BMC Medicine

RESEARCH

Open Access

Integration of non-randomized studies with randomized controlled trials in meta-analyses of clinical studies: a meta-epidemiological study on effect estimation of interventions

Fan Mei[1,2,3†], Minghong Yao[1,2,3†], Yuning Wang[1,2,3], Jiayidaer Huan[1,2,3], Yu Ma[1,2,3], Guowei Li[4,5,6], Kang Zou[1,2,3], Ling Li[1,2,3*] and Xin Sun[1,2,3,7*]

Abstract

Backgrounds Syntheses of non-randomized studies of interventions (NRSIs) and randomized controlled trials (RCTs) are increasingly used in decision-making. This study aimed to summarize when NRSIs are included in evidence syntheses of RCTs, with a particular focus on the methodological issues associated with combining NRSIs and RCTs.

Methods We searched PubMed to identify clinical systematic reviews published between 9 December 2017 and 9 December 2022, randomly sampling reviews in a 1:1 ratio of Core and non-Core clinical journals. We included systematic reviews with RCTs and NRSIs for the same clinical question. Clinical scenarios for considering the inclusion of NRSIs in eligible studies were classified. We extracted the methodological characteristics of the included studies, assessed the concordance of estimates between RCTs and NRSIs, calculated the ratio of the relative effect estimate from NRSIs to that from RCTs, and evaluated the impact on the estimates of pooled estimates when NRSIs are included.

Results Two hundred twenty systematic reviews were included in the analysis. The clinical scenarios for including NRSIs were grouped into four main justifications: adverse outcomes ($n=140$, 63.6%), long-term outcomes ($n=36$, 16.4%), the applicability of RCT results to broader populations ($n=11$, 5.0%), and other ($n=33$, 15.0%). When conducting a meta-analysis, none of these reviews assessed the compatibility of the different types of evidence prior, 203 (92.3%) combined estimates from RCTs and NRSIs in the same meta-analysis. Of the 203 studies, 169 (76.8%) used crude estimates of NRSIs, and 28 (13.8%) combined RCTs and multiple types of NRSIs. Seventy-seven studies (35.5%) showed "qualitative disagree" between estimates from RCTs and NRSIs, and 101 studies (46.5%) found "important difference". The integration of NRSIs changed the qualitative direction of estimates from RCTs in 72 out of 200 studies (36.0%).

†Fan Mei and Minghong Yao contributed equally to the study and are joint first authors.

*Correspondence:
Ling Li
liling@wchscu.cn
Xin Sun
sunxin@wchscu.cn
Full list of author information is available at the end of the article

Conclusions Systematic reviews typically include NRSIs in the context of assessing adverse or long-term outcomes. The inclusion of NRSIs in a meta-analysis of RCTs has a substantial impact on effect estimates, but discrepancies between RCTs and NRSIs are often ignored. Our proposed recommendations will help researchers to consider carefully when and how to synthesis evidence from RCTs and NRSIs.

Keywords Randomized controlled trials, Non-randomized studies of interventions, Systematic review, Meta-analysis, Meta-epidemiology

Background

Systematic reviews of randomized controlled trials (RCTs) are an established approach to synthesizing evidence on the relative effects of health interventions [1, 2]. Recently, non-randomized studies of interventions (NRSIs) have gained increasing attention in health decision-making. Indeed, NRSIs may provide valuable insights into the relative effects of health interventions, for example, by including more diverse study participants [3] and providing useful information in cases where RCTs are less likely (e.g., for adverse or long-term outcomes), or not feasible (e.g., unethical to conduct) [4], or provide additional evidence to assess population heterogeneity [5]. A recent international survey showed that 84% of experts from academic authoritative bodies (e.g., the Cochrane, Guideline International Network) agreed with the inclusion of NRSIs when assessing the effect of adverse or long-term outcomes, and 71.5% agreed with the inclusion of NRSIs as a surrogate evidence when RCTs are unavailable, of poor quality, or unethical [6].

