCTs and observational studies published up to April 2012. RCTs were identified with searches in Cochrane CENTRAL while cohort studies were searched for in AMED, Embase, MEDLINE, CINAHL, and handsearches of systematic reviews. **Inclusion criteria:** RCTs and prospective observational cohort studies conducted for primary care treatment for low back pain (LBP) (e.g. analgesia, exercises, manipulation therapy) amongst individuals aged 18 or over. Studies had to provide baseline and follow-up data on the designated primary outcome measure of pain intensity, measured on a numerical rating scale (NRS) or visual analogue scale (VAS). Studies published in English. **Exclusion criteria:** Studies conducted amongst patients with specific low back pain (e.g. cancer or inflammatory arthritis), post-operative or post-traumatic back pain, or back pain associated with pregnancy or labour.
Comparisons	Pooled mean differences and standard mean differences were extracted for RCTs and observational studies
Outcomes	1 outcome of relevance for this review: mean pain intensity score at different time points
Notes	Since Artus 2014 reported continuous outcomes, the study did not contribute to the primary meta-analysis. Reported results: low back pain symptoms followed a similar course in RCTs and cohort studies. Funding: "This study was part of a larger research project for the PhD conducted by MA, supervised by DvdW and KPJ. The PhD project was funded through an Arthritis Research UK Primary Care Fellowship, number 17890." Notes: The authors list the 'Hawthorne effect', the 'care effect' or the unique strict adherence to the treatment protocol, i.e. the 'protocol effect' as reasons why participants in trials might experience a larger benefit from interventions than participants in observational studies.

Risk of bias		
Item	**Authors' judgement**	**Support for judgement**
Explicit criteria?	Yes	Quote: "Included were studies (RCTs and prospective observational cohort studies) conducted for primary care treatment for LBP (e.g. analgesia, exercises manipulation therapy) among individuals aged 18 or over. Studies had to provide baseline and follow-up data on the designated primary outcome measure of pain intensity, measured on a Numerical Rating Scale (NRS) or Visual Analogue Scale (VAS). Only studies published in English were included."

Artus 2014 *(Continued)*

		Comment: explicit inclusion and exclusion criteria were reported in the methods section.
Investigator agreement?	Unclear	Quote: "screening of citations/abstracts ad selection of RCTs and cohort studies applying the inclusion criteria was conducted by MA, DVdW & KPJ" Comment: study selection was conducted by three reviewers; there is no information about whether study selection was done independently.
Complete sample?	Yes	Quote: "The Cochrane Central Register of Controlled Trials (CENTRAL) was therefore chosen as a sufficient data source for RCTs. This search was an update (up to April 2012) of a strategy previously used and described elsewhere [4]. For observational studies, a literature search was conducted for the same time period using the databases of AMED, EMBASE, MEDLINE and CINAHL [...]." Comment: both study designs were systematically searched.
Bias assessed?	Unclear	Comment: no information on risk of bias assessment reported.
Control for differences?	Yes	Quote: "Firstly, RCTs as a single group were compared with observational studies. Secondly, RCTs were sub-grouped into efficacy and pragmatic trials, based on whether the trial included a placebo, sham or no treatment, with such trials being grouped as efficacy trials. RCTs that included comparator treatment of usual care or waiting list arms were classified as pragmatic trials. To compare studies groups that are similar with regard to the type of treatment, a separate analysis was conducted to compare cohort studies with RCT arms that received 'usual care'. Each RCT sub-group was compared separately with observational studies". Comment: the authors conducted separate analyses by study aim and split RCTs into efficacy RCTs and pragmatic RCTs. Also, usual care arms in RCTs were analysed separately. Cohort studies were the only observational study design included.
Heterogeneity addressed?	Yes	Quote: "RCTs were sub-grouped into efficacy and pragmatic trials, based on whether the trial included a placebo, sham or no treatment, with such trials being grouped as efficacy trials. Each RCT sub-group was compared separately with observational studies." "They are comparable in terms of age distribution, gender composition and mean baseline pain intensity (Table 3). It appears that compared with observational studies, RCTs included a larger percentage of participants described as having chronic low back pain (57% in RCTs vs 11% in cohorts). However, these figures need to be interpreted with caution as observational studies often included a mixture of patients with acute and chronic back pain (19% in RCTs vs 63% in cohorts)." Comment: there is some heterogeneity amongst participants.
Similar outcomes?	Yes	Quote: "Studies had to provide baseline and follow-up data on the designated primary outcome measure of pain intensity, measured on a Numerical Rating Scale (NRS) or Visual Analogue Scale (VAS)." Comment: low back pain was the primary outcome measure. Other types of back pain were excluded. Limited set of tools for measuring back pain were listed as included in the inclusion criteria. Studies had to provide baseline and follow-up data on the designated primary outcome measure of pain intensity, measured on a numerical rating scale (NRS) or visual analogue scale (VAS). p. 2, left column.
No selective reporting	Yes	No selective reporting suspected. Analyses reported in the methods section are also reported in the results section.

Artus 2014 *(Continued)*

Absence of evidence of bias from other sources?	Yes	Comment: no other bias suspected.

Beks 2018

Study characteristics

Methods	Systematic review
Data	7 RCTs and 15 observational studies (number of participants not reported) examining operative versus nonoperative treatment of proximal humeral fractures. The review included RCTs and observational studies published up to September 2017 in MEDLINE, Embase, CENTRAL, and CINAHL. **Inclusion criteria:** Proximal humeral fracture, operative versus nonoperative treatment, and reporting of functional outcomes, as well as complications. **Exclusion criteria:** Language other than English, Dutch, or German; no availability of full text; inclusion of patients younger than 18 years; letters, meeting proceedings, and case reports; and external osteosynthesis as operative treatment.
Comparisons	Pooled mean differences and relative risks extracted for RCTs and observational studies
Outcomes	4 outcomes of relevance for this review: functional outcome as measured with Constant-Murley score, major reinterventions, nonunion, avascular necrosis
Notes	Reported results: pooled effects of observational studies were similar to those of RCTs; including observational studies led to more generalisable conclusions. Funding: "The authors, their immediate families, and any research foundations with which they are affiliated have not received any financial payments or other benefits from any commercial entity related to the subject of this article." Notes: The duration of follow-up was often reported to be too short in RCTs, but in sensitivity analysis with high quality studies, contrasting result regarding avascular necrosis did not yield, and pooled effects of both study types were similar.

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "The eligibility criteria were proximal humeral fracture, operative versus nonoperative treatment, and reporting of functional outcomes, as well as complications. The exclusion criteria were language other than English, Dutch, or German; no availability of full text; inclusion of patients younger than 18 years; letters, meeting proceedings, and case reports; and external osteosynthesis as operative treatment." Comment: the search syntax is provided in Appendix S1. Both RCTs and observational studies were included.
Investigator agreement?	Unclear	Quote: "Two reviewers (R.B.B. and Y.O.) independently searched the MEDLINE, Embase, CENTRAL (Cochrane Central Register of Controlled Trials), and CINAHL (Cumulative Index to Nursing and Allied Health Literature) databases

Beks 2018 *(Continued)*

		[...]"; "After screening of the titles and abstracts of identified records, studies were independently assessed based on full text." Comment: unclear if titles and abstracts were screened independently by two reviewers.
Complete sample?	Yes	Comment: search strategy, eligibility criteria as well as flow chart reported.
Bias assessed?	Yes	Quote: "Two reviewers (R.B.B. and H.F.) independently assessed the methodologic quality of all included studies with the Methodological Index for Non-Randomized Studies (MINORS).[ref 39] The MINORS is a validated instrument for methodologic quality assessment and clear reporting of observational studies of surgical interventions.[ref 39] Other quality-assessment tools focus on a specific study design, while the MINORS is externally validated on RCTs by comparison with the CONSORT statement, making it a suitable instrument for metaanalyses of different study designs." Comment: study quality was independently assessed with a validated instrument by two reviewers.
Control for differences?	No	Quote: "Several sensitivity analyses were performed for study quality, year of publication, osteosynthesis by (locking) plate fixation and arthroplasty, and Neer classification etc." Comment: no sensitivity analyses for study design were conducted.
Heterogeneity addressed?	Unclear	Quote: "All studies but 1 included displaced proximal humeral fractures. The majority of the included studies excluded patients with pathologic fractures, patients with open fractures, fractures in skeletally immature patients, and patients with other injuries sustained on the affected side. Most studies (n = 18, 82%) used the Neer classification and included patients with Neer 2-, 3-, or 4-part proximal humeral fractures. In 7 studies, at least 80% of patients were treated with a locking plate.10,11,15,17,30,32,35,38 In 4 studies, arthroplasty was investigated, with hemiarthroplasty in 3 and reverse shoulder arthroplasty in 14,31,36,42; 3 studies assessed proximal humeral nails 9,24,46; and 8 studies used fixation by means of Kirschner wires, screws, a tension-band technique, or a combination of techniques." Comment: diverse operative treatments are used. Little information about the homogeneity of participants.
Similar outcomes?	No	Quote: "In 14 studies (64%, n = 817), the Constant-Murley score was reported after at least 1 year of follow-up." Comment: only 14 out of 22 included studies reported the primary outcome considered in the review.
No selective reporting	Yes	No selective reporting suspected. All outcomes and comparisons reported in methods section are also reported in results section.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Benson 2000

Study characteristics	
Methods	Systematic review

Benson 2000 *(Continued)*

Data	Initial search for observational studies in MEDLINE and CDSR covering the period from 1985 to 1998. Subsequent searches for RCTs (N = 83 included) and observational studies (N = 53 included) (number of participants not reported) published between 1966 and 1998 investigating effect sizes of the same treatments as the initially identified observational studies in MEDLINE **Inclusion criteria:** For initial retrieval of observational studies: studies that were not experimental; that is, treatments were not assigned for the purposes of research, the study assessed the difference between two treatments or between one treatment and no treatment. The treatments were implemented by physicians. Study included a control group. For subsequent retrieval of observational studies and RCTs: comparing the same two treatments (or the same treatment and no treatment), used the same outcome measure, and used the same inclusion criteria for patients as the initially identified observational studies. **Exclusion criteria:** Studies of diet, exercise, lifestyle changes, or non-prescription medication were not included, since the type of bias in these studies differs from the type of bias in studies of physician-implemented treatment.
Comparisons	Pooled odds ratios for RCTs and observational studies were extracted
Outcomes	2 outcomes of relevance for this review: graft survival after kidney transplantation, wound infection
Notes	Reported results: "In most cases, the estimates of the treatment effects from observational studies and randomized, controlled trials were similar. In only 2 of the 19 analyses of treatment effects did the combined magnitude of the effect in observational studies lie outside the 95 percent confidence interval for the combined magnitude in the randomized, controlled trials." Funding: "Supported in part by grants from the Health Services and Resources Administration (PD15 PE87007 and 5D32PE10195-02) and the National Heart, Lung, and Blood Institute (2T35HL07485-21)."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "Observational studies were found by systematically searching [...]." Comment: four inclusion criteria for observational studies reported. Observational studies were matched to RCTs.
Investigator agreement?	Unclear	Quote: "We reviewed the abstracts of these articles and selected only those that met four criteria." Comment: insufficient information reported.
Complete sample?	No	Quote: "Observational studies were found by systematically searching Medline and the Cochrane Database of Systematic Reviews for studies reported from 1985 through 1998. Although Medline is now indexed for highly sensitive searches for randomized, controlled trials, “observational studies” is not an indexable concept in Medline, and there is no search term for observational studies." Comment: the authors could have missed observational studies due to poor indexing.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.

Benson 2000 *(Continued)*

Control for differences?	No	Exemplary quote: "Alternatively, differences may exist between RCTs and observational studies in the care and attention provided." Comment: methodological differences noted, but not controlled for.
Heterogeneity addressed?	No	Quote: "We did not select articles to reduce the heterogeneity of the results or to ensure high quality [...]." Comment: noted, but not controlled for.
Similar outcomes?	No	Quote: "In the selection of corresponding studies, there may have been differences in how some of the treatments were administered (e.g., evaluations by geriatric assessment units) or in how some of the outcomes were assessed [...]." Comment: a few exceptions where outcomes were not similar were noted.
No selective reporting	Unclear	Quote: "Five studies included in our analysis did not report a confidence interval for the magnitude of the effect. For three of these studies, we estimated the confidence interval from the magnitude of the effect and the P value." Comment: all outcomes and comparisons reported in methods section are also reported in results section. However, in studies that did not report 95% confidence intervals, the confidence interval was only imputed for 3 of 5 studies. It is unclear why this was not done in the remaining 2 studies.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Beynon 2008

Study characteristics	
Methods	Systematic review
Data	114 RCTs and 71 observational studies on 19 diverse topics with mortality as the main outcome, published between June 2012 and June 2013, through searches of CENTRAL. **Inclusion criteria:** Not reported **Exclusion criteria:** Not reported
Comparisons	Ratio of relative risks (RRR) extracted for each outcome
Outcomes	1 outcome of relevance for this review: mortality (with various PICOs)
Notes	Reported results: "Non-randomised studies overestimate treatment effects by 12% on average compared with RCTs." Funding: not reported.

Risk of bias

Item	Authors' judgement	Support for judgement

Beynon 2008 *(Continued)*

Explicit criteria?	No	Quote: "We did a meta-epidemiological study, pooling comparisons of RCTs with NRS [non-randomised studies] across clinical topics. We identified topics by randomly selecting RCTs [...]." Comment: RCTs were identified by outcome, then observational studies were matched to an RCT. Eligibility criteria are not reported in this conference abstract.
Investigator agreement?	Unclear	Comment: no information on study selection reported
Complete sample?	No	Comment: topic (all-cause mortality) selected at random, only one database searched.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	No	Comment: the authors mentioned selection bias of observational studies but did not control for this. Different study designs of non-randomised studies are not analysed nor reported separately.
Heterogeneity addressed?	Yes	Quote: "Overall, intervention effect estimates tended to be more beneficial in NRS [non-randomised studies] [...], with some evidence of between-meta-analysis heterogeneity in bias [...]." Comment: heterogeneity was addressed and found to be low. RCTs and non-randomised studies were matched based on PICO criteria.
Similar outcomes?	Yes	Comment: the primary outcome was all-cause mortality in all studies.
No selective reporting	Unclear	Comment: too little information reported in abstract to assess reporting bias.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Bhandari 2004

Study characteristics	
Methods	Systematic review
Data	27 studies included: 14 RCTs and 13 observational studies comparing internal fixation and arthroplasty in people with femoral neck fracture, published between 1962 and 2002. Searches of MEDLINE, Science Citation Index, manual search of table of contents of four major orthopaedic journals (*Journal of Bone and Joint Surgery* (American/British), *Journal of Orthopaedic Trauma*, *Clinical Orthopaedics and Related Research*) from 1998 to June 2002; major trauma textbooks in orthopaedics (Rockwood and Green - *Fractures in Adults*, Browner-Jupiter-Swiontkowski, *Skeletal Trauma*), title review of presentations/posters in programmes of three major orthopaedic meetings (American Academy of Orthopaedic Surgery, Orthopaedic Trauma Association, Canadian Orthopaedic Association) from 1996 to 2002, and contacting content experts. **Inclusion criteria:** Target population: people with displaced femoral neck fractures; Intervention: internal fixation (pin and side plates or multiple screws) versus arthroplasty (hemiarthroplasty, bipolar or total hip arthroplasty); Outcome measure: mortality data available; Methodological criteria: published or unpublished, randomised or non-randomised comparisons in the English literature. **Exclusion criteria:**

Bhandari 2004 *(Continued)*

	Not reported
Comparisons	Pooled relative risks extracted for all outcomes
Outcomes	2 outcomes of relevance for this review: mortality, revision rates
Notes	Reported results: "Non-randomized studies overestimated the risk of mortality by 40% when compared with the results of randomized trials. [...] non-randomized studies underestimated the relative benefit of arthroplasty by 19.5%." Funding: "No funding was received for the preparation of this manuscript."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "We identified articles that met the following eligibility criteria: (1) target population, patients with displaced femoral neck fractures; (2) intervention, internal fixation (pin and side plates or multiple screws) vs arthroplasty (hemi-arthroplasty, bipolar or total hip arthro plasty); (3) outcome measure, mortality data available; (4) methodological criteria, published or unpublished, randomized or non-randomized comparisons in the English literature." Comment: four explicit criteria covering PICO criteria on a focused topic are reported.
Investigator agreement?	Unclear	Quote: "Two of us reviewed the reference lists of all key articles for additional eligible articles." Comment: two researchers reviewed the reference lists of key articles for additional references. No further information about study selection or whether selection was conducted independently.
Complete sample?	Yes	Comment: complete sample on a focused topic. Multiple search strategies used.
Bias assessed?	No	Comment: in table 1, the assessment of possible confounders is reported. No further assessment of study quality was conducted for observational studies. No formal assessment of study quality or risk of bias in RCTs.
Control for differences?	No	Exemplary quote: "Investigators should be aware of the potential differences in results between non-randomized and randomized studies evaluating the evidence for hip fracture management." Comment: methodological differences were discussed, but not controlled for. Non-randomised, observational studies were analysed in aggregate.
Heterogeneity addressed?	Unclear	Comment: heterogeneity was not reported. Eligibility criteria might likely allow for a homogeneous sample.
Similar outcomes?	Yes	Quote: "We identified articles that met the following eligibility criteria: (1) target population, patients with displaced femoral neck fractures; (2) intervention, internal fixation (pin and side plates or multiple screws) vs arthroplasty (hemi-arthroplasty, bipolar or total hip arthro plasty); (3) outcome measure, mortality data available; [...]" Comment: definition of the outcome was part of the selection criteria.

Bhandari 2004 *(Continued)*

No selective reporting	Yes	No selective reporting suspected. Analyses reported in the methods section are also reported in the results section.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias detected.

Borkowska 2018

Study characteristics	
Methods	Systematic review
Data	211 RCT and 165 observational studies included (number of participants not reported) comparing patient-reported outcomes in RCTs and in observational studies for overactive bladder published between 2000 and 2017. Searched MEDLINE, Embase, clinical trial registries, and conference websites (not further specified). **Inclusion criteria:** Not reported **Exclusion criteria:** Not reported
Comparisons	No quantitative data for extraction and analysis reported
Outcomes	2 outcomes of relevance reported: change in scores in King's Health Questionnaire, change score in Overactive Bladder Questionnaire
Notes	This review was reported as a conference abstract and not included in the meta-analysis because it reported scant data. Reported results: observational studies suggest that improvements in health-related quality of life may be higher than the changes estimated in RCTS. Unclear whether this difference is due to selection bias or other factors. Funding: not reported

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	No	Comment: not reported in the abstract
Investigator agreement?	Unclear	Comment: not reported in the abstract
Complete sample?	Unclear	Comment: not reported in the abstract
Bias assessed?	Unclear	Comment: no risk of bias assessment reported in the abstract
Control for differences?	Unclear	Comment: not reported in the abstract
Heterogeneity addressed?	No	Comment: no information about heterogeneity reported
Similar outcomes?	Unclear	Comment: the five most frequently used instruments were reported for RCTs and for observational studies. Comparison of effects when treatment was as-

Borkowska 2018 *(Continued)*

		sessed with the same questionnaire (King's Health Questionnaire scale and Overactive Bladder Questionnaire) is mentioned, but details are not reported in the abstract.
No selective reporting	Unclear	Comment: too little information reported in abstract to assess reporting bias
Absence of evidence of bias from other sources?	Unclear	Comment: too little information reported in abstract to assess other biases

Bröckelmann 2022

Study characteristics	
Methods	Overview of reviews
Data	727 RCTs (n = 2,415,906) and 986 observational studies (n = 14,944,986) evaluating the agreement of effect estimates between bodies of evidence from RCTs and cohort studies in general medicine were included. Authors searched MEDLINE for the period between 2010 and 2019. **Inclusion criteria:** Methods: systematic review of interventions/exposure including RCTs and cohort studies; equivalent search for RCTs and cohort studies; performing quantitative meta-analysis for at least one BoE BoE-pairs: BoE-pair with a BoE from RCTs and a BoE from cohort studies evaluating the same medical research question (e.g. association of exenatide with pancreatitis; effect of vitamin D on hypertension; comparing total knee arthroplasty with unicompartimental knee arthroplasty for range of movement of the knee) Population: all populations (e.g. primary prevention, secondary prevention, general population, adults, children) Intervention/Exposure: all types of medical interventions and exposures (e.g. drugs, invasive, procedures, nutrients, vaccines) Comparator: all types of comparators (e.g. placebo, drugs, invasive, procedures, nutrients, vaccines) Outcomes: patient-relevant outcomes (e.g. mortality, cancer outcomes, cardiovascular outcomes, obstetrical outcomes) and intermediate disease markers (e.g. low-density lipoprotein cholesterol) Study design: randomised controlled trials (e.g. parallel, cluster, factorial, cross-over); cohort studies (e.g. prospective cohort, retrospective cohort, observational cohort analysis of RCT) **Exclusion criteria:** Methods: umbrella reviews, narrative reviews, systematic reviews of diagnostic test accuracy, individual patient data meta-analysis; no quantitative meta-analysis. BoE-pairs: single small study (n < 1000 participants) for one BoE (RCT or cohort studies); BoE-pair with one BoE using a continuous outcome and the other BoE using a binary outcome (e.g. risk of hypertension versus mean difference of systolic blood pressure). Study design: quasi-RCTs, non-randomised controlled trials, case-control studies, cross-sectional studies, ecological studies
Comparisons	RRRs were extracted
Outcomes	1 outcome of relevance for this review: composite outcome (overall comparison of effect estimates from RCTs and observational studies)
Notes	Reported results: pooling RRRs across BoE-pairs with binary outcomes resulted in a pooled RRR of 1.04 (95% CI 0.97 to 1.11; n = 120) with considerable statistical heterogeneity (I^2 = 69%; $\tau 2$ = 0.061; 95% prediction interval 0.63 to 1.71) (Figure 1 and Table 4 in overview publication). Differences of MDs in continuous outcomes (n = 9) were mostly small, except for operation duration for two types of knee prostheses where clear disagreement was shown (42) (Figure 2). In subgroup analyses, degree of PI/ECO-

Bröckelmann 2022 *(Continued)*

similarity, type of intervention, and type of outcome, the pooled RRR indicated that on average, differences between both BoE were small.

Funding: funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Projektnummer 459430615. Open Access funding enabled and organised by Projekt DEAL.