Despite the potential benefits, the inclusion of NRSIs in systematic reviews increases methodological complexity, and the decision to include NRSIs and the subsequent implementation are often challenging [7–10]. International communities have presented distinct perspectives on when to include NRSIs. For instance, the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) group recommended that NRSIs should be considered when RCTs are unable to address clinical questions or have serious indirectness [4, 11]. The US Agency for Healthcare Research and Quality (AHRQ) additionally advised researchers to include NRSIs when extending the generalizability of RCT findings [12]. However, there is a lack of understanding of the rationale for including NRSIs in empirical studies.

The other major issue is how to synthesis evidence from NRSIs into the systematic review of RCTs, for two main reasons. One is that such studies—which are heterogeneous in study design—are inherently subject to bias due to the lack of randomization [13]. The other is that the approaches to synthesizing randomized and non-randomized studies—which vary widely in current practice [6, 14, 15]—may also produce conflicting results [16]. Although there are significant concerns arise at all stages of the synthesis of RCTs and NRSIs, such as which types of NRSIs are included and how data from RCTs and NRSIs are analyzed, current research continues to combine RCTs and NRSIs without distinction or further consideration [17–19]. Previous efforts have summarized the reporting of studies that included NRSIs, which were not complete, restricted to reviews published in high-impact journals, or only evaluated the general systematic review process, limiting the generalizability of their findings [17, 18]. To date, no study has yet systematically examined the rationale and characteristics of NRSIs inclusion and whether the potential impact of NRSIs on the body of evidence was considered in meta-analyses.

There is a clear gap in reconciling different study designs and, more urgently, in integrating evidence from NRSIs and RCTs. We therefore conducted a meta-epidemiological study to systematically examine when and how NRSIs are included in systematic reviews of RCTs, to quantify their impact on estimates, and to provide recommendations for maximizing the value of "best available" evidence.

Methods

Design

This study is part of a larger project to assess the impact of including NRSIs in a meta-analysis of RCTs [20]. Our published protocol includes detailed information on the definitions, eligibility criteria, literature search, study process, study screening, data abstraction, and data analysis [20].

Search strategies and selection of eligible systematic reviews

We searched PubMed for clinical systematic reviews published between 9 December 2017 and 9 December 2022 (search strategies see Additional file 1: Supplementary Methods [Search Strategy]). We randomly selected 220 journal articles, with a 1:1 stratification based on journal type (Core and non-Core), as defined by the US National Library of Medicine and the National Institutes of Health [21]. We included a systematic review if it included RCTs and the following types of NRSIs for at least one outcome: nonrandomized controlled trials, cohort studies, case–control studies. Network meta-analyses, individual

participant meta-analyses, and dose–response meta-analyses were not considered. The outcome of the meta-analysis is binary and indicates the benefit or safety of a treatment or prevention intervention. The title and abstract were screened independently by two reviewers (FM, YW), and the full text of eligible reviews was then examined. Disagreements were resolved by discussion or adjudicated by one of two arbitrators (MY, LL).

Data extraction

Paired reviewers, trained in the methodology, abstracted the data independently and in duplicate. We used a pilot-tested, standardized data abstraction form, together with detailed instructions for title, abstract, and full-text screening and data extraction. To ensure reliability, we performed calibration exercises before data abstraction. Disagreements were resolved by one of two arbitrators (MY, LL).

A primary outcome was selected for each review, according to a previously published strategy [20]: if a systematic review reported a single primary outcome, we selected this as the primary outcome for our analyses; if a systematic review reported more than one eligible primary outcome, we selected the first one reported in the results that met the eligibility criteria.

To determine the completeness of the items, we developed extracted items according to the guidelines provided by the AHRQ [22] and the Cochrane Handbook [23]. The following three categories of items were extracted from each eligible systematic review:

- *Study characteristics*: name of the first author, location of the first author (WHO region), number of NRSIs and RCTs included, number of participants for RCTs and NRSIs, epidemiologists or statisticians involved, reporting guideline endorsement, area of diseases, type of outcome, type of intervention (pharmacological/surgery/medical device/other), type of journal (Core/non-Core journal), type of NRSI, type of funding, conflict of interest, patient and public involvement, etc. As for the area of diseases, we extracted diseases reported by systematic reviewers and then matched them to the disease category in the Medical Subject Headings (MeSH) categories. We identified whether a systematic review included a section on patient and public involvement in the main text [24].
- *The justifications for the inclusion of RCTs and NRSIs*: whether the rationale for the inclusion of NRSIs was provided, clinical scenarios for the inclusion of NRSIs in systematic reviews. For studies that were not reported in detail, we assessed the justification for including NRSIs according to the wording used in the reports and categorized them into different groups (Additional file 2: Table S1).
- *Process of conducting systematic reviews*:

Planning and identification of NRSI inclusion: availability of protocol/ registration, whether authors prespecified NRSI study design in the protocol or eligibility criteria, and whether specific search filters were used to identify NRSI.
Risk of bias and strength of evidence: tools used to assess the risk of bias (RoB) in RCTs and NRSIs, final strength of evidence rating.
Synthesis of results from RCTs and NRSIs:

(1) Assessment before conducting a meta-analysis: estimates of NRSIs used in meta-analysis, consideration of discrepancies between NRSIs and RCTs. We defined that discrepancies were considered when authors performed subgroup analyses or sensitivity analyses based on study type, explored sources of heterogeneity, or combined RCTs and NRSIs separately.
(2) The manner of NRSIs integrated into a meta-analysis: type of effect measure (risk ratio [RR], odds ratio [OR], hazard ratio [HR], risk difference [RD]), which types of NRSIs were included in the meta-analysis of RCTs, how RCTs and NRSIs were combined and how different types of NRSI designs were combined in the same metaanalysis, analytical approaches (analysis strategy, statistical methods, effect measure) used for the meta-analysis.
(3) Additional analyses: any analyses (heterogeneity tests, subgroup analyses, sensitivity analyses, publication bias assessments) employed to assess the effect of different study designs between RCTs and NRSIs and the RoB of NRSIs on the estimates.

Interpretation of results and conclusions: whether the potential impact of including NRSIs was explicitly stated in the discussion section, whether positive results were reported when RCT and NRSI results were inconsistent (for example, when RCTs and NRSIs were combined separately, the RCT results did not reach statistical significance, the NRSI achieved significant results, but the conclusion of the systematic review reported only significant results).

We extracted the effect estimates with the corresponding standard errors for each RCT and NRSI included in a meta-analysis when the NRSIs and RCTs were pooled using the aggregate data (e.g., inverse variance [IV] method); we abstracted the number of participants and the number of events in each group for each RCT and NRSI included in a meta-analysis when the NRSIs and RCTs were pooled using the event data (e.g., Mantel–Haenszel [MH] method). The detailed information extracted can be found in the published protocol [20].

Statistical analysis

We used descriptive analysis to summarize the general characteristics of the eligible systematic reviews. For categorical variables, we report frequencies and percentages. For continuous variables, we presented means (standard deviation) or medians (interquartile range [IQR]), which were not normally distributed. We compared the characteristics between Core and non-Core journals using the χ^2 test or Fisher's exact test for categorical variables, and the *t*-test or Mann–Whitney *U* test for continuous variables. We used R statistical software (version 4.1.1) for all analyses. All comparisons were two-tailed, and a *P* value of 0.05 or less than 0.05 was considered statistically significant.