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Comment: detailed PI/ECO-scheme reported that is the basis for eligibility criteria (Table 1, p.2).
Investigator agreement?	No	Quote: "The title and abstract screening was conducted by one reviewer (NB), and potentially relevant full texts were screened by two reviewers independently (NB, LS). Any discrepancy was resolved by a third reviewer (JJM)." Comment: title/abstract screening only conducted by one reviewer; only full text screening in duplicate.
Complete sample?	Yes	Quote: "MEDLINE (PubMed) was searched for relevant systematic reviews in the 13 medical journals with the highest impact factor (according to the Journal Citation Report 2018; category: general and internal medicine)." Comment: a complete sample of systematic reviews is expected.
Bias assessed?	No	Quote: "we did not evaluate the methodological quality of the included systematic reviews, but given that we focused on high-impact journals, we assumed that published systematic reviews are of reasonably high methodological quality" Comment: methodological quality of included SRs not evaluated. However, information on risk of bias or study quality of primary studies extracted.
Control for differences?	Yes	Quote: "We performed pre-specified and post hoc subgroup analyses to explore factors potentially related to the disagreement of effect estimates. The study protocol specified subgroup analysis by degree of PI/ECO-similarity and intervention type (drug, invasive procedure, nutrient, vaccine). Post hoc subgroup analyses were performed by the type of binary effect estimate (RR, OR, HR), type of intervention stratified by degree of PI/ECO-similarity, and type of outcome (e.g., CVD outcomes, cancer outcomes). We performed a post hoc multivariable metaregression among “similar but not identical” BoE pairs with binary outcomes. For each PI/ECO-domain, the average effect on the pooled RoR of the category “similar but not identical” was evaluated as compared to the reference category “more or less identical.” We performed two post hoc sensitivity analyses: First, by including only the BoE pair from each systematic review with the highest number of RCTs (if the number of RCTs was equal, we primarily included the BoE with the highest number of participants, followed by the highest number of events, followed by the highest number of cohort studies) and second, by direction of cohort study summary effect estimate (HR, OR, RR <1 vs. HR, OR, RR ≥1)." Comment: differences were investigated by analyses and considered in these.
Heterogeneity addressed?	Yes	Quote: "We evaluated the statistical heterogeneity of effect estimates across all BoE-pairs with binary outcomes and across BoE pairs using the same continuous outcomes with the I2 and τ2 statistics [69, 70]. To estimate τ2, we used Paule and Mandel method [71, 72]. We computed 95% prediction intervals (PIs) to estimate the extent of differences between results of BoE from RCTs and BoE from cohort studies likely to occur in future comparisons. Meta-analy-

Bröckelmann 2022 *(Continued)*

		ses were performed with the R package meta (73) using random-effects models (69)." Comment: bodies of evidence were assessed for similarities and differences between PICOs of RCTs and observational studies and analysed according to their degree of similarity. Different study types were synthesised separately. Statistical heterogeneity was assessed.
Similar outcomes?	Unclear	Comment: mainly objective outcomes included (mortality, cardiovascular disease, cancer). Subgroup analyses according to the similarity of PICOs were conducted.
No selective reporting	Yes	Comment: no selective reporting suspected. Analyses reported in the methods section are also reported in the results section.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Dahabre 2012

Study characteristics	
Methods	Systematic review
Data	63 RCTs and 21 observational studies included (number of participants not reported) investigating effect sizes of observational studies using propensity score methods and RCTs for acute coronary syndromes. Authors searched for RCTs in the Cochrane Database of Systematic Reviews, MEDLINE, guidelines from the American Heart Association and American College of Cardiologists, reference lists, and published systematic reviews. Authors searched for observational studies in MEDLINE, considering only the top 8 journals (by impact factor as reported by the Institute of Scientific Information, Thomson Reuters) in 'Cardiac and cardiovascular systems', the top four journals in 'Medicine, general and internal' that publish primary clinical research studies. Searches were conducted from inception to 11 February 2011. **Inclusion criteria:** "Studies using Propensity Score methods to obtain estimates of treatment efficacy for therapeutic interventions administered to patients with acute coronary syndrome. Acute coronary syndrome was defined as Acute Myocardial Infarction [ST-elevation myocardial infarction or non-ST-elevation myocardial infarction] or Unstable Angina; we accepted disease definitions as provided by each study. Observational studies of interventions for acute coronary syndrome that used Propensity Score methods to estimate treatment effects on short- or long-term mortality." **Exclusion criteria:** Not reported
Comparisons	Pooled ratios of risk ratios and pooled risk ratios for RCTs and cohort studies extracted
Outcomes	1 outcome of relevance for this review: mortality
Notes	This review was only considered in the sensitivity analyses of our review. Reported results: estimates from PS analyses differed statistically significantly from randomised evidence in two instances; however, observational studies reported more extreme beneficial treatment effects compared with RCTs in 13 of 17 instances. Funding: "This study was supported in part by grant UL1 RR025752 from the National Center for Research Resources to Tufts-Clinical Translational Science Institute. The content is solely the responsi-

Dahabre 2012 *(Continued)*

bility of the authors and does not represent the official views of the National Center for Research Resources or the National Institutes of Health. The funder did not participate in the design, conduct, analysis, interpretation of the study, or the decision to submit the manuscript for publication."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "Our searches for observational studies identified 599 citations, of which 70 were considered to be potentially eligible and were retrieved in full text. Figure 1 presents the search strategy flow along with reasons for exclusion for studies reviewed in full text. Forty-nine observational studies using PS methods were considered eligible for inclusion, of which 21 were successfully matched to 63 RCTs and were considered further." Comment: explicit inclusion criteria are reported in detail.
Investigator agreement?	Yes	Quote: "Two reviewers (I.J.D. and G.D.K.) read potentially eligible studies in full text to determine eligibility; discrepancies were resolved by consensus" Comment: study selection was conducted by two reviewers working independently.
Complete sample?	Unclear	Quote: "the subset of matched observational studies may not be representative of all observational studies using propensity score methods". Comment: the search for studies as well as inclusion and exclusion criteria seem to have resulted in a complete sample of studies. However, the authors report that the sample of observational studies may not be representative.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	Yes	Quote: "we used a binomial (sign) test to evaluate whether a particular design tended to produce favourable results for the experimental treatment more often than would be expected by chance"; "Analyses based on the single largest study (observational or randomized) for each topic yielded results similar to our main analysis (1 significant discrepancy in the 17 topics, i.e. one of the discrepancies in the main analysis was eliminated by considering only the largest available RCT for invasive strategies in non-ST segment elevation acute coronary syndrome reporting on long-term outcomes). One or more RCTs enrolling at least 1000 participants were available in 7 of the 17 topics (2 for short-term and 5 for long-term mortality." Comment: sensitivity analyses were conducted; RCTs and cohort studies were analysed separately.
Heterogeneity addressed?	Yes	Comment: cohort studies and randomised trials were matched according to PICO criteria.
Similar outcomes?	Yes	Quote: "We considered only studies reporting on either short-term (typically within 30 days of ACS diagnosis) or long-term (more than 30 days following ACS diagnosis) mortality because of its clinical importance and the fact that it is less prone to misclassification compared with other outcomes." Comment: mortality (short-term and long-term) was selected as outcome in both study types.
No selective reporting	Yes	Comment: no selective reporting suspected. Analyses reported in the methods section are also reported in the results section.

Dahabre 2012 *(Continued)*

Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias detected.

Demissie 1998

Study characteristics

Methods	Systematic review
Data	9 RCTs (n = 458,428) and 6 case-control studies (n = 4877) investigating results of randomised controlled trials and case-control studies in evaluating the effectiveness of screening mammography. Authors searched MEDLINE for the period between January 1966 and September 1996. **Inclusion criteria:** No further inclusion criteria beyond the search process described: "The following text words were used in locating the articles: breast neoplasm, breast cancer, mortality, mass screening, female, mammography, and screening mammography. Only papers written in English were scrutinized." **Exclusion criteria:** Not reported
Comparisons	Pooled ratios of risk ratios and pooled risk ratios for RCTs and observational studies extracted
Outcomes	1 outcome of relevance for this review: mortality after mammography for breast cancer screening
Notes	Since there was substantial overlap with studies included in Schmidt 2013, we excluded Demissie 1998 from the main analysis. Reported results: "[...] comparison of the summary risk estimates of the RCTs with that of the case-control studies showed RCTs to have a significantly higher summary risk estimate than case-control studies." Funding: "This study was supported by Grant No. P20CAS7142 from the National Cancer Institute, Grant No. ES-05022 from the National Institute of Environmental and Health Sciences, and Grant No. 5-T32-PE10011-04 from the Health Resources and Services Administration."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	No	Comment: no eligibility criteria reported.
Investigator agreement?	Unclear	Comment: no information reported.
Complete sample?	No	Comment: studies were systematically searched for in electronic databases, but some RCTs were also analysed as case-control studies. The extent of double-counting of studies and participants is unclear. Only one database searched.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	Yes	Quote: "A total of 445 articles were identified. Of these, 34 were reports of RCTs and 12 were reports of case-control studies."

Demissie 1998 *(Continued)*

		Comment: only RCTs and case-control studies included. These were analysed separately.
Heterogeneity addressed?	Yes	Comment: studies were searched with the same search terms and the topic of the review was narrow, so heterogeneity was relatively simple to consider. Subgroup analyses with different age groups and interventions were conducted, but studies and participants were probably double-counted. In two of three instances, the statistical heterogeneity was found to be low.
Similar outcomes?	Yes	Quote: "[...] in order to identify studies that assessed the efficacy of screening mammography in reducing the risk of death from breast cancer." Comment: mortality as outcome of interest allows for sufficient similarity.
No selective reporting	Yes	Comment: no selective reporting suspected. Reported analyses are in accordance with those proposed in the methods section.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Edwards 2012

Study characteristics	
Methods	Systematic review
Data	12 RCTs and 26 observational studies (n total = 32,969) investigating breast cancer surgery. Authors searched MEDLINE, Embase, and Cochrane Databases (not further specified) from 2003 to 2008. **Inclusion criteria:** RCTs had to be truly randomised and had to involve comparisons of two surgical procedures used for the treatment of breast cancer. Non-RCTs had to be published in English, use a non-randomised study design, and compare two groups that were comparable to those represented in the relevant RCT. **Exclusion criteria:** Studies with historical control group
Comparisons	Pooled relative risks of RCTs and observational studies extracted
Outcomes	3 outcomes of relevance for this review: sensory deficit, recurrence, mortality
Notes	Reported results: "Randomized controlled trials comparing surgical procedures for breast cancer may demonstrate clinically relevant differences in effect estimates in 20%-40% of cases relative to those generated by non-randomised trials, depending on which metric is used." Funding: "W.A. Ghali is supported by a Canada Research Chair in Health Services Research and by a Senior Health Scholar Award from the Alberta Heritage Foundation for Medical Research."

Risk of bias		
Item	**Authors' judgement**	**Support for judgement**
Explicit criteria?	Yes	Quote: "Our first goal was to identify all existing English-language RCTs published between January 2003 and May 2008 that included 2 surgical arms for the treatment of breast cancer."

Edwards 2012 *(Continued)*

		Comment: inclusion and exclusion criteria for RCTs and observational studies are reported.
Investigator agreement?	Yes	Quote: "Two reviewers (J.P.E., and either E.J.K. or A.J.G.) independently screened citations by title and abstract to identify studies for full-text review"; " Individual full text articles were then independently reviewed to determine eligibility" Comment: study selection was done by two reviewers working independently.
Complete sample?	Unclear	Quote: "We limited our search to this 5-year period with the hope of identifying RCTs for which there would likely be a large cohort of earlier nRCTs available for comparison." Comment: literature search and flow chart are reported. The selective search may have introduced bias by not selecting all available literature. Search for non-randomised studies only through related-article function and through screening of reference lists of eligible RCTs.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	No	Comment: no sensitivity analyses by study design conducted.
Heterogeneity addressed?	Yes	Comment: the authors calculated the heterogeneity within each meta-analysis. Studies were matched by an algorithm to ensure comparability. Studies were analysed based on their intervention.
Similar outcomes?	Unclear	Quote: "The outcomes of interest were determined after identifying matched groups of RCTs and nRCTs in a hierarchical manner. If mortality or recurrence were assessed in both the RCTs and nRCTs, these were employed as the outcome for analysis. If these data on mortality or recurrence were not available from both study types, we used objectively measured outcomes found in both study types. Finally, if objectively measured outcomes were not available, we used subjectively assessed or self-reported outcomes as a last resort." Comment: mortality and recurrence were used as primary outcomes. If those were not reported, other objective outcomes were used, and if these were lacking, subjective measures were used.
No selective reporting	Yes	Comment: no selective reporting suspected. Reported analyses are in accordance with those proposed in the methods section.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Furlan 2008

Study characteristics	
Methods	Systematic review
Data	8 RCTs and 17 observational studies (number of participants not reported) allowing for the comparison of results from RCTs and NRSs of interventions for low back pain. Authors searched MEDLINE and Embase up to May 2005 for NRSs meeting the inclusion criteria. Matching RCTs were identified though searching the reference lists of systematic reviews, CENTRAL, MEDLINE, and Embase. **Inclusion criteria:**

Furlan 2008 *(Continued)*

	Non-randomised study comparing two or more interventions for low-back pain. Only observational studies with a comparison group were included; any intervention (prevention or treatment); any type of low-back pain (acute or chronic, non-specific aetiology, disc herniation, spinal stenosis, osteoarthritis, etc.); any type of outcome measures **Exclusion criteria:** Not reported
Comparisons	Pooled odds ratios for RCTs and observational studies extracted
Outcomes	1 outcome of relevance for this review: low back pain at different time points
Notes	Reported results: NRSs frequently either agree with RCTs or underestimate the effects compared with RCTs. Funding: "Supported by the Canadian Institute of Health Research (CIHR) (to A.F.), Canada Research Chair in Knowledge Transfer for Musculoskeletal Care (to C.B.), and Canada Research Chair in e-Health Innovation (to A.J.)."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "Comprehensive searches were conducted in MEDLINE and EMBASE up to May 2005, for articles meeting the following inclusion criteria: NRS comparing 2 or more interventions for low back pain." Comment: observational studies identified according to specific criteria then matched to RCTs
Investigator agreement?	Unclear	Comment: no information reported.
Complete sample?	No	Quote: "Because there were many interventions, we selected the 3 interventions with the most nonrandomized studies." Comment: selected interventions with the most observational studies.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	No	Quote: "To determine the influence of study design (randomized or not) and other potential effect modifiers, we performed subgroup analyses." Comment: no control for differences in study design amongst observational studies; observational, non-randomised studies were analysed in aggregate.
Heterogeneity addressed?	Unclear	Comment: different interventions were subgrouped and analysed separately. Very little information about control for heterogeneity amongst populations.
Similar outcomes?	Unclear	Quote: "Included in the database were only observational studies with a comparison group, of any intervention, any type of low back pain, and any type of outcomes measure" Comment: grouped by intervention not outcome; outcome might differ between the included studies
No selective reporting	Yes	Comment: no selective reporting suspected, all analyses described in methods section are reported in the results section.

Furlan 2008 *(Continued)*

Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Golder 2011

Study characteristics

Methods	Overview of reviews 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 investigating adverse effects searched in databases from inception up to November 2009 from the CDSR, the Cochrane Methodology Register, Database of Abstracts of Reviews of Effects, Embase, Health Technology Assessment Database, Health Management Information Consortium, Index to Theses, Library, Information Science & Technology Abstracts and MEDLINE. The authors conducted handsearches in the journals BMC Clinical Pharmacology, BMC Medical Research Methodology, Drug Safety, Health Information and Libraries Journal (formerly Health Libraries Review), Journal of Clinical Epidemiology, Journal of Information Science, Journal of Librarianship and Information Science, Journal of the Medical Library Association (formerly the Bulletin of the Medical Library Association), Pharmacoepidemiology & Drug Safety. The following conference proceedings were handsearched: Cochrane Colloquia, HTAi, Pharma-Bio-Med Conference and Exposition, Symposium on Systematic Reviews. The following web sources were searched: Agency for Healthcare Research and Quality, Health Technology Assessment Programme. **Inclusion criteria:** Systematic reviews that evaluated studies of more than one type of design (for example, RCTs versus cohort or case-control studies) on the identification and/or quantification of adverse effects of health-care interventions; meta-analyses that reported pooled estimates of the risk of adverse effects according to study designs that the authors stated as RCTs as opposed to analytic epidemiologic studies such as case-control and controlled cohort studies; possible to compare the pooled risk ratios or odds ratios from RCTs against those from other study designs. **Exclusion criteria:** Not reported
Comparisons	Pooled ratios of odds ratios extracted comparing RCTs versus cohort studies, RCTs versus case-control studies, and RCTs versus studies labelled as observational studies
Outcomes	1 outcome of relevance for this review: adverse effects
Notes	Reported results: "there is no difference on average in the risk estimate of adverse effects of an intervention derived from meta-analyses of RCTs and meta-analyses of observational studies." Funding: "This research was undertaken by Su Golder as part of an MRC fellowship."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Unclear	Quote: "A meta-analysis or evaluation study was considered eligible for inclusion in this review if it evaluated studies of more than one type of design (for example, RCTs versus cohort or case-control studies) on the identification and/or quantification of adverse effects of health-care interventions. We were principally interested in meta analyses that reported pooled estimates of the risk of adverse effects according to study designs that the authors stated as RCTs,

Golder 2011 *(Continued)*

		as opposed to analytic epidemiologic studies such as case-control and controlled cohort studies (which authors may have lumped together as a single ''observational'' category). Our review focuses on the meta-analyses where it was possible to compare the pooled risk ratios (RRs) or odds ratios (ORs) from RCTs against those from other study designs." Comment: studies were identified from published meta-analyses in 5 journals.
Investigator agreement?	No	Quote: "In particular, one reviewer (S. G.) undertook a detailed hand search [...]. A second reviewer (Y. K. L.) checked the included and excluded papers that arose from this hand search." Comment: one reviewer did the handsearching, which was checked by a second reviewer.
Complete sample?	Unclear	Comment: electronic databases and other sources were searched; studies were selected based on relatively broad inclusion criteria appropriate for this topic.
Bias assessed?	No	Quote: "The following criteria were used to consider the validity of comparing risk estimates across different study designs. (1) Presence of confounding factors" Comment: no standardised, formal risk of bias/quality assessment; authors only checked if studies controlled for confounding, heterogeneity, and the statistical analysis used for meta-analyses.
Control for differences?	Yes	Comment: different observational study designs were analysed separately.
Heterogeneity addressed?	Unclear	Quote: "There was considerable heterogeneity between the comparisons of different studies, suggesting that any differences may be specific to particular types of interventions or adverse effects." Comment: heterogeneity amongst participants and interventions was described descriptively.
Similar outcomes?	No	Comment: only one outcome had multiple studies addressing it.
No selective reporting	Yes	Comment: no selective reporting suspected; analyses described in methods section are reported in the results section
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Gu 2020

Study characteristics	
Methods	Systematic review
Data	43 RCTs (n = 183,752) and 41 observational studies (n = 1,879,428) included, examining risk of major gastrointestinal bleeding with new versus conventional oral anticoagulants. Authors searched MEDLINE, Embase, the Cochrane Library, and ClinicalTrials.gov up to October 2018, as well as the reference lists of included articles. **Inclusion criteria:**

Gu 2020 *(Continued)*

	"RCTs or real-world studies that compared new versus conventional oral anticoagulants and reported data on major gastrointestinal bleeding were eligible for inclusion. [...] The language restriction of English, focus on the highest-quality real-world studies [RWSs], nationwide or health insurance database studies that reported adjusted or matched major gastrointestinal bleeding results by using authorized method to minimize confounding (propensity score adjustment, propensity score matching, inverse probability of treatment weighting, and covariate adjustment). When several RWSs used the same data source from an overlapping period, we only included the one that reported adjusted gastrointestinal bleeding data with the longest study period." **Exclusion criteria:** Studies that compared new oral anticoagulants with placebo, studies that reported only crude results or published only in conference abstract or letter form.
Comparisons	Pooled RR for RCTs and cohort studies extracted
Outcomes	3 outcomes of relevance for this review: major gastrointestinal bleeding, upper major gastrointestinal bleeding, lower major gastrointestinal bleeding
Notes	Results: "The pooled major rates of GIB [gastrointestinal bleeding] for patients on NOACs [non-vitamin K antagonist oral anticoagulants] (1.19%) vs conventional treatment (0.92%) did not differ significantly (RR from randomized controlled trials, 1.09; 95% CI, 0.91–1.31 and aHR [adjusted hazard ratio] from real-world studies, 1.02; 95% CI, 0.94–1.10)." Funding: "Supported by grants from the National Key Research and Development Program of China (2018YFC1312800), National Science Fund for Distinguished Young Scholars (81625002), National Natural Science Foundation of China (no. 71804109, no. 81502991, and no. 81803841), Shanghai Outstanding Academic Leaders Program (18XD1402400), Research Funds of Shanghai Health and Family Planning commission (20184Y0022), Clinical Pharmacy Innovation Research Institute of Shanghai Jiao Tong University School of Medicine (CXYJY2019ZD001, CXYJY2019QN004), and Program for Key but Weak Disciplines of Shanghai Municipal Commission of Health and Family Planning (2016ZB0304) Innovative research team of high-level local universities in Shanghai."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "To focus on the highest-quality RWSs, we only included nationwide or health insurance database studies that reported adjusted or matched major GIB results by using authorized method to minimize confounding (propensity score adjustment, propensity score matching, inverse probability of treatment weighting, and covariate adjustment).1 When several RWSs used the same data source from an overlapping period, such as the Danish health insurance data set from 2011 to 2014 (Danish Civil Registration system, National Patient Register, and National Prescription Registry), we only included the one that reported adjusted GIB data with the longest study period. Studies that reported only crude results or published only in conference abstract or letter form were excluded." Comment: eligibility criteria are broad but clearly reported.
Investigator agreement?	Yes	Quote: "Three reviewers (Z. G., A. W., C. Z.) independently assessed all study titles and abstracts for determining eligibility, and then full articles were retrieved and assessed according to inclusion criteria, with any disagreements being resolved by corresponding authors (J. P., H. L.)." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Yes	Quote: "We did not observe potential publication bias by qualitative funnel plots as well as Begg's test and Egger's test" (Supplementary Figures 1 and 2).

Gu 2020 *(Continued)*

		Comment: a complete sample is expected.
Bias assessed?	Unclear	Quote: "Comment: Risk of bias was assessed with the help of the Cochrane RoB tool." Comment: risk of bias assessment of observational studies unclear.
Control for differences?	Yes	Quote: "To test the robustness of the primary results, we conducted a series of sensitivity analyses by sequential elimination of each study from the pool or excluding studies that involved special clinical scenarios." Comment: differences in the studies were investigated through sensitivity analyses.
Heterogeneity addressed?	Yes	Quote: "For RWSs, we pooled aHRs and their 95% CIs by using random-effects models and performed subsequent subgroup analyses according to indications (AF, VTE, and other special clinical scenarios), controls (VKAs and antiplatelet agents), dosage (standard dose and low dose), gender (men and women), age (elderly patients, <75 years, and >75 years), VKA switchers, and population (United States, Canada, Europe, Taiwan, and New Zealand)." Comment: statistical heterogeneity was assessed and reported in text for the analyses.
Similar outcomes?	Yes	Quote: "The primary outcome was major GIB, defined as a decrease in hemoglobin level of 2 g/dL or greater within a 24-hour period, or leading to a transfusion of 2 or more units of packed red cells, or requiring an additional endoscopy intervention, according to the International Society on Thrombosis and Hemostasis criteria for RCTs and International Classification of Disease revision 9 or 10 codes of major GIB for RWSs. The secondary outcomes were upper and lower major GIB, with the same definition as the primary outcome." Comment: similar outcomes were used and were already reflected in inclusion criteria.
No selective reporting	Yes	Comment: no selective reporting suspected.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Guyatt 2000

Study characteristics	
Methods	Systematic review
Data	13 RCTs and 17 observational studies (number of participants not reported) investigating adolescent pregnancy preventions. Authors searched the period between 1970 and May 1993 in CATalog onLINE, CINAHL, conference papers index, Dissertation abstracts online, Embase, Educational Resources Information Center, MEDLINE, National Technical Information Services, POPulation information onLINE, PsycINFO (PSYCHOLOGICAL ABSTRACTS), and SOCIOLOGICAL ABSTRACTS. Authors searched the table of contents of the following journals for 1992 and 1993: *Family Planning Perspectives*, *Adolescence*, *Journal of Adolescent Health Care*, *American Journal of Public Health*, *Youth and Society*, *Journal of Adolescent Research*, *Journal of Early Adolescence*, and *Journal of Research on Adolescence*. **Inclusion criteria:**

Guyatt 2000 *(Continued)*

	"Randomized trials and observational studies focusing on adolescents 18 years of age or less, evaluating a variety of primary prevention programs including sex education classes, school-based clinics, free-standing clinics, physician/nurse practitioner practice-based service, improved access, and community-based programs. Primary studies were included if they reported initiation of sexual intercourse, birth control use, or pregnancy and had been conducted in North America, Australia, New Zealand, United Kingdom, Europe (excluding Eastern Europe) or Scandinavia. Both published studies and dissertations, conference proceedings, technical reports, and other unpublished documents." **Exclusion criteria:** Studies from Eastern Europe.
Comparisons	Pooled odds ratios for RCTs and observational studies extracted
Outcomes	8 outcomes of relevance for this review: initiation of intercourse (males), initiation of intercourse (females), pregnancy (females), responsible sexual behaviour (males), birth control use (males), birth control use (females)
Notes	Reported results: "The difference between the results of the observational studies and randomized trials was statistically significant in two of the eight outcomes. Observational studies yield systematically greater estimates of treatment effects than randomized trials [...]." Funding: not reported

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "We included randomized trials and observational studies focusing on adolescents 18 years of age or less, evaluating a variety of primary prevention programs including sex education classes, school-based clinics, free-standing clinics, physician/nurse practitioner practice-based service, improved access, and community-based programs. We included primary studies if they reported initiation of sexual intercourse, birth control use, or pregnancy and had been conducted in North America, Australia, New Zealand, United Kingdom, Europe (excluding Eastern Europe) or Scandinavia. We included both published studies and dissertations, conference proceedings, technical reports, and other unpublished documents that met our eligibility criteria. We performed all analyses separately by sex." Comment: detailed eligiblity criteria reported.
Investigator agreement?	Yes	Quote: "Two individuals independently rated each citation in every search to determine whether it met eligibility criteria for retrieval. We retrieved any article that either rater thought, on the basis of the title, might be relevant to the overview. Once retrieved, we rated, independently and in duplicate, the relevance of the full manuscripts. Disagreement was resolved by consensus." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Yes	Quote: "We searched the following computerized databases: CATLINE (CATalog onLINE), CINAHL (Cumulative Index to Nursing and Allied Health Literature), CONFERENCE PAPERS INDEX, DISSERTATION ABSTRACTS ONLINE, EMBASE, ERIC (Educational Resources Information Center), MEDLINE, NTIS (National Technical Information Services), POPLINE (POPulation information onLINE), PsycINFO (PSYCHOLOGICAL ABSTRACTS), and SOCIOLOGICAL ABSTRACTS (search strategies provided on request). We reviewed the reference lists of all papers for relevant citations. In addition, we reviewed the table of contents of the following journals for 1992 and 1993; Family Planning Perspectives, Adolescence, Journal of Adolescent Health Care, American Journal of

Guyatt 2000 *(Continued)*

		Public Health, Youth and Society, Journal of Adolescent Research, Journal of Early Adolescence, and Journal of Research on Adolescence. If we found any relevant articles, we extended the hand search back to 1988." Comment: a complete sample is expected.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	No	Comment: no sensitivity analyses reported, cohort and case-control studies were collapsed as observational studies.
Heterogeneity addressed?	Unclear	Quote: "Although few interventions were identical, there was considerable overlap among many." Comment: heterogeneity between participants and interventions was noticed but not controlled for systematically. Males and females were analysed separately.
Similar outcomes?	Yes	Quote: "We included randomized trials and observational studies that evaluated the impact of primary prevention interventions including sex education classes, school-based clinics, free-standing clinics, physician/nurse practitioner practice-based service, improved ac cess, and community-based programs on four outcomes: sexual intercourse, birth control use, responsible sexual behavior, or pregnancy in adolescents." Comment: outcomes seem similar across all studies.
No selective reporting	Yes	Comment: no selective reporting suspected; all analyses described in methods section are reported in the results section
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Hong 2021

Study characteristics	
Methods	Overview of reviews
Data	30 reviews (number of included studies not reported) that investigating how relative treatment effects of pharmaceuticals differ between observational studies and randomised controlled trials. Authors searched MEDLINE and Embase from January 1990 to January 2020. **Inclusion criteria:** Study design: published systematic literature reviews designed to compare relative treatment effects from observational studies with the corresponding effects from RCTs; or published systematic literature reviews that reported subgroup analyses stratified by RCT and observational study design; and observational studies included in these reviews had to be retrospective or prospective cohort studies, or case-control studies. Population: humans Intervention(s) and comparator(s): any active or placebo-controlled pharmaceutical or biopharmaceutical intervention Outcome(s): efficacy/effectiveness or safety outcomes, pooled relative treatment effect estimates for both observational studies and RCTs **Exclusion criteria:** Systematic reviews that compared absolute outcomes, such as event rates, between non-comparative observational studies and RCTs; non-pharmaceutical-based studies, e.g. surgical procedures, tradition-

Hong 2021 *(Continued)*

	al medicine, vitamin/herbal supplements, etc.; non-English language; abstracts or conference proceedings
Comparisons	Pooled ratios of odds ratios extracted
Outcomes	Pooled ratios of odds ratios extracted across various outcomes
Notes	Conference abstract Reported results: "There was no statistically significant difference (based on the 95% CI) in relative effect estimates between RCTs and observational studies in 79.7% of pairs. There was an extreme difference (ratio < 0.7 or > 1.43) in 43.2% of pairs, and, in 17.6% of pairs, there was a significant difference and the estimates pointed in opposite directions. There is significant variation in about 20% of comparisons". Funding: no funding was received for this study.

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "The search focused on comparative assessments of pharmaceuticals in the English language. We included reviews designed to compare relative treatment effects from observational studies with the corresponding effects from RCTs, and reviews that conducted subgroup analyses by study design" Comment: inclusion criteria are broad but seem very inclusive and adequate.
Investigator agreement?	Unclear	Quote: "three authors (JG, YH and LO) screened the titles and abstracts to identify relevant reviews. Once complete, LO verified the screening for accuracy. Following the title and abstract screen, full text articles were obtained for all potentially relevant reviews. Full text articles were then assessed to determine if they meet the selection criteria for final inclusion in the review." Comment: eligibility was assessed by two reviewers in title and abstract screening. The eligibility process in not clearly described for full-text screening.
Complete sample?	Yes	Comment: eligibility criteria were broad, search seems systematic, hence a complete consecutive sample can be expected.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	Yes	Comment: RCTs and observational studies were analysed separately. No further differentiation between types of observational studies.
Heterogeneity addressed?	No	Comment: no measure of heterogeneity reported; assessment of heterogeneity not reported in methods section.
Similar outcomes?	Unclear	Quote: "Seventy-four relative effect estimate pairs (hazard ratios, risk ratios or odds ratios) from 29 reviews that reported pooled relative effect estimates for RCTs and observational studies, including a variety of outcomes, intervention-comparators, and indications, comprised our final analysis sample." Comment: mostly, RCTs and observational studies used similar outcome measures. Occasionally, effect estimates from RCTs and observational studies were reported with different effect measures.
No selective reporting	Yes	Comment: no selective reporting suspected.

Hong 2021 *(Continued)*

Absence of evidence of bias from other sources?	Yes	Comment: no other forms of bias suspected.

Hoshino 2021a

Study characteristics

Methods	Systematic review
Data	5 RCTs, 20 cohort studies, and 10 case-control studies (number of participants not reported) investigating surgical outcomes of rectal cancer by study design. Authors searched MEDLINE, Scopus, and Cochrane CENTRAL up to June 2019. **Inclusion criteria:** Studies comparing laparoscopic anterior resection to open laparoscopic anterior resection for rectal cancer were eligible. When multiple surgical procedures were included in a study, studies in which over 70% of participants underwent laparoscopic anterior resection were included. Study design was restricted to RCT, case-matched study, or cohort study. Both prospective and retrospective studies were included, and the method of randomisation or matching was not restricted. The language was restricted to English. **Exclusion criteria:** Small studies that included fewer than 50 participants for each intervention group.
Comparisons	Pooled relative risks, risk differences, and mean differences were extracted for RCTs, cohort studies, and case-matched cohort studies.
Outcomes	11 outcomes of relevance for this review: postoperative overall complications, anastomotic leakage, mortality, reoperation, length of hospital stay, operative time, estimated blood loss (mL), positive circumferential resection margins, 3-year overall survival, 3-year disease-free survival, 3-year local recurrence rate
Notes	Reported results: findings did not differ between RCT and case-matched studies for most outcomes. Funding: this review was supported by a grant from Kondou Kinen Medical Foundation.

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "Studies in which laparoscopic LAR was compared with open LAR for rectal cancer were eligible. When multiple surgical procedures were included in a study, studies in which over 70% of patients underwent LAR were included. Small studies that included less than 50 patients for each intervention group were excluded. Study design was restricted to RCT, case-matched study, or cohort study. Both prospective and retrospective studies were included, and the method of randomization or matching was not restricted. The language was restricted to English." Comment: eligibility criteria are described in sufficient detail.
Investigator agreement?	Yes	Quote: "Two review authors (NH and YF) independently screened the titles and abstracts of studies identified by literature search, and then assessed the full texts of potential eligible articles. Disagreement was resolved by discussion."

Hoshino 2021a *(Continued)*

		Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Yes	Comment: eligibility criteria were broad, the search seems systematic, hence a complete consecutive sample can be expected.
Bias assessed?	No	Quote: "Also, this review included published data only and did not assess study quality." Comment: risk of bias was not assessed.
Control for differences?	Yes	Comment: RCTs, cohort studies, and case-matched studies were analysed separately.
Heterogeneity addressed?	No	Quote: "A random-effects model was used for all meta-analyses because of presumed heterogeneity in the surgical quality of LAR across the included studies." Comment: heterogeneity only controlled for through random-effects meta-analysis.
Similar outcomes?	Yes	Quote: "Short-term outcomes were the incidence of postoperative overall complications, the incidence of anastomotic leakage, mortality, reoperation rate, length of stay, operative time, estimated blood loss, and rate of positive circumferential resection margins. Long-term outcomes were 3-year overall survival (OS), 3-year disease-free survival (DFS), and 3-year local recurrence rate (LRR)." Comment: outcome measures were similar across included studies.
No selective reporting	Yes	Comment: no selective reporting suspected; all outcomes reported in methods section are also reported in the results section.
Absence of evidence of bias from other sources?	Yes	Comment: no other forms of bias suspected.

Hoshino 2021b

Study characteristics	
Methods	Systematic review
Data	7 RCTs and 52 observational studies investigating short- and long-term outcomes of laparoscopic versus open low anterior resection for rectal cancer. Authors searched MEDLINE, Scopus, and Cochrane CENTRAL up to June 2019. **Inclusion criteria:** Studies comparing robotic versus laparoscopic surgery for rectal cancer; RCTs, case-matched studies, and cohort studies: both prospective and retrospective studies were included in non-RCT studies. No restrictions were imposed regarding methods of randomisation or matching. **Exclusion criteria:** Studies of transanal surgery.
Comparisons	Pooled relative risks, risk differences, and mean differences were extracted for RCTs, cohort studies, and case-matched cohort studies.

Hoshino 2021b *(Continued)*

Outcomes	9 outcomes of relevance for this review: postoperative overall complications, anastomotic leakage, mortality, duration of hospital stay, conversion rate, duration of operation, estimated blood loss, positive circumferential resection margins, quality of total mesorectal excision
Notes	Reported results: "Case-matched studies occasionally overestimated the effects of interventions compared with RCTs." Funding: the study received no funding.

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "Eligible studies were those comparing robotic versus laparoscopic surgery for rectal cancer. Studies of transanal surgery were excluded. RCTs, case-matched studies, and cohort studies were subjected to analysis. Both prospective and retrospective studies were included in non-RCT studies. No restrictions were placed regarding methods of randomization or matching." Comment: inclusion criteria are broad but seem adequate and systematic
Investigator agreement?	Yes	Quote: "Two authors independently screened the extracted publications according to title and abstract, and then reviewed the full text of potentially eligible articles. Disagreement was resolved by discussion." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Yes	Comment: eligibility criteria were broad, the search seems systematic, hence a complete consecutive sample can be expected
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	Yes	Comment: RCTs, cohort studies and case-matched cohort studies were analysed separately.
Heterogeneity addressed?	No	Comment: no information reported about controlling for heterogeneity.
Similar outcomes?	Yes	Quote: "Primary outcomes were: incidence of postoperative overall complications, incidence of anastomotic leakage, and mortality. Secondary outcomes were: duration of hospital stay, conversion rate, duration of operation, estimated blood loss, rate of positive circumferential resection margins, and quality of total mesorectal excision." Comment: outcome measures seem similar across all included studies.
No selective reporting	Yes	Comment: no selective reporting suspected; all outcomes reported in methods section are also reported in the results section.
Absence of evidence of bias from other sources?	Yes	Comment: no other forms of bias suspected.

Ioannidis 2001

Study characteristics

Methods	Systematic review

Ioannidis 2001 *(Continued)*

	Identified meta-analyses that considered both RCTs and observational studies published before 2000
Data	45 topics identified from 240 RCTs and 168 observational studies investigating evidence of treatment effects in randomised and non-randomised studies identified through: review of the previous literature on comparisons on RCTs and NRSs until mid-1998; reference lists of identified articles; a search in the authors' personal database of meta-analyses between 1991 and 1997 in the *Journal of the American Medical Association*, *The Lancet*, *British Medical Journal*, *Annals of Internal Medicine,* and *Archives of Internal Medicine*; searches in MEDLINE until March 2000, the Cochrane Database of Systematic Reviews, and screening of meta-analyses performed by investigators in the authors' network **Inclusion criteria:** "Meta-analyses in which both randomized and nonrandomized studies were cited with at least 1 primary outcome being in binary form. Data on the binary outcome had to be presented in the meta-analysis. Binary data for the same outcome had to be available on at least 1 randomized trial and at least 1 nonrandomized study" **Exclusion criteria:** Not reported
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	Reported results: "Despite good correlation between randomized trials and nonrandomized studies - in particular, prospective studies - discrepancies beyond chance do occur and differences in estimated magnitude of treatment effect are very common." Funding: "The work was supported by grant PENEO ED27 (974) from the Program for Support of Research Potential, Greek Secretariat for Research and Technology, funded through the European Union. Also supported in part through the New England Medical Center Research Fund."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "we identified meta-analyses that had considered both randomized and non-randomized evidence"; "From all these sources, we selected the meta-analyses in which both randomized and non-randomized studies were cited with at least 1 primary outcome being in binary form." Comment: very explicit for meta-analyses identified and studies within the meta-analyses
Investigator agreement?	Unclear	Comment: the authors do not report explicit information on the process of eligibility except for saying "we".
Complete sample?	No	Comment: the search for eligible studies might have missed relevant studies. MEDLINE, Cochrane Database of Systematic Reviews, and other generally unstandardised databases were searched.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	Yes	Comment: different study designs were subgrouped.
Heterogeneity addressed?	No	Comment: primary studies were pooled in included reviews. No matching or control for heterogeneity reported by review authors.

Ioannidis 2001 *(Continued)*

Similar outcomes?	Yes	Comment: reviews of studies investigating the same outcome were included. The most important clinical outcome of each review was included in this review.
No selective reporting	Yes	Comment: no selective reporting suspected.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias detected.

Jainaud 2021

Study characteristics	
Methods	Overview of overviews of systematic reviews
Data	Unclear number of RCTs and observational studies investigating evidence on putative risk and protective factors in 57 topics. Authors searched up to November 2020 in PubMed (database unclear). **Inclusion criteria:** "All umbrella reviews including meta-analyses of observational studies assessing putative risk or protective factors were eligible. We considered all putative factors (i.e., any attributes, characteristics, or exposure of an individual that may either increase or decrease the occurrence of any type of health outcomes)." **Exclusion criteria:** "Umbrella reviews not assessing any putative risk or protective factors in observational settings or not using any of seven proposed standardized criteria to assess the evidence."
Comparisons	Ratios of odds ratios extracted
Outcomes	32 outcomes of relevance for this review: 30-day mortality, all-cause mortality, cancer, colorectal cancer, hip fracture, gastric cancer, haematological cancer, in-hospital mortality, infection, infection-related mortality, liver cancer, myopathy, pancreatitis, prostate cancer, sustained virological response, acute respiratory distress mortality, bladder cancer, breast cancer, gynaecologic cancer, kidney cancer, pancreatic cancer, skin cancer, lung cancer, femoral bone mineral density, hip bone mineral density, spine bone mineral density, fracture, acute kidney injury, diabetes, cataract, ventilator-free days, contrast-induced nephropathy
Notes	Reported results: "The differences between the meta-analyses estimates of observational studies and RCTs were beyond chance for 43.5% (27/62) associations." Funding: "METRICS is supported by a grant from the Laura and John Arnold Foundation. The work of JPAI is supported by an unrestricted gift from Sue and Bob O'Donnell. ET is supported by a CRUK Career Development Fellowship (C31250/A22804)."

Risk of bias		
Item	**Authors' judgement**	**Support for judgement**
Explicit criteria?	Yes	Quote: "All umbrella reviews including meta-analyses of observa tional studies assessing putative risk or protective factors were eligible"; "Umbrella reviews not assessing any putative risk or protective factors in obser vational settings or not using any of the seven previously proposed standardized criteria (Table 1) to assess the evidence were excluded."

Jainaud 2021 *(Continued)*

		Comment: eligibility criteria are reported in sufficient detail to identify umbrella reviews.
Investigator agreement?	No	Quote: "One author (PJ) screened all resulting articles from the literature search for inclusion criteria and consulted with a second author (JPA) when in doubt." Comment: study selection was done by one reviewer.
Complete sample?	No	Comment: eligibility criteria were broad but maybe sufficient to identify umbrella reviews; search was conducted in PubMed only.
Bias assessed?	Unclear	Comment: assessment of evidence of observational studies conducted with the help of seven standardised criteria. No formal assessment of study quality or review quality.
Control for differences?	No	Comment: observational studies were not assessed separately by study type. Overall, unclear attribution of studies to analyses.
Heterogeneity addressed?	No	Quote: "The estimates across different designs were paired according to outcome, exposure, comparison, and population." Comment: heterogeneity was noted but not controlled for.
Similar outcomes?	Unclear	Comment: outcomes reported multiple times. Differences between PICO for each outcome unclear. No definitions reported.
No selective reporting	Yes	Comment: no selective reporting suspected.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Kimachi 2021

Study characteristics	
Methods	Overview of reviews
Data	204 RCTs and 418 observational studies investigating systematic differences in effect estimates between observational studies and randomised control trials in meta-analyses in nephrology. Authors searched between 2006 and 2016 in MEDLINE and Embase. **Inclusion criteria:** "All comparative observational studies in nephrology that assess the trends and characteristics of systematic reviews of observational studies in nephrology in the past decade." Studies of kidney disease were selected on the following criteria: (1) participants with kidney disease; (2) studies with primary outcomes related to kidney diseases. Meta-analyses which combined observational studies and RCTs and compared two specific interventions. **Exclusion criteria:** Studies with participants with extra-renal diseases including ureteral, urethral, and urinary bladder diseases. Studies in which kidney diseases were treated as a composite outcome (e.g. composite outcome of kidney, pancreas, and liver cancers).
Comparisons	Ratios of odds ratios extracted

Kimachi 2021 *(Continued)*

Outcomes	1 outcome of relevance for this review: systematic differences in the effect estimates between observational studies and RCTs in meta-analyses combining both types of study in nephrology
Notes	Reported results: ratios of odds ratios with a 95% confidence intervals revealed that effect estimates were, on average, consistent between the two study designs. Funding: "A.O. reports personal fees from Chugai, personal fees from Ono Pharmaceutical, personal fees from Eli Lilly, personal fees from Mitsubishi-Tanabe, personal fees from Asahi-Kasei, personal fees from Takeda, personal fees from Pfizer, grants from Advantest, outside the submitted work; A.T. reports personal fees from Mitsubishi-Tanabe, personal fees from Dainippon-Sumitomo, and personal fees from Otsuka, outside the submitted work; T.A.F. reports grants and personal fees from Mitsubishi-Tanabe, personal fees from MSD, personal fees from Shionogi, and outside the submitted work; M.K. and K.K. declare that they have no relevant financial interests."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "We included systematic reviews of all comparative observational studies in nephrology to assess the trends and characteristics of systematic reviews of observational studies in nephrology in the past decade. We included systematic reviews published from 2006 to 2016 to assess the influence of reporting assessment tools including PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) published in 2009 and the risk of bias (RoB) tools including the Newcastle–Ottawa Scale (NOS) in 2007 and the Cochrane Risk of Bias Assessment Tool for Non-Randomized Studies of Interventions (ACROBAT-NRSI) in 2014. We selected studies of kidney disease based on the following two criteria: 1. We included studies on participants with kidney diseases. Kidney diseases were defined as diseases that occurred in the renal parenchyma, such as acute or chronic kidney injury, kidney neoplasms, and nephrolithiasis, based on the MeSH search builder of the term 'Kidney Diseases'. Studies were excluded if they had participants with extra-renal diseases including ureteral, urethral, and urinary bladder diseases. 2. We included studies with primary outcomes related to kidney diseases. We used the same definition of kidney diseases as above. We excluded studies in which kidney diseases were treated as a composite outcome (e.g. composite outcome of kidney, pancreas, and liver cancers)." Comment: detailed eligiblity criteria reported.
Investigator agreement?	Yes	Quote: "Two authors (M.K., K.K.) independently performed full screening to capture the trends and characteristics of systematic reviews of observational studies in the past decade." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Yes	Quote: "The literature searches were conducted in January 2017 using EMBASE and MEDLINE. We searched studies published from January 2006 to December 2016 with no language limitation. The search strategy was developed with the assistance of a medical information specialist and included key words related to 'observational study', 'systematic review', and 'kidney disease' (see Supplement Table 1). Search terms relevant to this review were collected through expert opinion, literature review, controlled vocabulary—including Medical Subject Headings (MeSH) and Excerpta Medica Tree—and a review of the primary search results." Comment: the selective search may have introduced bias by not selecting all available literature.

Kimachi 2021 *(Continued)*

Bias assessed?	Yes	Quote: "In addition, two authors (M.K., A.O.) independently graded each review for overall confidence as high, moderate, low, and critically low using the AMSTAR 2 tool." Comment: assessment of quality of included records was conducted.
Control for differences?	No	Comment: observational studies include cohort studies and case-control studies. These study types were not analysed separately.
Heterogeneity addressed?	Yes	Comment: heterogeneity was noted but not controlled for. Eligibility criteria were strict and likely allowed for a homogeneous sample.
Similar outcomes?	Unclear	Quote: "We included studies with primary outcomes related to kidney diseases. We used the same definition of kidney diseases as above. We excluded studies in which kidney diseases were treated as a composite outcome (e.g. composite outcome of kidney, pancreas, and liver cancers)." Comment: outcomes were selected by topic area but it remains unclear whether all outcomes are sufficiently similar.
No selective reporting	Yes	Comment: no selective reporting suspected; all outcomes reported in methods section are also reported in the results section.
Absence of evidence of bias from other sources?	Yes	Comment: no other forms of bias suspected.

Kirson 2013

Study characteristics	
Methods	Systematic review
Data	5 RCTs (n = 2983) and 8 retrospective and prospective cohort studies (n = 10,577) investigating the efficacy and effectiveness of depot versus oral antipsychotics in schizophrenia. Authors searched from 2000 to 2011 in PubMed (database unclear). **Inclusion criteria:** Full-text was available in English, the study was of human subjects, schizophrenia was the primary disease area investigated, both depot and oral formulations of antipsychotics of the same generation (i.e. first or second generation of antipsychotics) were available, and the publication was not a review article. Only studies that reported findings on relapse, hospitalisation, or all-cause discontinuation. **Exclusion criteria:** Not reported
Comparisons	Ratio of risk ratios and pooled relative risks for RCTs and observational studies extracted
Outcomes	1 outcome of relevance for this review: hospitalisation due to schizophrenia
Notes	Reported results: "We found that observational designs tend to show favorable outcomes for depot therapy, whereas randomized controlled trials tend to find no differences between oral and depot formulations." Funding: research support was provided to Analysis Group, Inc by Otsuka America Pharmaceutical, Inc.

Kirson 2013 *(Continued)*

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "(1) full text was available in English, (2) the study was of human subjects, (3) schizophrenia was the primary disease area investigated, (4) both depot and oral formulations of antipsychotics of the same generation (ie, first or second generation of antipsychotics) were available, and (5) the publication was not a review article." Comment: eligibility criteria relatively broad and briefly described.
Investigator agreement?	Yes	Quote: "Abstract of all of the studies meeting these search criteria were screened by 2 independent researchers. The full manuscripts were retrieved for studies with abstracts that met these criteria. The studies were again reviewed by the 2 reviewers to ensure that all of the above criteria were met on the basis of the full text." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	No	Quote: "The PubMed database was queried for publications between January 1, 2000, and December 31, 2011. The time frame was chosen to reflect research focused on the newer generation of antipsychotic agents." Comment: only one database searched.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	Yes	Quote: "Various sensitivity analyses conducted" Comment: additional stratified analyses for prospective versus retrospective cohort studies were conducted.
Heterogeneity addressed?	Yes	Quote: "Using reported gender and age baseline, we reweighted endpoints to account for demographic differences across treatment arms. Gender was adjusted on the basis of the gender distribution of schizophrenia patients in the general population"; "Adjustment calculations were performed separately for the depot and oral treatments groups." Comment: population characteristics and interventions accounted for in analyses.
Similar outcomes?	Unclear	Quote: "In addition, in order to compute comparable endpoints across studies, only studies that reported findings on relapse, hospitalization, or all-cause discontinuation were included." Comment: all outcomes were analysed in aggregate in one meta-analysis. No clear definition of outcomes reported. Heterogeneity between outcomes suspected.
No selective reporting	Yes	Comment: no selective reporting detected.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Kitsios 2015

Study characteristics	
Methods	Systematic review
Data	58 RCTs (n = 24,096) and 21 propensity-score-adjusted cohort studies (n = 69,012) investigating effect sizes of RCTs and observational studies in various critical care topics. Authors searched MEDLINE (top five critical care journals), reference lists of systematic reviews, the CDSR, PubMed, reference lists, and issued a call to experts for literature published up to July 2012. **Inclusion criteria:** For observational studies: "critically ill adult patients admitted to an intensive care unit - Patients enrolled in study while in the intensive care unit - Any type of intensive care unit for adults (general or subspecialty) considered eligible (Medical, Coronary, Trauma, Surgical, Neurological etc.) - Pharmacological or nonpharmacological (i.e., procedures, ventilator strategies, treatment bundles or protocols) - Timing of delivery of intervention after admission to intensive care unit (so that such interventions could be evaluated in a RCT enrolling an intensive care unit population). - Any comparator treatment considered to be the standard of care by individual studies to allow for the comparison of assessment of the efficacy of the experimental treatment (active therapy or placebo) Mortality as outcome - Short term (intensive care unit, in hospital, 30 day and 90 day) - Long-term (>90 day) Observational studies with a propensity score model build to assess the factors associated with assignment of the intervention of interest." For RCTs: RCT population eligibility criteria as for propensity score studies: "RCT should have examined a population with similar case-mix to the index propensity score study. Similarity assessed on definition of index disease (e.g. Sepsis, cardiac arrest) and enrollment from similar intensive care unit type (e.g. Medical vs. surgical, etc.) - Same pharmacological or nonpharmacological intervention applied in the same clinical setting - Similarity of intervention assessed as follows: (a) for medications, based on dosing schemes and same class of action; (b) for interventions, based on similar protocols and timing of application. - Any comparator treatment considered to be the standard of care by individual studies to allow for the comparison of assessment of the efficacy of the experimental treatment (active therapy or placebo) Mortality as outcome - Short term (intensive care unit, in hospital, 30 day and 90 day) - Long-term (> 90 day) - Data extracted from intention-to-treat analyses, as reported by original RCTs. - RCT as study design." **Exclusion criteria:** "Elective post-operative admissions in post-anesthesia care units or surgical intensive care units e.g., routine admissions after cardiac surgery. - Quasi-randomized trials"
Comparisons	Pooled relative risks for RCTs and observational studies extracted
Outcomes	3 outcomes of relevance for this review: short-term mortality on the intensive care unit, short-term mortality, short-term hospital mortality for various comparisons
Notes	Reported results: "Across diverse critical care topics, propensity score studies published in high-impact journals produced results that were generally consistent with the findings of randomized clinical trials. However, caution is needed when interpreting propensity score studies because occasionally their results contradict those of randomized clinical trials and there is no reliable way to predict disagreements." Funding: "Dr. Dahabreh received contract funding from the Patient-Centered Outcomes Research Institute (PCORI) Methods Research Award (ME1306-03758; Principal Investigator). Dr. Callahan is employed by the Medical University of South Carolina (Pulmonary Critical Care Fellowship Program). Dr. Paulus received support for article research from the National Institutes of Health (NIH). Her institution received grant support from the NIH."