We assessed the concordance of the evidence from NRSIs and RCTs. The results were considered to be "qualitatively consistent" if both RCTs and NRSIs found the same direction of effect, that is, a statistically significant increase, a statistically significant decrease, or no statistically significant difference [11]. If the results did not agree qualitatively, we compared the consistency of the direction of the effect. If the effect estimates (RR, OR, HR) from RCTs and NRSIs were not the same, we expressed both estimates in the same measure (OR) using an assumed control risk (ACR) [25]: $\mathrm{RR} = \frac{OR}{1-ACR(1-OR)}$

To quantify the magnitude of the difference between effect estimates from RCTs and NRSIs, we calculate the ratio of odds ratio (RoR) for the pooled effects from RCTs and NRSIs contributing to the meta-analysis of interest, with the pooled evidence from RCTs serving as the referent [11, 26]. RoRs indicated an "important difference" (<0.70 or >1.43) or not ($0.7 \leq \mathrm{RoR} \leq 1.43$) [27, 28]. In addition, we recorded the details of studies with larger or smaller differences (e.g., adjusted or unadjusted estimates of NRSI, type of NRSI). We further assessed the influence of combining RCTs and NRSIs on estimates by calculating the proportion of meta-analyses in which the inclusion of NRSIs changed the qualitative direction of estimates from RCTs.

Results

Of the 16,690 records identified during the literature search from 9 December 2017 to 9 December 2022, 1036 were excluded as duplicates, and 13,020 were excluded during the initial screening based on title and abstract. After the full-text screening, 255 systematic reviews were included. We then randomly selected 110 studies from Core journals and 110 studies from non-Core journals for inclusion in our study (Additional file 1: Supplementary Results, Additional file 3: Fig. S1).

General characteristics of systematic reviews and meta-analyses

Table 1 shows the characteristics of the 220 included clinical systematic reviews. The most common geographical region was the Western Pacific Region ($n=122$, 55.5%), followed by Europe ($n=46$, 20.9%) and South-East Asia Region ($n=37$, 16.8%). Only three studies (1.4%) stated the patient and public involvement.

The median number of studies included in the eligible reviews was 7 (IQR 5–10), of which the median number of studies was 3 (IQR 1–4) for RCTs and 4 (IQR 2–7) for NRSIs. The reviews included a median of 882 participants (IQR 292–2934), with a median of 413 participants (IQR 155–1157) for RCTs and 1605 participants (IQR 667–4707) for NRSIs. The intervention was classified as pharmacological ($n=90$, 40.9%) and surgical ($n=92$, 41.8%) in most reviews, and as a medical device in 12 reviews (5.5%). The most commonly selected primary outcome was morbidity ($n=104$, 47.3%). Regarding the type of NRSI, cohort studies were the most common type of NRSI included ($n=126$, 57.3%), followed by nonrandomized controlled trials ($n=26$, 11.8%) and case–control studies ($n=13$, 5.9%). For reviews with more than one type of NRSI, 47 (21.4%) mixed two types of NRSI, and 8 (3.6%) mixed at least three types of NRSI. Full details of the distribution of NRSI types were provided in Additional file 3: Fig. S2. No statistically significant differences were found when comparing these characteristics between Core and non-Core systematic reviews (Table 1).

Rationale for the inclusion of NRSIs

The justification of clinical scenarios is shown in Fig. 1. Only 30.0% ($n=66$) of the systematic reviews provide a rationale for the inclusion of NRSIs. Among these 66 systematic reviews, the most common clinical scenarios including NRSIs were adverse outcomes ($n=41$, 62.1%), long-term outcomes ($n=5$, 7.6%), and the applicability of RCT results to broader populations ($n=6$, 9.1%). We categorized other situations that may vary with the clinical question, such as a limited number of RCTs, as "others" ($n=14$, 21.2%). After reclassification of