Risk of bias		
Item	**Authors' judgement**	**Support for judgement**

Kitsios 2015 *(Continued)*

Explicit criteria?	Yes	Quote: "We excluded studies conducted in post-anesthesia care units or surgical ICUs with primarily elective postoperative admissions (e.g., routine admissions after cardiac surgery) because such admissions are brief and these patient populations may not reflect the clinical acuity of patients hospitalized in general ICUs. We considered both pharmacological and nonpharmacological interventions, provided that they were administered after admission to an ICU, so that they could be evaluated in an RCT enrolling participants from an ICU population. We considered all-cause mortality as the only outcome of interest, given that other clinical or surrogate outcomes are more prone to measurement error and misclassification." Comment: explicit and detailed criteria for including cohort studies and RCTs are listed in the supplementary material of the publication.
Investigator agreement?	Unclear	Quote: "Two reviewers (G.D.K., S.C.) read potentially eligible studies in full text to determine eligibility" Comment: study selection was conducted by two or more reviewers. Unclear whether the assessment was conducted independently.
Complete sample?	No	Quote: "our search strategy was not exhaustive but was designed to provide an adequate sample of PS studies published in high-impact journals. [...] we did not perform de novo searches and did not use multiple bibliographic databases in tandem to identify relevant studies for all topics". Comment: sample was generated consecutively, but the completeness of the sample is limited.
Bias assessed?	Yes	Quote: "Based on surveys of the methodological features of studies using propensity score methods, we identified a set of items as potentially indicative of the validity of propensity score-based analyses (supplemental document, Supplemental Digital Content 1, http://links.lww.com/CCM/B327). We also evaluated the methodological quality of included RCTs with the Cochrane Risk of Bias tool." Comment: risk of bias of all included studies was assessed systematically, i.e. Cochrane risk of bias tool for RCTs and indicator set (derived from literature) for cohort studies.
Control for differences?	Yes	Quote: "We performed the following sensitivity analyses: 1) we repeated all comparisons by using the single largest study available for each study design (PS or RCT) instead of meta-analysis estimates, and 2) we performed a comparison limited to RCTs enrolling at least 200 participants." Comment: effect sizes were calculated and meta-analysed separately for different study designs, i.e. RCTs and cohort studies.
Heterogeneity addressed?	Yes	Quote: "We systematically matched propensity score studies to randomized clinical trials based on patient selection criteria, interventions, and outcomes." Comment: RCTs and cohort studies were matched based on PICO criteria.
Similar outcomes?	Yes	Comment: mortality was used as outcome in all comparisons.
No selective reporting	Unclear	Comment: the number of participants was not reported. Also, there were some discrepancies in information between the supplementary material and the published paper.

Kitsios 2015 *(Continued)*

Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Kuss 2011

Study characteristics

Methods	Systematic review
Data	51 RCTs and 28 observational studies that employed propensity scores (number of participants not reported) examining effect estimates of meta-analyses of RCTs compared to effect estimates of meta-analyses of propensity score analyses in cardiac surgery. Authors searched MEDLINE up to February 2006. **Inclusion criteria:** Cohort studies: "propensity score analyses comparing off- and on-pump coronary artery bypass graft gave descriptive information on the propensity score study publication and at least one of the 10 short-term binary clinical outcomes [of] death, stroke, myocardial infarction, atrial fibrillation, acute renal failure, inotropic support, red blood cell transfusion, wound infection, reoperation for bleeding, or intra-aortic balloon pump support." "RCTs were included if they gave descriptive information on the RCT study publication and at least one of the binary clinical outcomes mentioned above." **Exclusion criteria:** Not reported.
Comparisons	Pooled odds ratios for RCTs and observational studies and ratios of odds ratios extracted
Outcomes	10 outcomes of relevance for this review: mortality, stroke, myocardial infarction, atrial fibrillation, inotropic support, wound infection, reoperation for bleeding, renal failure, intra-aortic balloon pump support, red blood cell transfusions
Notes	Reported results: For all outcomes, effect estimates from RCTs and propensity score analyses were mostly in agreement. 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. Funding: there was no external funding for this study.

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "RCTs were included if they gave descriptive information on the RCT study publication and at least one of the binary clinical outcomes mentioned above." Comment: the authors included all studies with propensity score analyses comparing off- and on-pump coronary artery bypass graft.
Investigator agreement?	Yes	Quote: "All RCT publications were gathered in full text and read independently by two reviewers (O.K., T.L.)" Comment: study selection was conducted independently by two or more reviewers. Methods for selection of non-RCTs are described in a previous publi-

Kuss 2011 *(Continued)*

		cation that states that the selection of non-RCTs was also conducted by two researchers independently.
Complete sample?	Unclear	Comment: it is possible that RCTs that were not previously identified in systematic reviews may have been missed.
Bias assessed?	Unclear	Quote: "Using some ad hoc measures (data not shown), we found that in the group of RCTs, the subgroup of meta-matched RCTs have a similar study quality, whereas in the group of PS analyses, the meta-matched PS analyses have a higher quality." Comment: not enough information reported about the assessment of risk of bias.
Control for differences?	Yes	Comment: confounder data were extensively collected.
Heterogeneity addressed?	No	Comment: heterogeneity not addressed.
Similar outcomes?	Yes	Quote: "PS analyses were included in the analysis presented here if they gave descriptive information on the PS study publication (e.g., average age, proportion of males etc., factors we will subsequently refer to as ‘‘meta confounders’’) and at least one of the 10 short-term binary clinical outcomes death, stroke, myocardial infarction (MI), atrial fibrillation, acute renal failure, inotropic support, RBC transfusion, wound infection, reoperation for bleeding or, intra-aortic balloon pump support." Comment: each analysis evaluated similar comparisons for disparate outcomes.
No selective reporting	Yes	Comment: no selective reporting suspected; all analyses described in methods section are reported in the results section.
Absence of evidence of bias from other sources?	Yes	Comment: no other biases suspected.

Li 2016

Study characteristics	
Methods	Systematic review
Data	5 RCTs (number of participants not reported) and 10 cohort studies that employed propensity scores (n = 494,964) investigating treatment effect estimates of non-vitamin K antagonist oral anticoagulants versus warfarin. Authors searched up to September 2015 in Cochrane CENTRAL, MEDLINE, Embase, abstract books or websites of four conference proceedings (European Society of Cardiology, American College of Cardiology, the International Society of Thrombosis and Haemostasis, and the American Society of Hematology Annual Meeting) and ClinicalTrials.gov. **Inclusion criteria:** People aged ≥ 18 years with non-valvular atrial fibrillation diagnosis; observational cohort studies using propensity score method to compare non-vitamin K antagonist oral anticoagulants versus warfarin for stroke prevention; phase III RCTs evaluating the efficacy and safety of non-vitamin K antagonist oral anticoagulants versus warfarin **Exclusion criteria:**

Li 2016 *(Continued)*

	Studies whose objectives were not to focus on effectiveness (benefit) or safety (harm) profiles of anticoagulants were excluded. Protocols and reviews that did not provide treatment effect estimates on benefit or safety profiles comparing non-vitamin K antagonist oral anticoagulants versus warfarin in patients with atrial fibrillation, studies which compared non-vitamin K antagonist oral anticoagulants versus warfarin only in patients for ablation or cardioversion of atrial fibrillation, given their short treatment duration and follow-up periods. Studies including non-warfarin vitamin K antagonists where data comparing NOACs and warfarin could not be isolated.
Comparisons	Pooled hazard ratios for RCTs and observational studies extracted
Outcomes	4 outcomes of relevance for this review: myocardial infarction, all-cause mortality, major bleeding, stroke or systemic embolism
Notes	Since there was substantial overlap with evidence included in Gu 2020, Li 2016 was not included in the meta-analyses of this review. Reported results: "No significant difference of treatment effect estimates between the PS studies and RCTs was observed." Funding: not reported

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "Patients aged >18 years with non-valvular AF diagnosis were eligible for inclusion. Observational cohort studies using PS method to compare NOACs versus warfarin for stroke prevention were eligible. Phase III RCTs evaluating the efficacy and safety of NOACs versus warfarin were also included. In this study, the comparisons were limited to NOACs including direct thrombin inhibitors (dabiga tran) and factor Xa inhibitors (rivaroxaban, apixaban, and edoxaban) versus oral warfarin, using warfarin as the comparator." Comment: explicit inclusion and exclusion criteria are reported.
Investigator agreement?	Yes	Quote: "Two reviewers (G.L. and Y.J.) independently screened and selected studies for possible inclusion" Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Yes	Comment: systematic literature search was conducted; explicit inclusion and exclusion criteria were applied
Bias assessed?	Yes	Quote: "For propensity score studies, we used the Cochrane Collaboration ROBINS-I assessment tool to evaluate the study quality. Each included study was rated based on the domains of confounding, selection of participants, classification of interventions, deviations from intended intervention, missing data, measurement of outcomes, and selection of result reporting. For RCTs, we also used the Cochrane Collaboration 'Risk of Bias' assessment tool which included sequence generation, allocation concealment, blinding, incomplete outcome data, selective outcome reporting and other issues, to assess the study quality of each included trial."
Control for differences?	No	Comment: sensitivity analysis with high-quality studies was conducted. RCTs and observational studies were analysed separately; observational studies were not split by study design. A priori sensitivity analyses were performed using a fixed-effects model.

Li 2016 *(Continued)*

Heterogeneity addressed?	Yes	Comment: subgroup analyses were conducted by type of intervention and trial duration.
Similar outcomes?	Yes	Comment: outcome measures were similar in RCTs and cohort studies, including stroke, major bleeding, mortality, myocardial infarction.
No selective reporting	Yes	Quote: "The protocol was registered in the Prospective Register of Ongoing Systematic Reviews (PROSPERO; identifier: CRD42015025940)." Comment: no selective reporting suspected. Analyses reported in the protocol and methods section are also reported in the results section.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Lonjon 2013

Study characteristics	
Methods	Systematic review
Data	94 RCTs and 70 observational studies that employed propensity scores investigating surgical procedures identified (search dates unclear) in MEDLINE **Inclusion criteria:** Non-randomised comparative studies assessing a surgical procedure in humans, with prospective recruitment and data collection (i.e. prospective cohort studies and administrative databases), and involving propensity score analysis with a binary outcome. **Exclusion criteria:** Reports of interventional procedures such as percutaneous coronary intervention or gastrointestinal endoscopy not performed by surgeons.
Comparisons	Ratios of odds ratios extracted
Outcomes	5 outcomes of relevance for this review: all-cause mortality, mortality, stroke, reoperation for bleeding, myocardial infarction
Notes	Reported results: "There was no statistically significant difference in treatment effect between NRSs with PS analysis and RCTs." The authors conclude that RCTs and propensity score analyses will likely yield similar results in surgery studies. Funding: "This study received source from Equipe "Espoirs de la Recherche" par la Fondation pour la Recherche Medicale (FRM)."

Risk of bias		
Item	**Authors' judgement**	**Support for judgement**
Explicit criteria?	Unclear	Quote: "Prespecified eligibility criteria were nonrandomized comparative studies assessing a surgical procedure in humans, with prospective recruitment and data collection (ie, prospective cohort studies and administrative databases), and involving PS analysis with a binary outcome."

Lonjon 2013 *(Continued)*

		Comment: 31 different clinical questions were included, although it is unclear if these questions were conceived a priori.
Investigator agreement?	No	Quote: "One of us, with surgical and methodological expertise, screened the title, abstract and full text of reports to identify eligible studies" Comment: study selection was conducted by one researcher.
Complete sample?	No	Comment: not all RCTs were selected for each research question--restricted to the 5 years preceding the study search.
Bias assessed?	Yes	Quote: "Performance bias, that is, bias due to departures from intended interventions. Because patient blinding is not possible in NRSs, we assessed whether contaminations could bias treatment effect estimates. For this purpose, we determine whether participants remained with the original intervention without contamination by the other intervention and we evaluated whether contamination was sufficiently low to avoid bias. • Detections bias, that is, bias in taking measurements. We rated studies with an objective outcome as having a low risk of bias and studies with a subjective outcome as having high risk of bias unless outcome assessors were blinded. • Attrition bias, that is, bias due to missing data. We rated studies as having low risk of bias if the proportion of missing data was low, with balanced number and reasons for missing data across intervention groups." Comment: performance, detection, and attrition biases were assessed as reported in the results section.
Control for differences?	Yes	Comment: sensitivity analyses performed
Heterogeneity addressed?	Yes	Quote: "The MEDLINE search of related RCTs yielded 4355 citations, from which we selected 94 reports of RCTs matching 70 reports." Comment: RCTs and observational studies were matched.
Similar outcomes?	Yes	Comment: the authors' primary outcome was all-cause mortality, which might allow for sufficient similarity.
No selective reporting	Unclear	Comment: as a result of not including all RCTs, selective reporting might be possible.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias detected.

MacLehose 2000

Study characteristics	
Methods	Systematic review
Data	12 RCTs, 6 cohort studies, and 9 case-control studies (number of participants not reported) examining effect sizes derived from randomised and non-randomised studies. Authors searched the period from 1966 to 1996 in the Cochrane Library, MEDLINE, Embase, Database of Abstracts of Reviews of Effects, and the Science Citation Index, and searched the references of relevant papers already identified, and contacted experts. **Inclusion criteria:**

MacLehose 2000 *(Continued)*

	"The intervention under consideration should be uniform across studies. The outcome by which the intervention under consideration was evaluated should be uniform across studies. The populations in which the intervention was evaluated should be uniform across studies." **Exclusion criteria:** Not reported
Comparisons	Pooled relative risks for RCTs and observational studies extracted
Outcomes	2 outcomes of relevance for this review extracted: mortality from breast cancer, prevention of neural tube defects in women without a previous history of a neural tube defect
Notes	Reported results: "Estimates from RCTs and cohort studies were not significantly different, but case-control studies gave significantly different estimates for both outcomes." Funding: not reported; Health Technology Assessment NHS R&D HTA Programme was the publisher of the report.

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "After further detailed MEDLINE searches the two interventions which best satisfied these criteria were selected from the seven: • mammographic screening (intervention) for women aged 50–64 years (UK guidelines; population) to reduce mortality from breast cancer (outcome) • periconceptional folic acid supplementation (intervention) for women trying to conceive (population) to prevent neural tube defects (outcome)". "Papers were eligible for strategy 2 if they reported primary evaluations of either of the interventions being reviewed, and if they matched the definitions for the intervention, population and outcome described above (see above)". Comment: explicit criteria for selecting studies based on PICO-scheme reported for three different population-intervention-outcome combinations.
Investigator agreement?	Unclear	Quote: "Abstracts of all papers identified by searches were read carefully and the full text of the original paper was obtained for any abstract that appeared relevant." Comment: insufficient information reported.
Complete sample?	Yes	Comment: systematic literature search in electronic databases and duplicate selection of records.
Bias assessed?	Yes	Quote: "The instrument used to assess the quality of a study required the main confounding factors for an intervention to be specified. An additional information sheet (see appendix 7) was therefore circulated with the instrument, providing details of: • the population, intervention and outcome which had been specified for the review (see appendix 4, questions 2, 3 and 4) • the four most common confounding variables (see appendix 4, questions 5 and 25) • up to four previously reported adverse effects of the intervention (see appendix 4, question 8). Confounding factors for each intervention were selected by identifying all the confounding factors included in analyses or considered in the articles reviewed. The four most frequently cited confounding factors were then chosen. A similar approach was taken in identifying possible adverse effects of the interventions."

MacLehose 2000 *(Continued)*

		Comment: risk of bias was assessed together with other quality parameters in a comprehensive instrument and quality scores were calculated.
Control for differences?	Yes	Comment: RCTs, cohort studies and case-control studies were analysed separately.
Heterogeneity addressed?	Unclear	Quote: "Variation in the intervention or exposure was a major source of heterogeneity between studies"; "The number of papers was limited by the strict criteria that we laid down in order to achieve homogeneity of the intervention, population and outcome investigated." Comment: heterogeneity was reduced by applying strict eligibility criteria. Heterogeneity was discussed but apparently not controlled for.
Similar outcomes?	Yes	Quote: "The second strategy was to compare estimates of effectiveness derived from RCT and QEO [quasi-experimental and observational] study designs for interventions for which the intervention, population and outcome investigated were anticipated to be homogeneous across studies." Comment: similar outcomes were used and were already reflected in inclusion criteria.
No selective reporting	Yes	Comment: no selective reporting detected.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Mathes 2021

Study characteristics	
Methods	Systematic review
Data	49 RCTs (n = 198,820) and 24 non-randomised controlled studies based on real-world data (n = 301,340) examining disagreements between real-world-data–based non-randomised controlled studies and randomised controlled trials. Authors searched up to February 2019 in PubMed (database unclear). **Inclusion criteria:** Studies that compared treatment effect estimates from non-randomised controlled studies using real-world data with treatment effect estimates from RCTs. Studies reporting data of non-randomised controlled studies based on real-world data and RCT(s) to assess comparative effectiveness of the same clinical question (same population, intervention, comparator, and outcome). Reporting overall mortality or a disease-specific binary outcome (e.g. disease-specific mortality, recurrence). Reporting data with which treatment effect estimates (risk ratios, odds ratio, or hazard ratio) and 95% confidence intervals for non-randomised controlled studies based on real-world data versus RCTs (i.e. a ratio of ratios) could be calculated. "We only included studies that analyzed a database that includes aggregated data of patient cohorts (registries, administrative/insurance databases). We included only publications written in English and German." **Exclusion criteria:** Studies based on electronic record reviews from a single institution.
Comparisons	Pooled hazard ratios extracted for RCTs and observational studies
Outcomes	3 outcomes of relevance for this review extracted: stroke, venous thromboembolism prophylaxis, and heart diseases (composite); cognitive impairment; atrial fibrillation

Mathes 2021 *(Continued)*

Notes	Reported results: "We found few disagreements that would probably have resulted in a different conclusion regarding the harm/benefit of the intervention in practice. Ninety-five percent confidence intervals overlapped for 12 of 15 treatment effect estimates." Funding: "This research did not receive any specific grant from funding agencies in public, commercial, or not-for-profit sectors."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "We applied the following eligibility criteria: Reporting data of nonrandomised controlled studies based on real-world data and RCT(s) to assess comparative effectiveness of the same clinical question (same PICO). Reporting overall mortality or a disease-specific binary outcome (e.g., disease-specific mortality, recurrence). Reporting data with which treatment effect estimates (risk ratios, odds ratio [OR], or hazard ratio [HR]) and 95% confidence intervals (95% CIs) for nonrandomised controlled studies based on real-world data vs. RCTs (i.e., a ratio of ratios) could be calculated." Comment: inclusion criteria were broad but explicit.
Investigator agreement?	Yes	Quote: "Two independent reviewers performed study selection and discussed all discrepancies until consensus was reached (T.R. and T.M.)." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	No	Quote: "We only included studies that analyzed a database that includes aggregated data of patient cohorts (registries, administrative/insurance databases). We excluded studies that were based on electronic record reviews from a single institution. We included only publications written in English and German. We performed a systematic literature search in PubMed on February 15, 2019." Comment: eligibility criteria were broad, search seems systematic, but was conducted in one database only
Bias assessed?	Yes	Quote: "For analyzing the causes of disagreement related to internal validity (risk of bias) and external validity, we developed a standardized assessment. The risk of bias assessment was derived from the ROBINS-I tool [9]. ROBINS-I guidance defines risk of bias as the ‘‘tendency for study results to differ systematically from the results expected from a randomized trial.’’ Comment: risk of bias was assessed with acceptable tools.
Control for differences?	Yes	Quote: "We used different measures to compare the treatment effect estimates from NRCS-RWDs and RCTs. We counted the number of conflicting effect directions, the number of 95% CIs that did not overlap, and the number of the 95% CIs of NRCS-RWDs that did not include the RCTs point estimate of the treatment effect estimate. In addition, we counted the number of clinical questions where the 95% CI of NRCS-RWDs did not include the null effect but the 95% CI of the RCT did include the null effect." Comment: RCTs and non-randomised studies were included. The sample of non-randomised studies seems to have included only cohort studies.
Heterogeneity addressed?	No	Comment: no measure of heterogeneity reported, assessment of heterogeneity not reported in methods section. Eligibility criteria were relatively broad, hence there might have been heterogeneity in the sample.

Mathes 2021 *(Continued)*

Similar outcomes?	No	Quote: "In half of the reports, we found a difference in outcome assessment or length of follow-up. Differences in missing data were often rated as unclear (5/12) because no information on missing outcome data for NRCS-RWDs was reported. In two of three comparisons, the outcome was "not restricted". Hence, the selection of relevant outcomes might be a variety of outcomes measures." Comment: outcomes seemed to have differed between study types.
No selective reporting	Yes	Comment: no selective reporting suspected; all outcomes reported in methods section are also reported in the results section.
Absence of evidence of bias from other sources?	Yes	Comment: no other forms of bias suspected.

Moneer 2022

Study characteristics	
Methods	Systematic review
Data	37 RCTs and 46 observational studies included (number of participants not reported) examining differences in effect estimates of RCTs and observational studies that investigated the effectiveness of COVID-19-related interventions. Authors searched until February 2021 in National Institutes of Health Covid-19 Treatment Guidelines, a living review and network meta-analysis published in the BMJ, a living systematic review with meta-analysis and trial sequential analysis in PLOS Medicine, and the Epistemonikos “Living OVerview of Evidence” (L·OVE) evidence database. **Inclusion criteria:** "RCTs in The BMJ’s living review that directly compared any of the three most frequently studied therapeutic interventions for covid-19 across all data sources (that is, hydroxychloroquine, lopinavir-ritonavir, or dexamethasone) for any safety and efficacy outcomes"; observational studies that evaluated the same interventions, comparisons, and outcomes that were reported in the BMJ’s living review. **Exclusion criteria:** Not in English; case study reports, case series, or cohort studies with a sample size of < 15; interventional studies or studies that did not include a comparator group; cross-sectional studies or case-control studies that did not evaluate the comparative effectiveness of an intervention
Comparisons	Pooled odds ratios and standard mean differences of RCTs and observational studies extracted
Outcomes	17 outcomes (some outcomes listed here were reported in multiple comparisons) of interest for this review extracted: mortality, mechanical ventilation, viral clearance, duration of hospital stay, time to viral clearance, time to symptom resolution
Notes	Results: "Overall, 21 (78%) of the 27 matched pairs had treatment effects that were in agreement. Among the 17 matched pairs consisting of meta-analyses of observational studies and meta-analyses of RCTs, 14 (82%) were in agreement; seven (70%) of the 10 matched pairs consisting of at least one observational study or one RCT were in agreement. The 18 matched pairs with treatment effects for dichotomous outcomes had a higher proportion of agreement (n=16, 89%) than did the nine matched pairs with treatment effects for continuous outcomes (n=5, 56%)." Funding: "JDW is supported by the National Institute on Alcohol Abuse and Alcoholism of the National Institutes of Health under award K01AA028258. The funders had no role in considering the study design

Moneer 2022 *(Continued)*

or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "To identify observational studies evaluating the same clinical questions as the RCTs included in The BMJ's living review (that is, matched pairs) [...]"; "To identify RCTs for the three interventions, we selected one source among the four sources used to locate the most prominent covid-19 interventions: a living systematic review and network meta analysis on drug treatments for covid-19 published in The BMJ." Comment: eligibility criteria are broad but clearly reported.
Investigator agreement?	Yes	Quote: "The resulting sample included 4774 records, which were imported into Covidence software to remove duplications and be screened by four investigators (OM, GD, JS, and JDW) at the title and abstract level. Two investigators (OM and GD) then evaluated potentially eligible records at the full text level to identify prospective or retrospective observational studies and case-control studies." Comment: study selection was conducted independently by two or more reviewers. RCTs were sampled from published systematic reviews.
Complete sample?	Yes	Comment: the sample seems systematically compiled and complete.
Bias assessed?	Yes	Quote: "For individual RCTs, we abstracted the risk of bias evaluations reported in The BMJ's living review, which were based on a revision of the Cochrane tool for assessing risk of bias in randomized trials (RoB 2.0)"; "For the individual observational studies, two authors (OM and JDW) conducted formal assessments of risk of material bias using the ROBINS-I tool for non-randomized studies." Comment: the review authors assessed risk of bias.
Control for differences?	Yes	Comment: RCTs and observational studies were analysed separately.
Heterogeneity addressed?	Yes	Quote: "Safety and efficacy outcomes from observational studies were identified and treatment effects for dichotomous (odds ratios) or continuous (mean differences or ratios of means) outcomes were calculated and, when possible, meta-analyzed to match the treatment effects from individual RCTs or meta-analyses of RCTs reported in The BMJ's living review with the same interventions, comparisons, and outcomes (that is, matched pairs). The analysis compared the distribution of study demographics and the agreement between treatment effects from matched pairs." Comment: included studies were matched.
Similar outcomes?	Yes	Quote: "To minimize the potential of selecting specific outcomes based on the direction and strength of the treatment effects, we recorded all safety or efficacy outcomes considered by The BMJ's living review"; "At least two individual authors (OM, JJS, GD, and JDW) independently screened and matched individual observational studies to individual RCTs if the observational studies and RCTs considered the same therapeutic intervention, comparator, and outcome measures." Comment: RCTs and observational studies used similar outcome measures.

Moneer 2022 *(Continued)*

No selective reporting	Yes	Comment: no selective reporting suspected.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Morfaw 2021

Study characteristics

Methods	Systematic review
Data	20 RCTs (n = 17,314) and 14 observational studies (n = 56,890) investigating the prevention of postpartum haemorrhage. Authors searched until February 2019 in MEDLINE, Embase, Cochrane CENTRAL, and handsearched the reference lists of relevant studies and previous reviews identified through the literature searches. **Inclusion criteria:** "Types of studies: We included RCTs and non-randomised studies (prospective, retrospective, and cross sectional) that evaluated the use of misoprostol compared to either placebo or no treatment in the prevention of post partum hemorrhage. We also included RCTs and NRS that evaluated the use of misoprostol plus oxytocin versus oxytocin alone in the prevention of PPH; Types of participants: Our study population included pregnant women delivering within hospital settings or in the community, who received misoprostol for the purposes of preventing post partum hemorrhage. Any studies with only a subset of the relevant participants were included and data was specifically extracted only for this subset of women; Type of intervention: We assessed the use of misoprostol given either orally, rectally or sublingually for the prevention of post partum hemorrhage, and this irrespective of the dose used; Comparisons: The control groups were expected to receive standard of care for the prevention of post partum hemorrhage applicable within the settings of the study. In cases where misoprostol was combined with oxytocin, the comparison group must have received an equal dose of oxytocin, and the only difference between the two treatment groups being the addition of misoprostol in the treatment group; Outcomes: The primary outcome was the number of cases of PPH as described by the authors in the different studies, irrespective of the criteria they used in ascertaining the outcome." **Exclusion criteria:** "We excluded studies on women who received misoprostol for induction of labor or for the treatment of confirmed post partum hemorrhage; single arm non-randomised studies."
Comparisons	Odds ratios of RCTs and observational studies extracted
Outcomes	1 outcome of relevance for this review extracted: number of cases of postpartum haemorrhage
Notes	Reported results: "The summary odds ratio (OR) from RCTs for the use of misoprostol in the prevention of post partum hemorrhage was 0.69 (95% confidence interval [CI]: 0.59 to 0.80). The summary OR from NRS was 0.46 (95% CI: 0.36 to 0.63). Classical and Bayesian approaches of combining the two study designs both showed benefit of misoprostol in preventing PPH, with similar effects." Funding: not reported

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "Types of studies: We included RCTs and NRS (prospective, retrospective, and cross-sectional) that evaluated the use of misoprostol compared to either placebo or no treatment in the prevention of PPH [postpartum haemorrhage]. We also included RCTs and NRS that evaluated the use of misoprostol

Morfaw 2021 *(Continued)*

		plus oxytocin versus oxytocin alone in the prevention of PPH. We excluded single arm NRS. Types of participants: Our study population included pregnant women delivering within hospital settings or in the community, who received misoprostol for the purposes of preventing PPH. Any studies with only a subset of the relevant participants were included and data was specifically extracted only for this subset of women. We excluded studies on women who received misoprostol for induction of labor or for the treatment of confirmed PPH. Type of intervention: We assessed the use of misoprostol given either orally, rectally or sublingually for the prevention of PPH, and this irrespective of the dose used. Comparisons: The control groups were expected to receive standard of care for the prevention of PPH applicable within the settings of the study. In cases where misoprostol was combined with oxytocin, the comparison group must have received an equal dose of oxytocin, and the only difference between the two treatment groups being the addition of misoprostol in the treatment group. Outcomes: The primary outcome was the number of cases of PPH as described by the authors in the different studies, irrespective of the criteria they used in ascertaining the outcome." Comment: eligibility criteria are clearly listed.
Investigator agreement?	Yes	Quote: "Two review authors (FM and BM) independently screened the titles and abstracts of the studies identified through the electronic searches in order to identify possible articles for inclusion, while excluding duplicates. Following this screening, the full texts of eligible articles were obtained and assessed by both reviewers based on the inclusion criteria cited above." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Yes	Comment: the sample seems systematically compiled and complete.
Bias assessed?	Yes	Quote: "Assessment of risk of bias in RCTs was done using the criteria outlined in the Cochrane Handbook for Systematic Reviews of Interventions. The risk of bias table was completed for each outcome by two review authors (FM and BM) working independently of each other. Studies were rated as being at either “high,” “low,” or “unclear” risk of bias. As much as possible, we avoided the term “unclear” in describing the risk of bias, except in the rare situations when the review authors could not make any judgment. We used the “Risk Of Bias In Non-randomized Studies—of Interventions” (ROBINS—I) tool to asses risk of bias in the NRS. Studies were rated as being at either “Low risk,” “Moderate risk,” “Serious risk,” and “Critical risk” of bias. For the purposes of this review, we merged the last three categories into a single category of “high risk of bias” for ease of comparability with the RCTs. We resolved any discrepancies in risk of bias by discussion or by consultation with a third author (LM)" Comment: the risk of bias assessment was conducted with the help of the Cochrane Risk of bias tool and ROBINS-I for RCTs and observational studies, respectively.
Control for differences?	Yes	Quote: "A sensitivity analysis for assessment of discrepancy between the two study designs was done using the difference in magnitude of treatment effect. This was done by assessing the ratio of the odds ratio with the threshold of discrepancy defined as the OR of the RCT being at least twice or less than half the odds ratio of the non-randomised studies. (page 201). We explored varied approaches of combining evidence from RCTs with that of NRS on this topic, each with its merits and demerits. Intuitively, a direct pooling of treatment effects using a classical approach makes us wonder whether we are not actually

Morfaw 2021 *(Continued)*

		'mixing apples and oranges'."; "This approach has the potential drawback of assigning more weight to the observational studies given their larger sample sizes. However, by comparing different methods of pooling the two study designs together, we were able to better assess the robustness of our conclusions" (page 205). Comment: RCTs and observational studies were analysed separately.
Heterogeneity addressed?	Yes	Comment: heterogeneity was clearly assessed and reported.
Similar outcomes?	Yes	Quote: "The primary outcome was the number of cases of PPH as described by the authors in the different studies, irrespective of the criteria they used in ascertaining the outcome" Comment: RCTs and observational studies used similar outcome measures.
No selective reporting	Yes	Comment: no selective reporting suspected.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Müeller 2010

Study characteristics	
Methods	Systematic review
Data	26 RCTs and 136 observational studies (n not reported) examining effect sizes in RCTs and non-randomised trials in cholecystectomy. Authors searched between 1993 and 2008 in PubMed (database unclear). **Inclusion criteria:** Studies that reported variables important for patients, such as mortality and parameters of morbidity, pain, health-related quality of life, return to work, length of hospital stay, and operative time, which are important for hospitals as well. Articles had to be written in English, German, French, Spanish, Italian, or Dutch. RCTs and non-RCTs (with concurrent or historical controls) comparing any kind of laparoscopic cholecystectomy with any kind of open cholecystectomy, including small-incision cholecystectomy, were included. **Exclusion criteria:** No abstract available, no clinical results (as defined later) were reported in the abstract; study results were based on external controls (i.e. literature controls); studies with selected participants, such as children or cirrhotic patients; meta-analyses, reviews, cohort studies without controls, and case series (less than 10 participants in one of the groups)
Comparisons	Relative risks of RCTs and non-randomised studies were extracted
Outcomes	9 outcomes of relevance extracted for this review: overall complications, wound infection, respiratory disorder, urological disorders, bile duct injury, bile leak, pneumonia, open cholecystectomy, urinary infection
Notes	Reported results: significant discrepancies between RCT- and non-RCT-based results were revealed for 3 of 15 variables. Funding: not reported.

Müeller 2010 *(Continued)*

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "studies (including registries) that fulfilled the following inclusion criteria were selected for detailed assessment: studies that reported variables important for patients, such as mortality and parameters of morbidity, pain, health-related quality of life, return to work, length of hospital stay, and operative time, which are important for hospitals as well. To be included, articles had to be written in English, German, French, Spanish, Italian, or Dutch. RCTs and nRCTs (with concurrent or historical controls) comparing any kind of LC with any kind of OC, including small-incision cholecystectomy, were included." Comment: identified RCTs and observational studies (cohorts) on a specific topic.
Investigator agreement?	Unclear	Comment: no information reported.
Complete sample?	No	Quote: "A combined literature search in the Medline database was performed to select both RCTs and nRCTS". Comment: only one database searched.
Bias assessed?	Yes	Quote: "Therefore, methodological quality parameters, such as the performance of randomization, blinding, and intention to treat (ITT) were categorized as adequate, nonadequate, or unclear. In nRCTs, patients were assumed to be analyzed adequately according to ITT if the numbers of converted patients or dropouts were listed." Comment: Cochrane RoB criteria plus additional criteria assessed.
Control for differences?	Yes	Comment: heterogeneity was addressed through sensitivity analysis.
Heterogeneity addressed?	Yes	Comment: sensitivity analysis
Similar outcomes?	Yes	Quote: "Because in most studies, even among the RCTs, a primary outcome criterion is not defined, studies (including registries) that fulfilled the following inclusion criteria were selected for detailed assessment: studies that reported variables important for patients, such as mortality and parameters of morbidity, pain, health-related quality of life, return to work, length of hospital stay, and operative time, which are important for hospitals as well." Comment: included studies with different outcomes, analysed by outcome.
No selective reporting	Yes	Comment: no risk of selective reporting suspected.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected

Naudet 2011

Study characteristics

Methods	Systematic review

Naudet 2011 *(Continued)*

Data	109 RCTs (n = 11,035) and 12 observational studies (n = 6757) examining antidepressant response in major depressive disorder (MDD). Authors searched between January 1989 and July 2009 in MEDLINE, the Cochrane library, Embase, clinicaltrials.gov, Current Controlled Trial, bibliographies, and by mailing key organisations and researchers. **Inclusion criteria:** "Types of participants. In the main analysis, we reviewed studies involving adults with a diagnosis of MDD (DSM IV, DSM IV-R, DSM III, DSM III-R, ICD 10, Feighner criteria, Research Diagnostic Criteria). Studies involving patients with other psychiatric or medical comorbidities were considered, except if these comorbidities were an explicit inclusion criterion for the study. Studies involving more than 20% bipolar disorder were excluded, as were studies exclusively involving elderly patients or patients with seasonal affective disorder, post partum depression, postmenopausal depression, atypical depression. As in ‘‘real-life’’ a wide range of depressive disorders is treated with antidepressants, a second analysis included studies involving patients with a diagnosis of anxious depression (criteria for both an anxious disorder and MDD) and/or minor depressive episode and/or dysthymia. Types of intervention. We focused our attention on fluoxetine and venlafaxine in oral mono-therapy for MDD firstline treatment. By choosing these two antidepressants, which are widely used, we were sure to have a large number of RCTs and observational studies. Types of outcome. The primary outcome measure was the difference between baseline and last assessment on the 17-item or 21-item Hamilton Rating Scale for Depression (HRSD) or the Montgomery and Asberg Rating Scale (MADRS). Studies not providing the desired information on these scales were included in the qualitative review. Types of study. In this review the studies considered were those designed to measure antidepressant efficacy or effectiveness, conducted between January 1989 and July 2009: on the one hand RCTs (antidepressant versus placebo or active treatment) and on the another hand observational cohorts (longitudinal nonrandomized and non-blinded studies). Studies designed to provide evidence on other issues such as physiological hypotheses were not retained. Only study reports in English, French and Spanish language were considered." **Exclusion criteria:** Studies involving more than 20% bipolar disorder; studies exclusively involving elderly patients or patients with seasonal affective disorder, postpartum depression, postmenopausal depression, atypical depression; studies designed to provide evidence on other issues.
Comparisons	Correlation coefficients for the outcome were extracted from RCTs, with observational studies as the reference group
Outcomes	1 outcome of interest for this review: difference between baseline and last assessment on the 17-item or 21-item Hamilton Rating Scale for Depression (HRSD) or the Montgomery and Asberg Rating Scale
Notes	Since Naudet 2011 reported continuous outcomes, the study did not contribute to the primary meta-analysis. Reported results: "Response to antidepressants is greater in RCTs than in observational studies." Funding: "This paper was supported by the Institut National de la Sante et de la Recherche Médicale (INSERM). The funding sources had no role in the design and conduct of the study, in the collection, analysis, and interpretation of the data, or in the preparation, review, or approval of the manuscript."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "Types of participants. In the main analysis, we reviewed studies involving adults with a diagnosis of MDD (DSM IV, DSM IV-R, DSM III, DSM III-R,

Naudet 2011 *(Continued)*

		ICD 10, Feighner criteria, Research Diagnostic Criteria). Studies involving patients with other psychiatric or medical comorbidities were considered, except if these comorbidities were an explicit inclusion criterion for the study. Studies involving more than 20% bipolar disorder were excluded, as were studies exclusively involving elderly patients or patients with seasonal affective disorder, post partum depression, postmenopausal depression, atypical depression. As in ''real-life'' a wide range of depressive disorders is treated with antidepressants, a second analysis included studies involving patients with a diagnosis of anxious depression (criteria for both an anxious disorder and MDD) and/or minor depressive episode and/or dysthymia. Types of intervention. We focused our attention on fluoxetine and venlafaxine in oral mono-therapy for MDD first line treatment. By choosing these two antidepressants, which are widely used, we were sure to have a large number of RCTs and observational studies. Types of outcome. The primary outcome measure was the difference between baseline and last assessment on the 17-item or 21-item Hamilton Rating Scale for Depression (HRSD) or the Montgomery and Asberg Rating Scale (MADRS). Studies not providing the desired information on these scales were included in the qualitative review. Types of study. In this review the studies considered were those designed to measure antidepressant efficacy or effectiveness, conducted between January 1989 and July 2009: on the one hand RCTs (antidepressant versus placebo or active treatment) and on the another hand observational cohorts (longitudinal non randomized and non-blinded studies). Studies designed to provide evidence on other issues such as physiological hypotheses were not retained. Only study reports in English, French and Spanish language were considered." Comment: the authors specified eligibility criteria based on PICO criteria.
Investigator agreement?	Yes	Quote: "Eligibility assessment was performed independently in blinded standardized manner by 2 reviewers." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Yes	Quote: "Eligible studies were identified from Pubmed/Medline, the Cochrane library, and Embase, including conference abstracts. In a first step, an initial search on Medline was undertaken to determine optimal keywords and include possible changes in the databases. The keywords used were double-checked before starting the main search. In a second step all identified keywords were used to search all the databases mentioned above. A third search was undertaken on the bibliographies of identified articles and previous meta-analyses." Comment: searched for all studies on a specific topic; seems thorough.
Bias assessed?	Yes	Quote: "Each paper was then assessed for methodological quality prior to inclusion in the review, using two appropriate standardized critical appraisal instruments [14], one for RCTs and one for observational studies (Appendix S1)." Comment: yes: authors used different instruments for RCTs and observational studies.
Control for differences?	Yes	Comment: some RoB items included in meta-regression; also did sensitivity analysis according to risk of bias
Heterogeneity addressed?	Yes	Comment: heterogeneity assessed in meta-regression.
Similar outcomes?	Yes	Quote: "The primary outcome measure was the difference between baseline and last assessment on the 17-item or 21-item Hamilton Rating Scale for Depression (HRSD) or the Montgomery and Asberg Rating Scale (MADRS). Stud-

Naudet 2011 *(Continued)*

		ies not providing the desired information on these scales were included in the qualitative review." Comment: two outcome measurement tools used that were converted to standardised scores.
No selective reporting	Unclear	Comment: limited evidence of publication bias based on funnel plots.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Otsuka 2022

Study characteristics	
Methods	Systematic review
Data	5 RCTs with 3690 participants, 10 cohort studies with 1779 participants, and 8 case-control studies with 9268 participants retrieved through searches in PubMed, Cochrane CENTRAL, and Web of Science from inception to July 2021 **Inclusion criteria:** RCTs, case-matched studies, or cohort studies; studies that compared laparoscopic distal gastrectomy versus open distal gastrectomy for advanced gastric cancer; studies that provided available outcome data; and articles written in English. **Exclusion criteria:** Studies without appropriate data; laboratory or animal studies; and papers identified as letters, comments, correspondence, editorials, or reviews
Comparisons	Pooled mean differences, odds ratios and hazard ratios were extracted for each outcome separately for RCTs and observational studies.
Outcomes	8 outcomes of relevance for this review: operative time, intraoperative blood loss, postoperative hospital stay, number of retrieved lymph nodes, postoperative complications, recurrence, 3-year disease-free survival, 3-year overall survival
Notes	Results: "There was no difference in estimated treatment effects between RCTs and case-matched studies for all outcomes except for the number of retrieved lymph nodes and postoperative complications. In terms of intraoperative blood loss, postoperative hospital stay, number of retrieved lymph nodes, and recurrence, observational studies tended to overestimate the treatment effects." Funding: not reported

Risk of bias		
Item	**Authors' judgement**	**Support for judgement**
Explicit criteria?	Yes	Quote: "The inclusion criteria were as follows: (1) RCTs, case matched studies, or cohort studies; (2) studies that compared LDG versus ODG for AGC; (3) studies that provided avail able outcome data; and (4) articles written in English. The exclusion criteria were as follows: (1) studies with out appropriate data; (2) laboratory or animal studies; and (3) papers identifed as letters, comments, correspondence, editorials, or reviews."

Otsuka 2022 *(Continued)*

		Comment: the PICO was specified in detail and sufficient eligibility criteria reported.
Investigator agreement?	Yes	Quote: "Two authors (R.O. and Y.M.) independently reviewed the title and abstract of articles after eliminating duplicates. The same authors then evaluated the full text according to the study eligibility criteria described below. In cases of disagreement, the authors discussed or consulted a third author until agreement was reached." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Yes	Comment: the sample seems systematically compiled and complete.
Bias assessed?	Yes	Quote: "The risk of bias assessed using the revised Cochrane risk-of-bias tool is shown in Table 2. For overall risk-of-bias judgement, all included RCTs were rated as low risk of bias. The quality of the included observational studies was assessed using the Newcastle–Ottawa quality assessment scale, and all studies were graded as a high quality (Table 3). In addition, we conducted a funnel plot analysis to assess the possibility of a publication bias (Fig. 2). The spread of the distribution of the effect sizes of the studies in the funnel plot was more pronounced in observational studies than in others." Comment: risk of bias was assessed.
Control for differences?	Yes	Comment: study types were analysed separately.
Heterogeneity addressed?	No	Comment: no information reported on heterogeneity in populations or interventions.
Similar outcomes?	Yes	Quote: "The following data were extracted: population characteristics (year of publication, study design, country in which the study was performed, number of patients), short-term outcome parameters (operative time, intraoperative blood loss, postoperative hospital stay, retrieved lymph nodes, postoperative complications), and long term outcome parameters (recurrence, 3-year disease free survival (DFS), 3-year overall survival (OS))." Comment: similar outcome measures for each outcome were used.
No selective reporting	Yes	Comment: listed outcomes were reported in subgroup analysis.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Papanikolaou 2006

Study characteristics	
Methods	Systematic review
Data	167 RCTs and 25 observational studies (number of participants not reported) examining evidence on harms of medical interventions in randomised and non-randomised studies were retrieved through searches in the CDSR and MEDLINE up to October 2004. **Inclusion criteria:** "All nonrandomized controlled studies in which the comparison (intervention versus no treatment, or intervention versus other intervention) was similar to that in the respective randomized trial(s), with

Papanikolaou 2006 *(Continued)*

	"no treatment" corresponding to "placebo." Nonrandomized studies that had selected participants with the same, overlapping or wider indications for the intervention as those used in the respective randomized trials." **Exclusion criteria:** "Noncontrolled studies (the absolute and relative risk conferred by the intervention per se cannot be estimated), unless the specific harm was so rare in the control population that randomized trials had recorded no events in control subjects (in which case the absolute risk among treated subjects would still be meaningful to compare between study designs). Nonrandomized studies in which the indications differed entirely from those in the randomized trial populations."
Comparisons	Pooled risk ratios were extracted for each outcome separately for RCTs and observational studies.
Outcomes	9 outcomes of relevance for this review: convulsions, hypotonic hyporesponsiveness, major extracranial bleed (with anticoagulant therapy or antiplatelet therapy), symptomatic intracranial bleed, visceral or vascular injury, wound infection, spontaneous miscarriage, multiple gestation, major bleed
Notes	Data reported in this review were included in Golder 2011 and consequently were not included in our quantitative analysis. Reported results: "There was no clear predilection for randomized or nonrandomized studies to estimate greater relative risks, but usually (75% [6/8]) the randomized trials estimated larger absolute excess risks of harm than the nonrandomized studies did." Funding: not reported

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "We included nonrandomized studies that had selected participants with the same, overlapping or wider indications for the intervention as those used in the respective randomized trials; we excluded nonrandomized studies in which the indications differed entirely from those in the randomized trial populations." Comment: authors matched observational studies to published RCTs on particular topics
Investigator agreement?	Unclear	Quote: "Two of us (P.P. and G.C.) searched MEDLINE (PubMed) independently (last search October 2004) for qualifying non randomized studies that would correspond to each of the 66 harms." Comment: unclear if titles and abstacts were screened independently and whether assessment was made by two reviewers.
Complete sample?	No	Comment: unclear whether authors were able to match observational studies to all the RCTs.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	No	Comment: not done.
Heterogeneity addressed?	Unclear	Quote: "Differences in data between observational and randomized trials may be due in part to differences in study populations; however, this information is often difficult to dissect, and important details about patient populations may not be transparent in published reports." Comment: heterogeneity discussed but not controlled for.