Table 1 General characteristics of included systematic review and meta-analyses

Characteristics	Systematic reviews, No. (%)			
	Overall (*N*=220)	Core journal (*n*=110)	Non-core journal (*n*=110)	*P* value
Median (IQR) No of studies included				
Total	7 (5–10)	7 (5–10)	8 (5–10)	0.928
RCTs	3 (1–4)	3 (2–4)	2 (1–4)	0.215
NRSIs	4 (2–7)	4 (2–6)	4 (2–7)	0.808
Median (IQR) No of participants included				
Total	882 (292–2934)	848 (294–3028)	916 (296–2920)	0.958
RCTs	413 (155–1157)	483 (163–1182)	376 (144–1038)	0.569
NRSIs	1605 (667–4707)	1605 (613–4854)	1604 (711–4313)	0.931
Epidemiologists or statisticians involved	49 (22.3)	20 (18.2)	29 (26.4)	0.195
Reporting guideline endorsement				0.904[a]
PRISMA	146 (66.4)	71 (64.5)	75 (68.2)	
MOOSE	3 (1.4)	2 (1.8)	1 (0.9)	
PRISMA and MOOSE	25 (11.4)	13 (11.8)	12 (10.9)	
QUORUM	1 (0.5)	0 (0.0)	1 (0.9)	
Not reported	45 (20.5)	24 (21.8)	21 (19.1)	
Assessment objectives of studies included				0.587
Efficacy/effectiveness	123 (55.9)	59 (53.6)	64 (58.2)	
Safety/harm	97 (44.1)	51 (46.4)	46 (41.8)	
Type(s) of disease				0.253[a]
Cardiology	29 (13.2)	11 (10.0)	18 (16.4)	
General	36 (16.4)	20 (18.2)	16 (14.5)	
Infectious diseases	14 (6.4)	3 (2.7)	11 (10.0)	
Neurology	12 (5.5)	6 (5.5)	6 (5.5)	
Oncology	14 (6.4)	7 (6.4)	7 (6.4)	
Orthopedics	34 (15.5)	19 (17.3)	15 (13.6)	
Other	81 (36.8)	44 (40.0)	37 (33.6)	
Type(s) of intervention/exposure				0.284[a]
Pharmacological	90 (40.9)	50 (45.5)	40 (36.4)	
Surgery	92 (41.8)	46 (41.8)	46 (41.8)	
Medical device	12 (5.5)	4 (3.6)	8 (7.3)	
Other	26 (11.8)	10 (9.1)	16 (14.5)	
Type(s) of outcome				0.804[a]
Mortality	44 (20.0)	22 (20.0)	22 (20.0)	
Morbidity	104 (47.3)	54 (49.1)	50 (45.5)	
Surrogate outcome	4 (1.8)	3 (2.7)	1 (0.9)	
Symptoms/Quality of life/Functional status	5 (2.3)	2 (1.8)	3 (2.7)	
Other	63 (28.6)	29 (26.4)	34 (30.9)	
Type(s) of NRSIs included				0.988
Cohort only	126 (57.3)	64 (58.2)	62 (56.4)	
Case-control only	13 (5.9)	6 (5.5)	7 (6.4)	
Nonrandomized controlled trials only	26 (11.8)	13 (11.8)	13 (11.8)	
Multiple types of NRSIs				
Mixed two types of NRSIs (e.g., cohort and case series)	47 (21.4)	22 (20.0)	25 (22.7)	
Mixed at least three types of NRSIs (e.g., nonrandomized controlled trials, cohort, and case-control)	8 (3.6)	4 (3.6)	4 (3.6)	
Location (WHO region)				0.099[a]
Western Pacific Region	122 (55.5)	69 (62.7)	53 (48.2)	
European Region	46 (20.9)	23 (20.9)	23 (20.9)	

Table 1 (continued)