Papanikolaou 2006 *(Continued)*

Similar outcomes?	No	Comment: "harms" broadly defined; could include multiple outcomes.
No selective reporting	Yes	Comment: no selective reporting suspected.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Safieddine 2021

Study characteristics	
Methods	Systematic review
Data	6 RCTs (n = 27,121) and 20 observational studies (n = 248,971) examining venous thromboembolism patients receiving direct oral anticoagulants or conventional treatment. Authors searched between January 2009 and August 2020 in PubMed, Cochrane Library (not further specified), and Google Scholar. **Inclusion criteria:** "Studies had to meet all the following criteria: (I) pivotal phase III RCTs, as well as observational studies, (2) comparing direct oral anticoagulants vs. vitamin-K antagonists administered for at least 3 months for the initial treatment of venous thromboembolisms. The comparisons were limited to direct oral anticoagulants approved or under regulatory review for the treatment of venous thromboembolisms, namely direct thrombin inhibitors (dabigatran) and factor Xa inhibitors (rivaroxaban, apixaban, and edoxaban) vs. oral warfarin or vitamin-K antagonists as the comparator. Descriptions of the observational studies. Three subcategories of observational studies were defined a priori in the internal protocol: (1) prospective cohort studies comprising observational studies with prospective recruitment and construction of a specific database for the study, (2) studies using living databases, comprising observational studies with prospective recruitment included in a living clinical database (i.e., a registry) or a living health administrative database, and (3) retrospective cohort studies with retrospective recruitment." **Exclusion criteria:** When duplicates were identified, the study with the largest sample size was selected. Studies with objectives not focusing on the effectiveness or safety profiles of anticoagulants, as well as protocols and reviews that did not provide estimates of treatment effects in terms of efficacy or safety profiles.
Comparisons	Pooled hazard ratios were extracted for each outcome
Outcomes	2 outcomes of relevance for this review: recurrent venous thromboembolism and major bleeding
Notes	Reported results: "In this clinical setting, an exaggeration of the treatment efficacy estimate was seen with observational studies compared with RCTs. Among observational studies, prospective cohort studies yielded the most favorable estimates of treatment effect compared with RCTs." Funding: scholarship from Tyr municipality, Lebanon.

Risk of bias

Item	**Authors' judgement**	**Support for judgement**
Explicit criteria?	Yes	Quote: "To be included in the metaanalysis, studies had to meet all the following criteria: (I) pivotal phase III RCTs, as well as observational studies, (2) comparing direct oral anticoagulants vs. vitamin-K antagonists administered for at least 3 months for the initial treatment of venous thromboembolism. The comparisons were limited to direct oral anticoagulants approved or under reg-

Safieddine 2021 *(Continued)*

		ulatory review for the treatment of venous thromboembolism, namely direct thrombin inhibitors (dabigatran) and factor Xa inhibitors (rivaroxaban, apixaban, and edoxaban) vs. oral warfarin or vitamin-K antagonists as the comparator. Three subcategories of observational studies were defined a priori. Studies with objectives not focusing on the effectiveness or safety profiles of anticoagulants, as well as protocols and reviews that did not provide estimates of treatment effects in terms of efficacy or safety profiles, were excluded." Comment: detailed eligibility criteria are reported
Investigator agreement?	Yes	Quote: "Two of the authors (MS and SL) independently evaluated studies for potential inclusion, disagreements being resolved by discussion." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Yes	Comment: eligibility criteria were sufficiently detailed, search seems systematic, hence a complete consecutive sample can be expected.
Bias assessed?	Unclear	Comment: no information on conducting risk of bias reported.
Control for differences?	Yes	Comment: sensitivity analyses for both outcomes (recurrent venous thromboembolism (VTE) and major bleeding) were performed by excluding the observational studies that did not involve the use of any methods of adjusting for confounders in the analysis phase. RCTs, cohort studies were analysed separately; cohort studies were further separated by prospective, retrospective, and database-based cohort studies.
Heterogeneity addressed?	Yes	Comment: altough I^2 was not reported, the review seems to have controlled for methodological differences between the studies by separating RCTs and cohort studies and sub-categorising cohort studies in 3 categories. Eligibility criteria were relatively narrow which might have allowed for a homogeneous sample across studies.
Similar outcomes?	Yes	Quote: "The primary efficacy outcome was used as the primary endpoint in all randomized studies evaluating direct oral anticoagulants in this indication, that is, recurrence of venous thromboembolisms. This efficacy outcome is a composite endpoint, including fatal pulmonary embolism (PE), nonfatal symptomatic pulmonary embolism, and symptomatic recurrence of deep-vein thrombosis (DVT). The main safety outcome was major bleeding according to the definition issued by the International Society on Thrombosis and Hemostasis and including fatal bleeding, bleeding causing a fall in hemoglobin level of 20 g/L or more or leading to transfusion of two or more units of whole blood or red blood cells, symptomatic bleeding in a critical area or organ (such as intracranial, intraocular, intraspinal, retroperitoneal, intra-articular or pericardia!, or intramuscular with compartment syndrome) [7]." Comment: outcome measures seem similar in all study types.
No selective reporting	Yes	Comment: no selective reporting suspected; all outcomes reported in methods section are also reported in the results section.
Absence of evidence of bias from other sources?	Yes	Comment: no other forms of bias suspected.

Schmidt 2013

Study characteristics	
Methods	Systematic review
Data	17 RCTs (n = 469,447), 7 cohort studies (n = 279,639), and 6 case-control studies (n = 94,895) examining the effect of mammography screening on mortality, and the effect of coronary bypass or statins on mortality. Searches were conducted in the register of the "IPD Cochrane Methods Group", MEDLINE, CENTRAL, and Scopus; search period unclear. **Inclusion criteria:** Studies investigating a similar PICO as previously identified; individual patient data meta-analysis; investigated similar subgroup specifics to allow for direct comparisons of treatment effects; allow calculation of point estimates and confidence interval of treatment effect; used an RCT or cohort or case-control design; in English language **Exclusion criteria:** Not reported
Comparisons	Pooled risk ratios were extracted for each outcome
Outcomes	1 outcome of relevance for this review: mortality
Notes	Reported results: "Main and subgroup-specific effects based on reported observational data were similar in direction to those from individual patient data meta-analyses." Funding: "The Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, has received unrestricted research funding from the Netherlands Organisation for Health Research and Development (ZonMW), Dutch Health Care Insurance Board (CVZ), Royal Dutch Pharmacists Association (KNMP), private-public funded Top Institute Pharma (www.tiphar-ma.nl, includes cofunding from universities, government, and industry), EU Innovative Medicines Initiative (IMI), EU 7th Framework Program (FP7), Dutch Medicines Evaluation Board, and Dutch Ministry of Health and industry (including GlaxoSmithKline, Pfizer, and others)."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "RCTs and observational studies were included when they (1) investigated similar patients, interventions, and outcomes as the IPDMA; (2) investigated similar subgroup-specifics that allowed for direct comparison of treatment effects; (3) allowed calculation of point estimates and CI of the treatments effects; (4) used an RCT, cohort study, or case-control design; and (5) were written in English". Comment: the inclusion criteria for each example are very broad.
Investigator agreement?	Unclear	Comment: no information reported.
Complete sample?	Unclear	Comment: the study sampling seems consecutive, based on the literature search and the eligibility criteria. However, eligibility criteria not detailed enough to allow repetition of study.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	No	Comment: no sensitivity analyses reported; cohort and case-control studies are collapsed as observational studies. There was larger heterogeneity in interventions and comparators for one outcome of the study.

Schmidt 2013 *(Continued)*

Heterogeneity addressed?	Yes	Quote: "follow-up duration ranges were 8.8 to 18 years in the mammography example, 5.6 to 10.4 years in the CABG [coronary artery bypass graft] example, and 0.5 to 8 years in the statin example. Furthermore, although treatment and outcomes were very similar in the mammography and coronary artery bypass graft examples (Appendix B at www.jclinepi.com), in the statin example, RCTs used placebo or active comparison groups, whereas observational studies used no or diminished treatment adherence as a comparator group." Comment: the review performed prespecified subgroup analyses.
Similar outcomes?	Unclear	Comment: mortality and cardiovascular endpoints were selected as outcomes in the included studies. These hard outcomes should allow for sufficient similarity in definition.
No selective reporting	Yes	Comment: deviations from the initial study protocol are reported transparently. All analyses described in the methods section are reported in the results section.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias detected.

Schwingshackl 2021

Study characteristics	
Methods	Overview of systematic reviews
Data	33 systematic reviews of RCTs and 46 systematic reviews of cohort studies evaluating agreement between bodies of evidence from randomised controlled trials and cohort studies in nutrition research. Authors searched between January 2010 and December 2019 in the CDSR and MEDLINE. **Inclusion criteria:** "General population; Intervention/Exposure: a. Dietary pattern: e.g., mediterranean diet, Dietary approaches to Stop Hypertension, low-carbohydrate diet. b. Food groups: the following food groups (macro-level), and foods (micro-level): E.g. grains, vegetables, fruit, milk and dairy products, meat, processed meat, fish, eggs, nuts, chocolate, oils were considered. c. Macronutrients: Carbohydrate (starch, fructose, glucose, sucrose); fat: e.g. n–3 fatty acids (EPA, DHA, a-linolenic acid); n–6 fatty acids (linoleic acid); monounsaturated fat; protein (e.g. amino acids). d. Micronutrients: Vitamins: beta-carotene; vitamins A, E, C (ascorbic acid), and D (cholecalciferol, ergocalciferol); B vitamins (thiamine, riboflavin, niacin, pyridoxine, cobalamin, folic acid. Minerals: magnesium, calcium, selenium, sodium, potassium, iron, zinc, copper, iodine. e. Other: Fibre (psyllium, inulin, cellulose); probiotics; prebiotics; and synbiotics. Control/Comparison: a. Low (no) intake (status) level of the above interventions/exposure. b. Placebo/Usual care. Outcomes: E.g. all-cause mortality, cardiovascular disease, coronary heart disease (myocardial infarction, ischemic heart disease, and acute coronary syndrome), stroke, cancer, type 2 diabetes, dementia, fractures, age-related macular degeneration, anthropometric outcomes; important intermediate disease markers such systolic blood pressure, and diastolic blood pressure, fasting glucose, and LDL-cholesterol. Study design: a. Systematic reviews of randomized controlled trials. b. Matching systematic reviews of cohort studies: cohort studies (if available prospective cohort studies were preferred)." **Exclusion criteria:** "In cases where a body of evidence reported effect estimates based on a pool of studies of variable design (i.e. case-control, cross-sectional studies, retrospective cohort studies, or quasi RCTs), the pooled effect estimates by excluding non-cohort/non-RCT studies were recalculated, while retaining the CSs/

Schwingshackl 2021 *(Continued)*

	RCTs fulfilling our inclusion criteria. Moreover, where a BoE reported effect estimates based on either "dietary intake and dietary supplements", "nutrient status (e.g. plasma selenium status) and dietary intake" or "nutrient status and dietary supplements" effect estimates whenever feasible to improve comparability between exposures in controlled studies and interventions in RCTs were recalculated. For example, in case a meta-analysis of RCTs investigated the effect of "selenium supplements", if the authors of the matched meta-analysis of controlled studies mixed "plasma selenium status" and "selenium supplements"; studies with "plasma selenium status" and recalculated the effect estimates based on "selenium supplements" only."
Comparisons	Ratio of risk ratios were extracted
Outcomes	1 outcome relevant for this review: composite outcome
Notes	Reported results: "On average, the difference in pooled results between estimates from bodies of evidence from RCTs and bodies of evidence from cohort studies was small." Funding: "Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 459430615 and Forschungskommission der Medizinischen Fakultät Freiburg"

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Comment: PI/ECO stated in Table 1. Systematic reviews of RCTs and matching systematic reviews of cohort studies published between 2010 and 2019 were searched for in the Cochrane Database for Systematic Reviews and MEDLINE.
Investigator agreement?	No	Quote: "Screening of titles/abstracts was done by one reviewer (LS), and was followed by a screening for inclusion of relevant full texts by two reviewers independently (LS, JZ)." Comment: title and abstract screening conducted by one reviewer; only full-text screening was conducted in duplicate.
Complete sample?	Yes	Comment: studies were searched systematically and explicit criteria for study selection were described. Studies in Cochrane Database for Systematic Reviews and MEDLINE published between 2010 and 2019. Flow charts reported.
Bias assessed?	No	Comment: methodological quality of included SRs not evaluated (e.g. AMSTAR/ROBIS). However, information on risk of bias or study quality of primary studies extracted.
Control for differences?	Yes	Quote: "We conducted a-priori planned subgroup analyses: type of dietary intervention/exposure, outcome, and PI/ECO similarity degree ("more or less identical", "similar but not identical", and "broadly similar"). We conducted two post-hoc sensitivity analyses excluding highly correlated outcomes. First, a very conservative sensitivity analysis was performed in which we included only one outcome per comparison (i.e. the outcome with the largest number of RCTs) from each Cochrane Review. Second, a sensitivity analysis, in which we included outcomes based on their ranking in the SoF tables in the identified Cochrane Reviews (from top to bottom). For example, for the intervention/exposure α-linolenic acid the outcomes coronary heart disease, cardiovascular disease, and cardiovascular disease mortality are likely highly correlated." Comment: differences accounted for in subgroup analyses.
Heterogeneity addressed?	Yes	Quote: "The pooled estimates were obtained through a random-effects meta-analysis model. We assessed heterogeneity through the I^2 and τ^2 statistics. The τ^2 was estimated by the Paule and Mandel method, which is the recom-

Schwingshackl 2021 *(Continued)*

		mended method for binary outcomes and performs well also with continuous ones. Furthermore, the 95% prediction intervals were obtained, in order to show the range of possible values for the difference in the results between bodies of evidence of RCTs and bodies of evidence of cohort studies that might be observed in future comparisons. All the meta-analyses were performed using the R package meta."
Similar outcomes?	Yes	Quote: "We conducted a priori planned subgroup analyses: type of dietary intervention or exposure, outcome, and PI/ECO similarity degree." Comment: mainly objective outcomes included (mortality, cardiovascular disease, cancer); RCTs and observational studies were matched according to PICO criteria. The review includes sensitivity analyses according to the degree of similarity of PICO criteria.
No selective reporting	Yes	Comment: no selective reporting suspected.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Shen 2020

Study characteristics	
Methods	Systematic review
Data	5 RCTs (n = 28,152) and 27 observational studies (n = 519,267) examining direct oral anticoagulants and vitamin-K antagonists in the elderly with atrial fibrillation. Authors searched up to July 2019 in MEDLINE, Embase, and the Cochrane Library (not further specified). **Inclusion criteria:** RCTs or observational studies; including elderly patients (≥ 75 years) with atrial fibrillation; compared direct oral anticoagulants with vitamin-K antagonists (warfarin, phenprocoumon et al.); and reported benefits and harmful outcomes. "For the highest quality observational studies, only nationwide or health insurance database studies that reported adjusted or matched data using an authorized method to minimize confounding [covariate adjustment, propensity score adjustment, propensity score matching, inverse probability of treatment weighting] were included. If multiple observational studies from the same data source were identified, the one that reported adjusted data with the longest study period was used." **Exclusion criteria:** Studies that reported only crude results or were published only in a conference abstract or letter.
Comparisons	Pooled risk ratios and hazard ratios were extracted for each outcome
Outcomes	5 outcomes relevant for this review: stroke, intracranial haemorrhage, major bleeding, gastrointestinal bleeding, all-cause mortality
Notes	Reported results: "No significant difference in treatment effect estimates was found between 27 observational studies and 5 RCTs." Funding: This study was supported by the National Natural Science Foundation of China (71974137), Program of General Scientific Project of Zhejiang Education Department (Y201941020), Shaoxing Science and Technology Bureau (2017B70010), Research Funds of Shanghai Health and Family Planning Commission (20184Y0022), Cultivation fund of clinical research of Renji Hospital (PY2018-III-06),

Shen 2020 *(Continued)*

and Clinical Pharmacy Innovation Research Institute of Shanghai Jiao Tong University School of Medicine (CXYJY2019ZD001 and CXYJY2019QN004).

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "Studies were eligible for inclusion if they were RCTs or observational studies; included elderly patients (75 years) with atrial fibrillation; compared direct oral anticoagulants with vitamin-K antagonists (warfarin, phenprocoumon et al.); and reported benefits and harmful outcomes. For the highest quality observational studies, only nationwide or health insurance database studies that reported adjusted or matched data using an authorized method to minimize confounding [covariate adjustment, propensity score adjustment, propensity score matching, inverse probability of treatment weighting] were included." Comment: detailed eligibility criteria are reported.
Investigator agreement?	Yes	Quote: "Two authors (N-NS and YW) independently reviewed each title and abstract, and assessed full texts of retrieved studies, with any disagreements resolved via consultation with the corresponding authors." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Yes	Comment: eligibility criteria were broad, search seems systematic, hence a complete consecutive sample can be expected.
Bias assessed?	Unclear	Quote: "The methodological quality of each included RCT was assessed according to the Cochrane Collaboration Risk of Bias Tool. Considering an inherently higher bias risk of observational studies relative to RCTs, the methodological quality of each observational studies was evaluated using the following items: (1) using authorized adjustment method to deal with selection bias; (2) potential for residual confounding; (3) using methods to handle time-varying covariates and information censoring; and (4) reporting baseline characteristics and outcome measures in detail." Comment: risk of bias assessed; assessment tool for observational studies suboptimal.
Control for differences?	Yes	Comment: RCTs and observational studies were analysed separately. Sample of observational studies seems to include cohort studies only.
Heterogeneity addressed?	Yes	Quote: "In addition, subgroup analyses of OSs were performed based on individual agents (dabigatran, rivaroxaban, apixaban, and edoxaban), gender (men and women), age (>80, >85, and >90 years), and population (U.S.A, Canada, Italy, Germany, Sweden, Danish, France, Spain, Korea, etc.)"; "Meta-regression analysis was performed to determine the potential bias of effect factors on outcomes." Comment: the effect of different factors was assessed with subgroup analyses and meta-regression analyses.
Similar outcomes?	Yes	Comment: outcome measures were similar in both study types.
No selective reporting	Yes	Comment: no selective reporting suspected; all outcomes reported in methods section are also reported in the results section.

Shen 2020 *(Continued)*

Absence of evidence of bias from other sources?	Yes	Comment: no other forms of bias suspected.

Tan 2017

Study characteristics

Methods	Systematic review (conference abstract)
Data	8 RCTs and 19 observational studies (number of participants not reported) examining effects of concurrent chemotherapy (CCT) with or without adjuvant chemotherapy (AC) versus radiotherapy alone in stage II-IVB nasopharyngeal carcinoma (NPC) in studies. Authors searched MEDLINE (no search date reported). **Inclusion criteria:** "studies determining the effect of the addition of concurrent chemotherapy with or without adjuvant chemotherapy in stage II-IVB nasopharyngeal carcinoma. The outcome of interest was overall survival." **Exclusion criteria:** Not reported
Comparisons	Pooled hazard ratios were extracted for each outcome
Outcomes	1 outcome relevant for this review: survival in stage II-IVB nasopharyngeal carcinoma
Notes	Conference abstract. This review was not included in the meta-analysis for it reported to little data. Reported results: "The survival benefit associated with chemotherapy was similar in both RCTs and observational studies focusing on stage III-IVB nasopharyngeal carcinoma. Similarly, there was no significant difference in overall survival for RCT and observational studies focusing on stage II nasopharyngeal carcinoma." Funding: not reported

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Unclear	Quote: "We searched MEDLINE for eligible studies determining the effect of addition of CCT [concurrent chemotherapy] with or without AC [adjuvant chemotherapy] in stage II-IVB NPC [nasopharyngeal carcinoma]. Outcome of interest was overall survival (OS)." Comment: insufficient information reported.
Investigator agreement?	Unclear	Comment: no information reported.
Complete sample?	No	Comment: literature search was limited to one database only.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	Unclear	Comment: no information reported.
Heterogeneity addressed?	No	Comment: no information reported about the control for heterogeneity.

Tan 2017 *(Continued)*

Similar outcomes?	Yes	Comment: the main outcome of interest was survival, which might allow for sufficient similarity in definition.
No selective reporting	Unclear	Comment: too little information reported in abstract to assess reporting bias.
Absence of evidence of bias from other sources?	Unclear	Comment: too little information reported in abstract to assess other biases.

Tzoulaki 2011

Study characteristics	
Methods	Overview of reviews
Data	36 RCTs and unclear number of observational studies examining prognostic effect size of cardiovascular biomarkers. Authors searched up to 2011 in MEDLINE, and searched for meta-analyses of individual participant data published by major consortia operating in the speciality. **Inclusion criteria:** Meta-analyses that examined any emerging biomarker, defined as any biological parameter other than those included in the Framingham risk score, in relation to cardiovascular disease, coronary heart disease, or cardiovascular mortality, included at least one meta-analysis examining the association between an eligible biomarker with an eligible outcome, and containing data from at least one dataset from an observational study and one from a randomised controlled trial, which was subsequently analysed as an observational study. Reviews were included regardless of the baseline characteristics (clinical setting) of the examined populations. If an article presented separate meta-analyses on more than one eligible biomarker or outcome or on participants with different clinical settings, these meta-analyses were kept separate. Reviews were also included regardless of whether the included studies used adjustment for some covariates or score (such as the Framingham risk score) or tested for association in unadjusted analyses. When we identified more than one meta-analysis examining the same biomarker and same outcome on the same clinical setting, we kept only the most recent one with eligible data. Meta-analyses were included regardless of whether they were meta-analyses of the published literature or of individual participant data. **Exclusion criteria:** Not reported
Comparisons	Pooled risk ratios were extracted for each outcome, using the same indicator that had been used in the meta-analysis of interest
Outcomes	20 outcomes relevant to this review: cardiovascular risk with selenium, C-reactive protein and cardiovascular risk, triglycerides and cardiovascular risk, non-HDL cholesterol and cardiovascular risk, Lp(a) lipoprotein and cardiovascular risk, lipoprotein-associated phospholipase A2 activity and cardiovascular risk, lipoprotein-associated phospholipase A2 mass, leucocyte count and cardiovascular risk without a history of cardiovascular risk, leucocyte count and cardiovascular risk with pre-existing history of cardiovascular risk, Chlamydia pneumoniae immunoglobulin G titre and cardiovascular risk, homocysteine and cardiovascular risk, apolipoprotein B and cardiovascular risk, apolipoprotein A I top third, apolipoprotein B:A ratio and cardiovascular risk, fibrinogen and cardiovascular risk, nighttime ambulatory blood pressure and cardiovascular risk, daytime ambulatory blood pressure and cardiovascular risk, B-type natriuretic peptide and cardiovascular risk, coronary artery calcium and cardiovascular risk, troponin

Tzoulaki 2011 *(Continued)*

Notes	Reported results: "For seven major biomarkers, the prognostic effect was significantly stronger in datasets from observational studies than in datasets from randomised controlled trials." Funding: there was no funding for this study.