Characteristics	Systematic reviews, No. (%)			
	Overall (*N*=220)	Core journal (*n*=110)	Non-core journal (*n*=110)	*P* value
South-East Asia Region	37 (16.8)	15 (13.6)	22 (20.0)	
Region of the Americas	8 (3.6)	1 (0.9)	7 (6.4)	
Eastern Mediterranean Region	4 (1.8)	1 (0.9)	3 (2.7)	
African Region	3 (1.4)	1 (0.9)	2 (1.8)	
Source of funding				0.757[a]
Private not for profit	34 (15.5)	14 (12.7)	20 (18.2)	
Private for profit	5 (2.3)	2 (1.8)	3 (2.7)	
Government	50 (22.7)	26 (23.6)	24 (21.8)	
No funded	60 (27.3)	33 (30.0)	27 (24.5)	
Funding not reported	71 (32.3)	35 (31.8)	36 (32.7)	
Conflict of interest claimed				0.544[a]
Present	8 (3.6)	5 (4.5)	3 (2.7)	
Not present	194 (88.2)	98 (89.1)	96 (87.3)	
Not declared	18 (8.2)	7 (6.4)	11 (10.0)	
Patient and public involvement stated	3 (1.4)	2 (1.8)	1 (0.9)	1

IQR interquartile range, *WHO* World Health Organization

[a] Fisher exact test

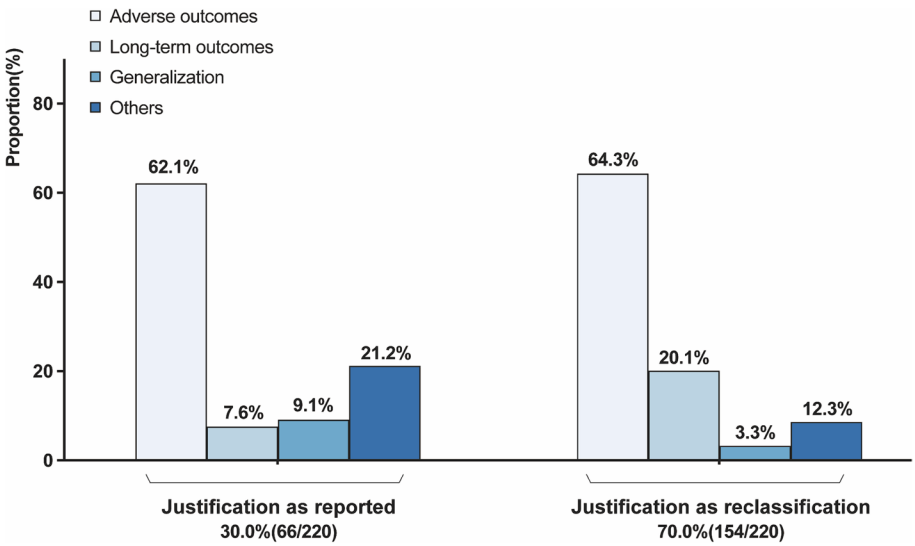

Fig. 1 Justifications for including NRSIs in the included studies (criteria see Additional file 2: Table S1)

154 reviews that did not report the rationale for the inclusion of NRSI, we obtained similar results that most studies considered evidence from NRSI in the assessment of adverse outcomes ($n=99$, 64.3%), followed by long-term outcomes ($n=31$, 20.1%), other ($n=19$, 12.3%), and generalizability ($n=5$, 3.3%). The details of reclassification are displayed in Additional file 2: Table S2. A similar distribution was found across different types of interventions and types of NRSI, with adverse outcomes accounting for the majority of studies (Additional file 3: Fig. S3).

Processes for conducting systematic reviews

The characteristics of the design and conduct of systematic reviews are presented in Table 2, the considerations for meta-analysis in Tables 3 and 4, and the interpretation and