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "Articles were eligible if they included at least one meta-analysis examining the association between an eligible biomarker with an eligible outcome and containing data from at least one dataset from an observational study and one from a randomised controlled trial, which was subsequently analysed as an observational study." Comment: explicit inclusion criteria described in methods section.
Investigator agreement?	Unclear	Comment: no information reported.
Complete sample?	Yes	Comment: studies were searched systematically and explicit criteria for study selection were described.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	Yes	Comment: subgroup analyses by study design found no difference between observational study designs. Overall, results of observational studies were reported in aggregate.
Heterogeneity addressed?	No	Comment: no information reported regarding control for heterogeneity amongst participants and interventions.
Similar outcomes?	Unclear	Comment: cardiovascular risk as main outcome; no information reported on whether definitions differed amongst studies.
No selective reporting	Yes	Comment: no selective reporting suspected.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Van de Wall 2020

Study characteristics

Methods	Systematic review
Data	2 RCTs (n = 150) and 10 observational studies (n = 1262) comparing conservative with operative treatment for humeral shaft fractures. Authors searched up to March 2019 in MEDLINE, Embase, CENTRAL, and CINAHL. **Inclusion criteria:** Studies about humeral shaft fracture, conservative treatment (cast immobilisation and/or functional bracing), operative treatment (minimally invasive or open plating, nail fixation, and external fixator), age 16 years or older, and reporting of outcomes of interest (nonunion, reintervention, time to union, radial nerve palsy, and functional outcomes). **Exclusion criteria:**

Van de Wall 2020 *(Continued)*

	Studies about pathologic fractures; treatment for delayed union or nonunion; studies with an average follow-up period of less than 6 months; languages other than English, French, German, or Dutch; no availability of full text; and letters, meeting proceedings, and case reports.
Comparisons	Pooled odds ratios were extracted for each outcome, using the same indicator that had been used in the meta-analysis of interest
Outcomes	2 outcomes relevant for this review: nonunion rate and intervention or reintervention rate
Notes	Reported results: "There appeared to be no difference in mean time to union and mean Disabilities of the Arm, Shoulder." Funding: "The authors, their immediate families, and any research foundations with which they are affiliated have not received any financial payments or other benefits from any commercial entity related to the subject of this article."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "The inclusion criteria were humeral shaft fracture, conservative treatment (cast immobilization and/or functional bracing), operative treatment (minimally invasive or open plating, nail fixation, and external fixator), age 16 years or older, and reporting of outcomes of interest (nonunion, reintervention, time to union, radial nerve palsy, and functional outcomes). The exclusion criteria were pathologic fractures; treatment for delayed union or nonunion; studies with an average follow-up period of less than 6 months; languages other than English, French, German, or Dutch; no availability of full text; and letters, meeting proceedings, and case reports." Comment: the PubMed/MEDLINE, Embase, CENTRAL (Cochrane Central Register of Controlled Trials), and CINAHL (Cumulative Index to Nursing and Allied Health Literature) databases were searched for studies; eligibility criteria reported in detail.
Investigator agreement?	Yes	Quote: "Two reviewers (B.J.M.v.d.W. and Y.O.) independently screened titles and abstracts for eligibility"; "The same 2 reviewers independently performed the full-text screening." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Yes	Eligibility criteria were sufficiently detailed, search seems systematic, hence a complete consecutive sample can be expected.
Bias assessed?	Yes	Quote: "Two reviewers (B.J.M.v.d.W. and Y.O.) independently assessed the methodologic quality of included studies using the Methodological Index for Non-Randomized Studies (MINORS). The MINORS is a validated instrument for assessing the methodologic quality of cohort studies, resulting in a score between 0 and 24. Randomized studies were appraised using the same tool to measure quality on the same scale as observational studies." Comment: risk of bias was assessed with the MINORS tool. The same tool was also used for RCTs.
Control for differences?	Yes	Comment: RCTs and cohort studies were analysed separately.
Heterogeneity addressed?	Unclear	Quote: "Only 3 studies - all observational studies - had a study population with a mean age older than 50 years. The pooled analysis did not demonstrate a dif-

Van de Wall 2020 *(Continued)*

		ference in nonunion rates between conservative and operative treatment (OR, 4.7; 95% CI, 0.8-26.1; I2 = 0%; Supplementary Fig. S7)." Comment: heterogeneity partially addressed.
Similar outcomes?	No	Quote: "This meta-analysis investigated the difference between conservative and operative treatment, irrespective of type of operative management (nail, plate, minimally invasive techniques). Finally, to increase the power of the pooled analysis, we used a compound endpoint for reintervention. In other words, we did not take the severity of the indication or reintervention itself into account." Comment: outcome measures might have been heterogeneous amongst studies in the review.
No selective reporting	Yes	Comment: no selective reporting suspected; all outcomes reported in methods section are also reported in the results section.
Absence of evidence of bias from other sources?	Yes	Comment: no other forms of bias suspected.

Van Heesewijk 2018

Study characteristics	
Methods	Systematic review
Data	2 RCTs (n = 185) and 25 observational studies (n = 16,103) examining surgical treatments for necrotising enterocolitis (NEC) in preterm infants. Authors searched up to 1990 and 2017 in PubMed and the CDSR. **Inclusion criteria:** Papers which: (1) included low birth-weight (< 2500 g) or preterm (< 37 weeks' gestation) infants, (2) compared surgical treatment of NEC with laparotomy versus peritoneal drainage, and (3) reported the primary outcome of mortality. All randomised trials, quasi-randomised trials, cohort studies, and case-control studies were included. No language restrictions were applied, and non-English articles were translated. **Exclusion criteria:** Not reported
Comparisons	Pooled odds ratios were extracted for each outcome, using the same indicator that had been used in the meta-analysis of interest.
Outcomes	1 outcome of relevance for this review: mortality
Notes	Reported results: "Meta-analysis of observational studies demonstrated significantly lower mortality after laparotomy, as compared to peritoneal drainage. In contrast, RCTs demonstrated no difference in mortality." Funding: not reported

Risk of bias

Item	Authors' judgement	Support for judgement

Van Heesewijk 2018 *(Continued)*

Explicit criteria?	Yes	Quote: "The PubMed search was performed using the Mesh terms "Necrotizing Enterocolitis" AND "surgery", and subsequently using Mesh terms "Necrotizing Enterocolitis" AND "low birth weight infant". Lastly, the reference lists of included articles were reviewed for additional eligible publications. No language restrictions were applied and non-English articles were translated." Comment: inclusion criteria were reported in detail.
Investigator agreement?	Yes	Quote: "All searches, assessments of methodology and bias, and extraction of data were performed by two authors (AVH, MR), with resolution of queries by a third party (SBD)." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Unclear	Quote: "Articles published between 1 January 1990 and 1 May 2017 were retrieved for further screening." Comment: limited search for studies might have yielded incomplete sample.
Bias assessed?	No	Quote: "Standardized checklists were used to assess methodological quality and risk of bias in both the observational studies and the RCTS. The methodological quality of reporting of observational studies assessed with the STROBE checklist. The CONSORT guidelines were used to assess the reporting quality of RCTs." Comment: reporting guidelines were used as a proxy to assess study quality.
Control for differences?	No	Comment: observational studies were not analysed separately by study design (i.e. cohort studies and case-control studies).
Heterogeneity addressed?	No	Comment: no information reported. There is not a consensus about the optimal definition of necrotising enterocolitis, either clinically or in the literature about its treatment.
Similar outcomes?	Yes	Comment: the main outcome was mortality, which might allow for sufficient similarity in definition.
No selective reporting	Yes	Comment: all results for the mentioned analyses were reported.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Virk 2019

Study characteristics	
Methods	Systematic review
Data	9 RCTs (n = 903) and 26 observational studies (n = 8919) comparing outcomes of atrial fibrillation ablation performed with versus without contact force guidance. Authors searched up to May 2018 in MEDLINE and Embase. **Inclusion criteria:** "RCTs or controlled observational studies were assessed according to the following eligibility criteria: (i) patients with atrial fibrillation (paroxysmal or persistent) undergoing pulmonary vein isolation with or without additional substrate ablation; (ii) patients in intervention arm undergoing radiofrequency

Virk 2019 *(Continued)*

	ablation guided by direct assessment of CF; (iii) patients in control arm undergoing radiofrequency ablation without contact force guidance (either with use of conventional non-contact force catheters or blinding operators to contact force information); (iv) reporting on at least one of: incidence of peri-procedural complications, freedom from atrial fibrillation at follow-up or procedural parameters (total procedure duration, fluoroscopy duration/exposure, radiofrequency ablation duration); (v) sample size of at least 10 patients in each arm." **Exclusion criteria:** "Studies comparing contact force-guided ablation with cryoballoon ablation or radiofrequency ablation with robotic or remote magnetic navigation systems were excluded. Studies reporting ablation of different arrhythmias were excluded if data was not separately reported for the atrial fibrillation subgroup. All publications were limited to those involving adult subjects and written in English. Conference abstracts were excluded if data was not subsequently published as a peer-reviewed journal article. When institutions published duplicate reports with accumulating numbers of patients or increased lengths of follow-up, only the most complete reports were included."
Comparisons	Pooled relative risks or (standardised) mean differences were extracted for each outcome, using the same indicator that had been used in the meta-analysis of interest
Outcomes	7 outcomes of relevance for this review: procedure duration, fluoroscopy duration, fluoroscopy exposure, radiofrequency ablation duration, major peri-procedural complications, freedom from atrial fibrillation, cardiac tamponade
Notes	Reported results: "Meta-analysis of randomized data demonstrated that CF guidance does not improve the safety or efficacy of AF ablation, despite initial observational data showing dramatic improvement." Funding: not reported

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "Randomized controlled trials (RCTs) or controlled OS were assessed according to the following eligibility criteria: (i) patients with AF (paroxysmal or persistent) undergoing pulmonary vein isolation (PVI) with or without additional substrate ablation; (ii) patients in intervention arm undergoing RFA guided by direct assessment of CF; (iii) patients in control arm undergoing RFA without CF guidance (either with use of conventional non-CF catheters or blinding operators to CF information); (iv) reporting on at least one of: incidence of peri-procedural complications, freedom from AF at follow-up or procedural parameters (total procedure duration, fluoroscopy duration/exposure, RFA duration); (v) sample size of at least 10 patients in each arm." Comment: detailed eligibility criteria are reported; few details reported with regard to eligibility of study designs
Investigator agreement?	Yes	Quote: "Two reviewers (S.A.V. and J.A.) independently screened the title and abstract of records identified in the search." Comment: study selection was conducted independently by two or more reviewers.
Complete sample?	Yes	Eligibility criteria were sufficiently detailed, search seems systematic, hence a complete consecutive sample can be expected.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	Yes	Comment: RCTs and cohort studies were analysed separately.

Virk 2019 *(Continued)*

Heterogeneity addressed?	Yes	Quote: "Meta-regression analyses were performed to assess the impact of the following covariates: study sample size, study publication date, mean age of participants, proportion of male participants, baseline left ventricular ejection fraction, baseline left atrial diameter, ablation strategy (PVI only vs. additional ablation), and mean CF in the CF guidance cohort." Comment: heterogeneity addressed in meta-regression.
Similar outcomes?	Yes	Comment: the studies measured similar outcomes following the interventions, and were grouped as such for analysis. Outcome measures were similar in both study types.
No selective reporting	Yes	Comment: no selective reporting suspected; all outcomes reported in methods section are also reported in the results section.
Absence of evidence of bias from other sources?	Yes	Comment: no other forms of bias suspected.

Yanik 2013

Study characteristics	
Methods	Overview of reviews (reported as conference abstract)
Data	12 systematic reviews included, including an unclear number of RCTs and observational studies (open topic). Authors searched between 2000 and 2011 in PubMed (database unclear). **Inclusion criteria:** Meta-analyses and systematic reviews that compared RCTs and nRCTs for at least 3 different exposure-outcome associations in a peer-reviewed journal. **Exclusion criteria:** Not reported
Comparisons	No numerical data were reported or extracted
Outcomes	1 outcome of relevance for this review: degree of concordance between results from RCTs and non-randomised studies
Notes	Reported results: "The non-RCT estimate and RCT estimate were on opposite sides of the null for 25% of associations. Among the associations in which the RCT and non-RCT estimates were on the same side of the null, 46% had a ratio of odds ratios greater than 1, indicating an nRCT estimate further away from the null than the RCT estimate. For 16% of associations RORs indicated the RCT estimate was statistically significantly different from the nRCT estimate with an alpha of 0.05. In 9% of the comparisons the ROR estimate was statistically significant and the RCT and nRCT estimates were on opposite sides of the null." Funding: not reported

Risk of bias		
Item	**Authors' judgement**	**Support for judgement**
Explicit criteria?	Unclear	Quote: "We systematically identified meta-analyses and systematic reviews that compared RCTs and nRCTs for at least 3 different exposure-outcome associations in a peer-reviewed journal through a search of the PubMed database."

Yanik 2013 *(Continued)*

		Comment: inclusion criteria mentioned in the abstract: meta-analyses and systematic reviews that compared RCTs and nRCTs for at least 3 different exposure-outcome associations, published in a peer-reviewed journal.
Investigator agreement?	Unclear	Comment: no information reported.
Complete sample?	No	Comment: literature search was restricted to one database only.
Bias assessed?	Unclear	Comment: no risk of bias assessment reported.
Control for differences?	Unclear	Comment: not reported.
Heterogeneity addressed?	No	Comment: not reported.
Similar outcomes?	Unclear	Comment: three main outcomes were selected, but apparently analysed in aggregate.
No selective reporting	Unclear	Comment: too little information reported in abstract to assess reporting bias.
Absence of evidence of bias from other sources?	Unclear	Comment: too little information reported in abstract to assess other biases.

Youn 2021

Study characteristics	
Methods	Systematic review
Data	6 RCTs (n = 5352) and 87 (n = 239,433) observational studies that examined transcatheter versus surgical aortic valve replacement. Authors searched up to June 2017 in MEDLINE, MEDLINE In-Process / ePubs, Embase, CENTRAL, CDSR, Scopus, and Web of Science. **Inclusion criteria:** All RCTs that "randomly assigned patients to transcatheter or surgical aortic valve replacement and followed patients over time". All comparative cohort studies that "reported primary data on outcomes of interest after transcatheter or surgical aortic valve replacement." **Exclusion criteria:** "Non-randomized studies that were not comparative cohort studies, defined the population by excluding the outcome of interest, combined patients from RCTs and non-randomized studies, conference abstracts, poster presentations, non-peer reviewed publications, unpublished literature, systematic reviews that lacked primary data, and studies that used other surgical aortic valve replacement methods (e.g., minimally invasive, sutureless)."
Comparisons	Pooled odds ratios, adjusted odds ratios, ratios of odds ratios, mean differences and differences in mean differences were extracted, using the same indicator that had been used in the meta-analysis of interest
Outcomes	2 outcomes of relevance for this review: postoperative mortality and length of hospital stay
Notes	Reported results: "Nonrandomized studies underestimated the benefit of transcatheter aortic valve implantation compared with RCTs. Nonrandomized studies using propensity score matching and regression modelling to adjust results estimated treatment effects closer to high quality RCTs."

Youn 2021 *(Continued)*

Funding: "The study was funded by the Canadian Institutes of Health Research (CIHR) operating Grant MOP-136787."

Risk of bias

Item	Authors' judgement	Support for judgement
Explicit criteria?	Yes	Quote: "We included all RCTs that randomly assigned patients to transcatheter or surgical aortic valve replacement and followed patients over time. We also included all comparative cohort studies that reported primary data on outcomes of interest after transcatheter or surgical aortic valve replacement." Comment: eligibility criteria reported.
Investigator agreement?	Unclear	Quote: "We used DistillerSR (Evidence Partners, Ottawa, Canada) to check for duplicate citations, and to screen titles, abstracts, and full text." Comment: insufficient information reported.
Complete sample?	Yes	Quote: "We searched Medline, Medline In-Process/ePubs, Embase, Cochrane Central Register of Controlled Trials, Cochrane Database of Systematic Reviews, Scopus, and Web of Science from inception to June 2017." Comment: systematic, broad search and eligibility criteria reported.
Bias assessed?	Yes	Quote: "RCTs were divided into high or low quality RCTs based on the Cochrane Risk Of Bias (ROB) tool [24] based on the content of the published articles; authors were not contacted for additional information. No RCT blinded study participants; hence RCTs that satisfied all other criteria were categorized as high quality. Non-randomized studies reported unadjusted estimates, adjusted estimates, or both. Non-randomized studies estimates were pooled into 3 groups: without adjustment, adjusted using PSM, and adjusted using regression. Finally, we previously developed a set of 41 non-randomized studies attributes that could bias studies (Additional file 1: Table S2). These attributes were based on existing frameworks of bias and quality assessment tools for non-randomized studies, and were extensively pilot tested and iteratively developed for clarity and reliability." Comment: risk of bias was assessed with Cochrane RoB tool.
Control for differences?	Yes	Quote: "We compared the pooled estimates of the effect measures between study categories, and also between nonrandomized studies with attributes hypothesized to be associated with bias. In all comparisons, ROR<1 and DMD<0 indicated that studies favored transcatheter aortic valve implantation." Comment: differences between studies accounted for by separating them by study type.
Heterogeneity addressed?	No	Comment: heterogeneity not controlled for.
Similar outcomes?	Yes	Quote: "We defined postoperative mortality as death due to any cause within 1-month or in hospital after the procedure regardless of location. We defined length of stay as the number of days the patient stayed in the hospital after the procedure. We extracted the necessary components of each outcome to calculate the pooled estimates of treatment effects. We calculated missing data points using given information where possible." Comment: outcomes were defined and similar between studies.

Youn 2021 *(Continued)*

No selective reporting	No	Comment: details of subgroup characteristics, such as total participants and heterogeneity of the meta-analysed data, are missing in some instances.
Absence of evidence of bias from other sources?	Yes	Comment: no other sources of bias suspected.

Ziff 2015

Study characteristics	
Methods	Systematic review
Data	8 RCTs and 27 observational studies (number of participants not reported) that examined treatment with digoxin compared with control (placebo or no treatment). Authors searched up to July 2014 in MEDLINE, Embase, and the Cochrane Library (not further specified). **Inclusion criteria:** All studies that examined comparative outcomes with digoxin and control (placebo or no treatment), regardless of study design. All cardiovascular outcomes and all populations were included. **Exclusion criteria:** Studies that did not provide comparative outcomes or were not published as full-text articles in English.
Comparisons	Pooled relative risks for RCTs, observational studies, and observational studies that employed propensity scores
Outcomes	2 outcomes of relevance for this review: all-cause mortality, all-cause hospitalisation
Notes	Reported results: "Digoxin is associated with a neutral effect on mortality in randomised trials and a lower rate of admissions to hospital across all study types. More pronounced effect was seen in (propensity score) adjusted cohort studies in 1 of three outcomes, no difference in the effect estimates was seen between RCTs and observational studies in 2 outcomes." Funding: "The study was funded by a grant from the Arthur Thompson Trust, University of Birmingham."

Risk of bias		
Item	**Authors' judgement**	**Support for judgement**
Explicit criteria?	Yes	Quote: "We evaluated all studies that examined comparative outcomes with digoxin and control (placebo or no treatment), regardless of study design. All cardiovascular outcomes and all populations were included. We excluded studies that did not provide comparative outcomes or were not published as full text articles in English." Comment: eligibility criteria are reported briefly, all study designs were included.
Investigator agreement?	Unclear	Comment: no information reported.
Complete sample?	Yes	Comment: eligibility criteria were sufficiently detailed, search seems systematic, hence a complete consecutive sample can be expected.

Ziff 2015 *(Continued)*

Bias assessed?	Yes	Quote: "We assessed the risk of bias with the Cochrane Collaboration's risk of bias tool for randomised controlled trials and the risk of bias assessment tool for non-randomised studies (RoBANS), both of which address key criteria such as selection bias, exposure measurement, blinding, completeness of outcome data, and selectivity of reporting. We assessed of risk of bias using these standardised tools independently from data extraction, with each study assessed by two authors and adjudication by a third when required." Comment: risk of bias assessed with Cochrane RoB tool.
Control for differences?	Yes	Comment: studies were analysed separately according to study type.
Heterogeneity addressed?	Yes	Quote: "Meta-regression was used to explore the impact of differences in key baseline characteristics between digoxin and control patients on all cause mortality in observational data. Studies with smaller differences in the percentage of patients with diabetes, as well as those receiving diuretics and anti-arrhythmic drugs, reported less difference in mortality between digoxin and control. At study level, baseline age and year of publication also significantly affected the comparative risk of death between patients treated with digoxin and control (table 3 and fig B in appendix 3)." Comment: heterogeneity in population assessed in meta-regression.
Similar outcomes?	Unclear	Quote: "Definitions of heart failure and atrial fibrillation in different studies varied, and we cannot exclude misclassification. Although some studies reported the stage of heart failure, left ventricular ejection fraction, and the type of atrial fibrillation, many studies did not." Comment: outcome measures were similar in both study types; definitions of outcome measures were adapted from the individual included studies.
No selective reporting	Yes	Comment: no selective reporting suspected; all outcomes reported in methods section are also reported in the results section or the appendix.
Absence of evidence of bias from other sources?	Yes	Comment: no other forms of bias suspected.

AMED: Allied and Complementary Medicine Database;**CABG:** coronary artery bypass graft; **CDSR:** Cochrane Database of Systematic Reviews; **CENTRAL:** Cochrane Central Register of Controlled Trials; **CI:** confidence interval; **CINAHL:** Cumulative Index to Nursing and Allied Health Literature; **HbA1c**: haemoglobin A1c; **MD:** mean difference; **nRCT:** non-randomised controlled study; **NRS:** non-randomised study; **PI/ECO:** population, intervention/exposure, comparison, and outcome; **PS:** propensity score; **RCT:** randomised controlled trial; **RoB:** risk of bias; **ROBINS-I:** risk of bias in non-randomised studies of interventions; **RR:** risk ratio; **RRR:** ratio of risk ratios; **SR:** systematic reviews;

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

Study	Reason for exclusion
Abraham 2010	No own MA performed
Ahren 2014	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Algra 2012	Incidental comparison of study types
Anderson 2019	No sample to compare observational studies and RCT defined a priori
Arditi 2016	Incidental comparison of study types

Study	Reason for exclusion
Atar 2015	No sample to compare observational studies and RCT defined a priori
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.
Beyerbach 2022	For one study type only existing SRs/MAs are included
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.
Collins 2012	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Concato 2000a	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Concato 2000b	No own MA performed
Coscia 2019	No own MA performed
Coulam 1994	From this study it was not possible to separate out uncontrolled, quasi-, or pseudo-randomized studies from other studies.
CRD42017058116	Incidental comparison of study types
CRD42018104452	Incidental comparison of study types
CRD42019130585	Incidental comparison of study types
da Silva 2017	Incidental comparison of study types
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
El-Hayek 2014	Incidental comparison of study types
Ewald 2020	Comparison with a single study

Study	Reason for exclusion
Flossmann 2007	An original meta-analysis with an incidental comparison of RCTs and observational studies.
Fukuta 2017	Incidental comparison of study types
Furlan 2008a	No own MA performed
Gray 2017	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Gyawali 2020	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
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.
Hundscheid 2021	Incidental comparison of study types
Ioannidis 2005	This is a qualitative comparison of high cited RCTs and observational studies and their initially stronger effects that are often later contradicted.
Jee 2016	Incidental comparison of study types
Khan 2019	Incidental comparison of study types
Kilcher 2018	No own MA performed
Kim 2020	Incidental comparison of study types
Kirk 2004	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Kishimoto 2021	Incidental comparison of study types
Kitsios 2015a	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Krogh 2021	No systematically compiled sample
Kunz 1998	No own MA performed
Kuss 2020	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Labrarere 2006	This is a methods paper that did not have a systematic selection of studies for identified outcomes or interventions.
Lai 2010	Incidental comparison of study types
LaTorre 2009	An original meta-analysis of harms outcomes among only observational studies.
Leichsenring 2008	Incidental comparison of study types

Study	Reason for exclusion
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.
Liu 2017	Incidental comparison of study types
Loke 2011	An original meta-analysis with an incidental comparison of RCTs and observational studies.
Mak 2009	An original meta-analysis with an incidental comparison of RCTs and observational studies.
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.
Mehyar 2021	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Melloni 2015	Incidental comparison of study types
Moreira 2012	No meta-analysis; RCT data included quasi-experimental.
Morgan 2014	Incidental comparison of study types
Moyer 2002	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Mugavero 2011	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Ni Chroinin 2013	An original meta-analysis with an incidental comparison of RCTs and observational studies.
Nigwekar 2009	Incidental comparison of study types
Nixdorf 2010	An original meta-analysis with an incidental comparison of RCTs and observational studies.
NN 2014	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Oliver 2010	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Ottenbacker 1992	A commentary and/or descriptive analysis.
Papageorgiou 2015	No own MA performed
Papanastassiou 2012	An original meta-analysis with an incidental comparison of RCTs and observational studies.
Pasala 2016	Incidental comparison of study types
Peinemann 2013	No own MA performed

Study	Reason for exclusion
Phillips 1999	This study had no systematic selection of meta-analyses; only included three large prospective studies that were the focus of the analysis.
Podmore 2021	Incidental comparison of study types.
Pratt 2012	No meta-analysis performed.
Pyorala 1995	An original meta-analysis with an incidental comparison of RCTs and observational studies.
Rivera-Caravaca 2018	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Rompen 2021	Incidental comparison of study types
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.
Shikata 2006	No own meta-analysis performed for data from RCTs.
Sison 2018	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Smeeing 2017	Incidental comparison of study types
Soni 2019	Comparison with a single study
Steinberg 1994	An analysis of previously published meta-analyses that aimed to compare effects between sources of controls within observational study designs.
Stuart 2017	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
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.
Sun 2007	This is not a meta-analysis or review of meta-analyses. There is no comparison of RCTs and observational data.
Syn 2020	No own MA performed
Theodoratou 2014	No own MA performed
Trikalinos 2012	No own MA performed
Vigil-De Gracia 2020	Incidental comparison of study types
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.

Study	Reason for exclusion
Watson 1994	An original meta-analysis with an incidental comparison of RCTs and observational studies; the authors include non-randomized as observational studies.
Wilkes 2010	Study designs not limited to RCTs and 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.
Yank 2009	Incidental comparison of study types
Zhang 2014a	For one study type only, existing SRs/MAs are included
Zhang 2014b	For one study type, only existing SRs/MAs are included

MA: meta-analysis; **RCT:** randomized controlled trial; **SR:** systematic review

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

CRD42014013478

Study name	Comparison of the associations of specific foods on body weight between RCTs and observational epidemiologic studies
Methods	Systematic review. Searches in MEDLINE, Scopus, Cochrane, and Pro Quest for studies published in English. Title and abstract screening by one reviewer. No information about staff for data extraction. For binary outcomes, the summary measures include RR, OR, or risk difference. For continuous outcomes, the summary measure is the difference in means, using the same scale. No subgroup analyses planned
Data	RCTs and cohort studies
Comparisons	Associations of specific foods on body weight
Outcomes	Weight change
Starting date	07 April 2014
Contact information	Ms Pufal, milene.pufal@gmail.com
Notes	

CRD42017059665

Study name	Comparison of cohort and controlled studies of suicide risk categorisation: a meta-analysis
Methods	Searches in PubMed for studies in English and papers with an abstract, complemented by hand-searching relevant review article citations. Independent data extraction by two researchers. Analysis by random-effects meta-analysis with mixed-effects meta regression, quantitative tests of publication bias and sensitivity analysis.

CRD42017059665 *(Continued)*

Data	Controlled, matched case-control, cohort, nested case-control
Comparisons	Suicide by psychiatric patients
Outcomes	Odds of suicide and area under the curve (psychometric properties) in the high risk groups according to cohort and control design. Additional outcome: the extent to which the number of initially examined variables and the number of variables in the high risk models explains heterogeneity in diagnostic odds
Starting date	19 March 2017
Contact information	Matthew Large, mmbl@bigpond.com
Notes	

CRD42017079569

Study name	Examining the treatment effect of antibiotic use to treat acute respiratory tract infections in primary care: a systematic review comparing the outcomes in randomised control trials (RCTs) and observational studies
Methods	Systematic search in electronic databases for RCTs and observational studies on effectiveness of antibiotics on acute respiratory tract infections. Within RCTs and observational studies, the results will be pooled in an aggregate data meta-analysis. A generic inverse-variance random-effects model will be used to pool the mean difference (MD) with 95% confidence interval (CI) for continuous outcomes to incorporate heterogeneity. When the units of the outcome measures used across studies are not consistent, the effects will be reported as standardised mean differences. For dichotomous data, a random-effects model will be used to pool the summary risk ratio with 95% CI. The estimates produced by RCTs and by observational studies will then be compared narratively. Subgroups: traditional methods to control for confounding versus studies that used enhanced methods (propensity scores/instrumental variable methods) to control for confounding
Data	All observational cohort studies and RCTs exploring the effectiveness of antibiotics when prescribed for acute respiratory tract infections will be included in the review.
Comparisons	Acute respiratory tract infections
Outcomes	Patient-reported severity of illness in the days following consultation; duration of illness following consultation, measured either directly by patient/clinician report or indirectly as the time to return to normal activities; and complications or worsening of illness requiring re-consultation and/or hospitalisation
Starting date	14 August 2017
Contact information	Beth Stuart, bls1@soton.ac.uk
Notes	

CRD42018062204

Study name	Comparison of treatment effects in randomised vs. non-randomised studies and the role of analytical methods to control for confounding: a meta-epidemiological study

CRD42018062204 *(Continued)*

Methods	Systematic review. Searches for meta-epidemiologic reviews in previously published systematic reviews and electronic databases. Search for primary studies including RCTs and non-RCTs and match all retrieved studies by medical topic. Study selection according to prespecified inclusion and exclusion criteria. Analysis: use meta-epidemiological methods to obtain pooled estimates of systematic discrepancies in treatment effects between randomised and non-randomised studies, and between different types of analytical methods used in non-randomised studies. Subgroup analysis: subgroup studies according to analytical method used in non-randomised study and conduct meta-analyses, compare results with those of RCTs.
Data	Randomised and non-randomised studies (the latter including non-randomised study with concurrent controls; non-randomised study with external controls; non-randomised study without controls)
Comparisons	Covering a broad range of pharmaceutical interventions and therapeutic areas
Outcomes	Main outcome: the discrepancy in treatment effects between randomised and non-randomised studies. Additional outcome: discrepancy in treatment effects between randomised studies and various types of analytical methods applied in non-randomised studies (types of analytical methods will be determined based on the methods used in included studies, but examples include propensity score matching, instrumental variable analysis, multivariate regression analysis)
Starting date	1 May 2018 (anticipated)
Contact information	Maximilian Salcher-Konrad, m.salcher@lse.ac.uk
Notes	

OR: odds ratio; **RCT:** randomised controlled trial; **RD:** risk difference; **RR:** risk ratio

APPENDICES

Appendix 1. Search strategy for electronic databases

MEDLINE (OVID) (R) ALL

#	Search
1	(systematic review or meta-analysis).pt. or exp *Meta-Analysis as Topic/ or Randomized Controlled Trials as Topic/ or *Propensity Score/ or exp *observational trials as topic/ or *Outcome Assessment, Health Care/mt
2	(methodological stud* or methodological review or research on research or meta-epidemiological or meta-stud* or meta-research or meta-analys* or metaanalys* or metasynthe* or meta-synthe* or systematic literature review or systematic review or meta-review).ti,ab. or (review* or overview*).ti.
3	1 or 2
4	((epidemiologic* adj2 stud*) or observational or cohort or cross-sectional or longitudinal or serial or nonexperimental or non-experimental or nonrandomi*ed or non-randomi*ed or NRSs or case-control or before-after or pre-post or case-cohort or case-crossover or natural experiment* or propensity score* or research design*).ti. and randomi*ed.ti,ab,kf.
5	((epidemiologic* adj2 stud*) or observational or cohort or cross-sectional or longitudinal or serial or nonexperimental or non-experimental or nonrandomi*ed or non-randomi*ed or NRSs or

(Continued)

	case-control or before-after or pre-post or case-cohort or case-crossover or natural experiment* or propensity score* or research design*).ab. /freq=4 and randomi*ed.ab. /freq=4
6	4 or 5
7	(similar* or dissimilar* or consisten* or inconsisten* or differen* or concordan* or discordan* or agree* or disagree* or heterogene*).ti,ab.
8	((compar* or assess* or estimat* or overestimat* or over-estimat* or examin* or meta* or match* or pooled or pooling) adj9 (treatment* or effect* or size* or magnitude* or harm* or risk*)).ti,ab.
9	7 and 8
10	3 and 6 and 9
11	limit 10 to yr="1990 -Current"

Search strategy for the Cochrane Database of Systematic Reviews, The Cochrane Library

#	Search
#1	((epidemiologic* near/2 stud*) or observational or cohort or cross-sectional or longitudinal or serial or nonexperimental or non-experimental or nonrandomi*ed or non-randomi*ed or NRSs or case-control or before-after or pre-post or case-cohort or case-crossover or natural experiment* or propensity score* or research design*) AND randomi*ed:ti,ab,kw
#2	(similar* or dissimilar* or consisten* or inconsisten* or differen* or concordan* or discordan* or agree* or disagree* or heterogene*):ti,ab,kw
#3	((compar* or assess* or estimat* or overestimat* or over-estimat* or examin* or meta* or match* or pooled or pooling) near/9 (treatment* or effect* or size* or magnitude* or harm* or risk*)):ti,ab,kw
#4	#1 AND #2 AND #3
	Filter: Methodology

Search strategy for EPISTEMONIKOS (www.epistemonikos.org)

title:((epidemiological OR observational OR cohort OR cross-sectional OR longitudinal OR serial OR nonexperimental OR non-experimental OR nonrandomi*ed OR non-randomi*ed OR NRSs OR case-control OR before-after OR pre-post OR case-cohort OR case-crossover OR natural experiment* OR propensity OR "research design*" OR "study design*" OR "effect estimate*" OR "effect size*" OR "treatment effect*") AND random*) AND (title:(Compar* OR "methodological study" OR "methodological studies" OR "methodological review" OR "methodological reviews" OR "research on research" OR "meta-epidemiological study" "meta-epidemiological studies" OR "meta-study*" OR "meta-studies" OR "meta-research" OR overview*) OR abstract:(Compar* OR "methodological study" OR "methodological studies" OR "methodological review" OR "methodological reviews" OR "research on research" OR "meta-epidemiological study" "meta-epidemiological studies" OR "meta-study*" OR "meta-studies" OR "meta-research" OR overview*))

Search strategy for Science Citation Index and Emerging Sources Citation Index, Web of Science Core Collection, Clarivate (*Timespan = 1990 to 2021)*

#5 #1 AND #4

#4 #2 AND #3

#3 ts= ((compar* or assess* or estimat* or overestimat* or over-estimat* or examin* or meta* or match* or pooled or pooling) near/7 (treatment* or effect* or size* or magnitude* or harm* or risk*))

#2 ts=(similar* or dissimilar* or consisten* or inconsisten* or differen* or concordan* or discordan* or agree* or disagree*

or heterogene*)

#1 ti=((epidemiologic* near/2 stud*) or observational or cohort or cross sectional or longitudinal or serial or nonexperimental or non-experimental or nonrandomi*ed or non-randomi*ed or NRSs or case-control or before-after or pre-post or case-cohort or case crossover or natural experiment* or propensity score* or research design*) and ti=randomi*ed

Appendix 2. Searching other resources

Google Scholar Search

Google Scholar was searched on 5 May 2021 using the Original Review as the source reference and analysing citing references.

411 citing references were found, these were further filtered by using the "search in articles with citations" function with the term "non randomised", returning 209 results. The first 30 were screened for relevancy, of which 4 appeared relevant to the topic, 3 of these were found by the bibliogrpahic database searches also (Medline, WoS), 1 was a unique hit, a relevant thesis on the review topic.

References to reviews that were included in forward and backward citation tracking

Allain 2017; Ankarfeldt 2017; Artus 2014; Beks 2018; Benson 2000; Beynon 2008; Bhandari 2004; Borkowska 2018; Bröckelmann 2022; Dahabre 2012; Demissie 1998; Edwards 2012; Furlan 2008; Golder 2011; Gu 2020; Guyatt 2000; Hong 2021; Hoshino 2021a; Hoshino 2021b; Ioannidis 2001; Jainaud 2021; Kimachi 2021; Kirson 2013; Kitsios 2015; Kuss 2011; Li 2016; Lonjon 2013; MacLehose 2000; Mathes 2021; Moneer 2022; Morfaw 2021; Müeller 2010; Naudet 2011; Otsuka 2022; Papanikolaou 2006; Safieddine 2021; Schmidt 2013; Schwingshackl 2021; Shen 2020; Tan 2017; Tzoulaki 2011; Van de Wall 2020; Van Heesewijk 2018; Virk 2019; Yanik 2013; Youn 2021; Ziff 2015

Appendix 3. Underlying data for the comparison of continuous effect estimates

Review ID	Outcome	Difference between meta-analyses of observational studies and RCTs
Ankarfeldt 2017	Change in HbA1c (intervention: glucagon-like peptide-1 analogs)	Observational studies and RCTs lead to similar results. Chi^2 not reported. MD_{RCTs}: -0.06 (95% CI -0.22 to 0.09, I^2 = 87.0%, 11 studies); MD_{OBS}: -0.02 (95% CI -0.57 to 0.52, I^2 not reported, 4 studies)
Ankarfeldt 2017	Change in HbA1c (intervention: dipeptidyl peptidase-4 inhibitors)	A higher change in HbA1c was observed in observational studies compared to RCTs. Chi^2 not reported. MD_{RCTs}: -0.01 (95% CI -0.08 to 0.05, I^2 = 38%, 12 studies); MD_{OBS}: -0.14 (95% CI -0.03 to -0.02, I^2 = 51%, 3 studies)

(Continued)

Artus 2014	Standard mean change in pain intensity at 52 weeks	Observational studies and RCTs lead to similar results. Chi^2 not reported. SMD_{RCTs}: 0.9 (95% CI 0.8 to 1.0, I^2 = 99%, unclear number of studies); SMD_{OBS}: 1.1 (95% CI 0.8 to 1.6, I^2 = 99%, 11 studies)
Beks 2018	Functional outcome as measured with Constant-Murley score	Observational studies and RCTs lead to similar results. Chi^2 not reported. MD_{RCTs}: 0.4 (95% CI -4.76 to 5.56, I^2 = 0%, 5 studies); MD_{OBS}: -1.50 (95% CI -7.33 to 4.33, I^2 = 80%, 9 studies)
Moneer 2022	Duration of hospital stay (hydroxychloroquine versus standard of care or placebo)	Observational studies and RCTs lead to similar results, a stronger effect was seen on evidence from observational studies. Chi^2 not reported. SMD_{RCTs}: 0.12 (95% CI -0.08 to 0.32, I^2 not reported, 16 studies); SMD_{OBS}: 0.35 (95% CI 0.16 to 0.54, I^2 not reported, 11 studies) difference in SMDs: 0.23 (95% CI -0.05 to 0.51, I^2 not reported)
Moneer 2022	Duration of hospital stay (lopinavir-ritonavir versus standard of care or placebo)	Observational studies and RCTs lead to similar results. Chi^2 not reported. SMD_{RCTs}: 0.11 (-0.10 to 0.32, I^2 not reported, 3 studies); SMD_{OBS}: 0.67 (-0.28 to 1.62, I^2 not reported, 4 studies); difference in SMDs: 0.56 (95% CI -0.42 to 1.54, I^2 not reported)
Moneer 2022	Duration of hospital stay (hydroxychloroquine and azithromycin versus standard of care or placebo)	Observational studies and RCTs lead to similar results, a stronger effect was seen on evidence from observational studies. Chi^2 not reported. SMD_{RCTs}: 0.07 (95% CI -0.11 to 0.25, I^2 not reported, 8 studies); SMD_{OBS}: 0.53 (95% CI 0.14 to 0.92, I^2 not reported, 2 studies); difference in SMDs: 0.46 (95% CI 0.02 to 0.90, I^2 not reported)
Moneer 2022	Time to viral clearance	Effect estimates from RCTs and OBS lie on opposite sides of the 0. Effect estimates are not statistically significant. Chi^2 not reported. ROM_{RCTs}: 0.82 (95% CI 0.56 to 1.21, I^2 not reported, 6 studies); ROM_{OBS}: 1.11 (95% CI 0.90 to 1.37, I^2 not reported, 5 studies); RROM: 1.35 (95% CI 0.87 to 2.10, I^2 not reported)
Moneer 2022	Time to symptom resolution	Observational studies and RCTs lead to similar results, a stronger effect was seen on evidence from observational studies.

(Continued)

		Chi^2 not reported. ROM_{RCTs}: 1.08 (95% CI 0.73 to 1.59, I^2 not reported, 2 studies); ROM_{OBS}: 1.17 (95% CI 0.98 to 1.40, I^2 not reported, 2 studies); RROM: 1.08 (95% CI 0.71 to 1.66, I^2 not reported)
Müeller 2010	Length of hospital stay (d)	A larger estimate was observed in observational studies compared to RCTs ($p < 0.001$). Chi^2 not reported. MD_{RCTs}: -2.0 (95% CI -2.6 to -1.4, I^2 = 98%, 20 studies); MD_{OBS}: -3.9 (95% CI -4.1 to -3.6, I^2 = 100%, 91 studies)
Müeller 2010	Duration of surgery (mins)	Observational studies and RCTs lead to similar results. Chi^2 not reported. MD_{RCTS}: 7.03 (95% CI -3.1 to 17.08, I^2 = 99%, 22 studies), MD_{OBS}: 7.05 (95% CI 0.3 to 14.07, I^2 = 100%, 59 studies)
Müeller 2010	Return to work (weeks)	Both RCTs' and observational studies' meta-analyses show a significant result in favour of the intervention, but a larger estimate was observed in observational studies compared to RCTs ($p = 0.018$). Chi^2 not reported. MD_{RCTs}: -1.2 (95% CI -1.9 to -0.6, I^2 = 98%, 10 studies); MD_{OBS}: -2.4 (95% CI -3.3 to -1.8, I^2 = 99%, 22 studies)
Naudet 2011	Difference between baseline and last assessment on the 17-item or 21-item Hamilton Rating Scale for Depression or the Montgomery and Asberg Rating Scale	Correlation coefficient suggesting that response to antidepressants is greater in RCTs than in observational studies. Correlation coefficient: 4.59 95% CI 2.61 to 6.56, 109 RCTs, 12 OBS
Otsuka 2022	Operative time	Observational studies and RCTs lead to similar results. Chi^2 not reported. MD_{RCTs}: 49.20 (95% CI 29.38 to 69.02, I^2 = 94%); MD_{cohort}: 47.85 (95% CI 29.37 to 66.33, I^2 = 96%); $MD_{case\text{-}control}$: 35.25 (95% CI 15.20 to 55.30, I^2 = 96%)
Otsuka 2022	Intraoperative blood loss	Observational studies and RCTs lead to similar results. Chi^2 not reported. MD_{RCTs}: -35.91 (95% CI -67.54 to -4.28, I^2 = 79%);

(Continued)

		MD_{cohort}: -179.3 ((95% CI -235.81 to -122.8, I^2 = 98%); $MD_{case\text{-}control}$: -44.89 (95% CI -64.65 to -25.12, I^2 = 81%)
Otsuka 2022	Postoperative hospital stay	Observational studies and RCTs lead to similar results. Chi^2 not reported. MD_{RCTs}: -0.73 (95% CI -1.28 to -0.19, I^2 = 26%); MD_{cohort}: -2.75 (95% CI -4.1 to -1.41, I^2 = 94%); $MD_{case\text{-}control}$: -2.49 (95% CI -3.84 to -1.13 I^2 = 81%)
Otsuka 2022	Number of retrieved lymph nodes	Observational studies and RCTs lead to similar results. Chi^2 not reported. MD_{RCTs}: -1.19 (95% CI -2.23 to -0.04, I^2 = 0%); MD_{cohort}: 0.21 (95% CI -2.16 to 2.58, I^2 = 83%); $MD_{case\text{-}control}$: -0.14 (95% CI -1.63 to 1.35, I^2 = 61%)
Virk 2019	Procedure duration	Comparison RCTS versus observational studies, test for subgroup differences: Chi^2 = 1.86, df = 1 (P = 0.17). Observational studies show a shorter procedure duration than RCTs. MD_{RCTs}: -7.87 (95% CI -17.48 to 1.74, I^2 = 61%, 7 studies); MD_{OBS}: -18.26 (95% CI -29.66 to -6.85, I^2 = 91%, 19 studies)
Virk 2019	Fluoroscopy duration	Comparison RCTS versus observational studies, test for subgroup differences: Chi^2 = 5.44, df = 1 (P = 0.02). Observational studies show a shorter fluoroscopy duration than RCTs. MD_{RCTs}: -1.73 (95% CI -3.67 to 0.22, I^2 = 0%, 6 studies); MD_{OBS}: -6.70 (95% CI -10.40 to -3.00, I^2 = 97%; 21 studies)
Virk 2019	Fluoroscopy exposure	Observational studies show a shorter fluoroscopy exposure than RCTs. Comparison RCTS versus observational studies, test for subgroup differences: Chi^2 = 0.78, df = 1 (P = 0.38). SMD_{RCTs}: -0.19 (95% CI -0.70 to 0.31, I^2 = 68%, 3 studies); SMD_{OBS}: -0.44 (95% CI -0.67 to -0.22; I^2 = 68%, 8 studies).
Virk 2019	Radiofrequency ablation duration	Observational studies show a shorter duration than RCTs. Comparison RCTS versus observational studies, test for subgroup differences: Chi^2 = 0.41, df = 1 (P = 0.52). MD_{RCTs}: 0.59 (95% CI -2.00 to 3.91, I^2 = 50%, 6 studies);

(Continued)

		MD_{OBS}: -4.70 (95% CI -8.28 to -1.13, I^2 = 89%, 16 studies)
Youn 2021	Length of hospital stay (number of days)	For length of stay, all study types except for propensity score-adjusted non-randomised studies significantly favoured transcatheter aortic valve implantation. Chi^2 not reported. MD_{RCTs}: -3.85 (95% CI -4.85 to -2.86, I^2 = 79%, 6 studies); MD_{OBS}: -3.01 (95% CI -6.01 to -0.00, I^2 = 99%, 10 studies).
	OBS: observational studies; ROM: ratio of means; RROM: ratio of ratio of means; SMD: standardised mean difference	

WHAT'S NEW

Date	Event	Description
4 January 2024	New search has been performed	New systematic searches were conducted with new search strategies; 35 new records were included; revised approach to statistical analysis; new analyses were conducted; assessment of the evidence with the GRADE approach; conclusions amended; implications revised; team of authors changed: new authors, DAS, JN, KG, LS, SB included; one previous author HH acknowledged.
4 January 2024	New citation required but conclusions have not changed	New citations were included in the review, leading to a slight amendment in the findings of the review, but not the conclusion.

HISTORY

Protocol first published: Issue 2, 2012
Review first published: Issue 4, 2014

CONTRIBUTIONS OF AUTHORS

All authors approve of the final version of the review.

Conception of the review: LB, AA

Design of the review: LB, AA, IT

Co-ordination of the review: IT

Search and selection of studies for inclusion in the review: IT, AA, JLZN, DA, KG, LS, LB

Collection of data for the review: IT, AA, JLZN, DA, LS, LB

Assessment of the risk of bias in the included studies: IT, AA, JLZN, DA, LS, LB

Analysis of data: SB, IT

Interpretation of data: IT, AA, SB, LS, LB

GRADEing of the evidence: IT, LS

Writing of the review: IT, AA, SB, LB

DECLARATIONS OF INTEREST

IT, AA, JLZN, DAS, KG and LB have no competing interests to declare.

LS and SB are authors of overviews of reviews that are included in this review. They were not involved in extracting data from nor assessing the quality of studies that they co-authored.

LB was not involved in the editorial process.

SOURCES OF SUPPORT

Internal sources

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

  The initial review had support from the Clinical and Translational Sciences Institute (CTSI), University of California, San Francisco (UCSF) for the review authors.

External sources

- The National Association of Statutory Health Insurance Funds, Germany

  The update of this Cochrane Methods review was partially funded by a grant from the National Association of Statutory Health Insurance Funds.

DIFFERENCES BETWEEN PROTOCOL AND REVIEW

The previously published version of this review searched for evidence from 01 January 1990 to 06 December 2013 in the Cochrane Methodology Register, Cochrane Database of Systematic Reviews, MEDLINE (via PubMed), Embase (via Embase.com), Literatura Latino-Americana y del Caribe en Ciencias de la Salud (LILACS), PsycINFO, and Web of Science/Web of Social Science (Anglemyer 2014).

Search strategy

For the review update, we developed and applied a revised search strategy for the searches in electronic databases. In addition, we revised the selection of databases for the search. The previous search strategy was very broad, yielding a very high number of retrieved records that seemed unreasonable to manage in an update search. By reproducing the PubMed search from Anglemyer 2014, 48,258 hits were retrieved for the given time frame (1990 to 2013). Extending the time frame to 2021, there were129,775 hits. The search methods reported in Anglemyer 2014 also did not report the retrieval of single database searches or the overall yield, but only the number of records after deduplication (n = 4406). We therefore revised the search strategy in an attempt to maintain and balance high sensitivity and specificity.

The decision to change from PubMed to Ovid MEDLINE was because of enhanced search functionalities that allowed more targeted queries (using the frequency and proximity operators in title and abstract).

Our current selection of databases is motivated by their focus on systematic reviews and their comprehensive coverage of health-related literature.

We explored the 'similar articles' feature of Ovid MEDLINE, and found that it increased the number of retrieved records but yielded a relatively low number of potentially relevant records. Thus, we decided not to use this feature in this update.

We did not search the Cochrane Methodology Register because it has not been systematically updated since 2011, according to the Cochrane Methods Group. We did not search LILACS, the Social Science Citation Index, Embase and PsycInfo because of the limited relevance of these databases' scope for the review topic and objective.

Data extraction

We expanded the data extraction tool to cover more information about the included reviews/overviews and their results. We used REDCap or MS Excel for data extraction.

Subgroup and sensitivity analyses

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

We added a subgroup analysis to analyse effect estimates from overviews of reviews separately. For the statistical analyses, we also used the ratio of ratios combining ratios of odds ratios, ratios of risk ratios, and ratios of hazard ratios. This is described in more detail in the Methods section.

We complemented the review's primary meta-analysis with two sensitivity analyses (see Sensitivity analysis): (1) we included findings from systematic reviews that employed a selective inversion approach in their analyses to test for the effect of their inclusion on the overall effect estimate; and (2) we included data on favourable outcomes in the analysis without the selective inversion.

We complemented the review's primary meta-analysis with a visual inspection of the dichotomous and continuous effect estimates from RCTs and observational studies.

INDEX TERMS

Medical Subject Headings (MeSH)

Bias; Case-Control Studies; *Delivery of Health Care; Observational Studies as Topic; Randomized Controlled Trials as Topic; Systematic Reviews as Topic

MeSH check words

Humans