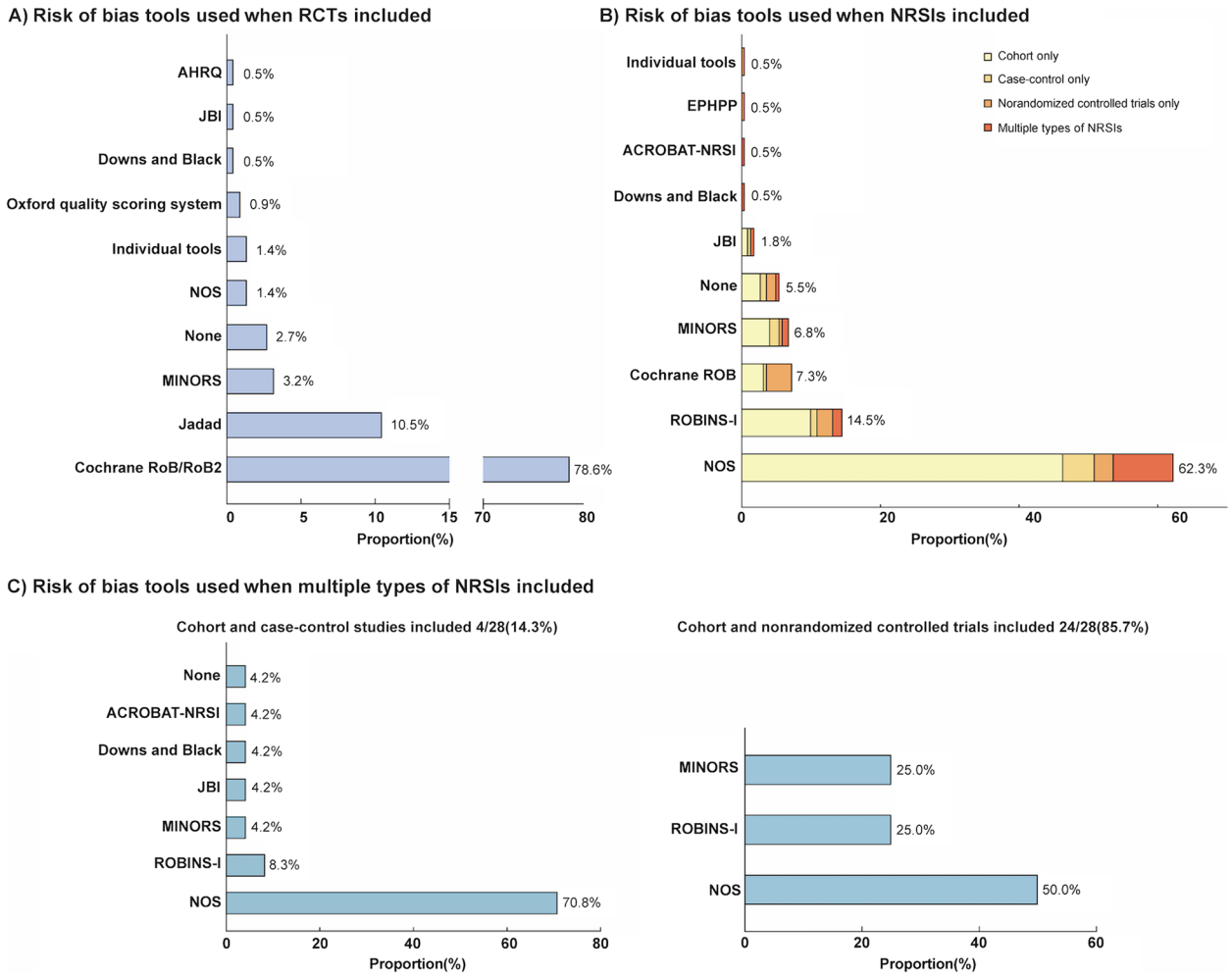

Fig. 2 Overview of the tools used for the assessment of risk of bias. AHRQ = Agency for Healthcare Research and Quality, JBI = Joanna Briggs Institute, NOS = Newcastle Ottawa Scale, MINORS = Methodological index for non-randomized studies, ACROBAT-NRSI = A Cochrane Risk Of Bias Assessment Tool for Non-Randomized Studies, EPHPP = Effective Public Health Practice Project, ROBINS-I = Risk Of Bias In Non-randomized Studies of Interventions. Individual tools refer to tools that the authors themselves have formulated

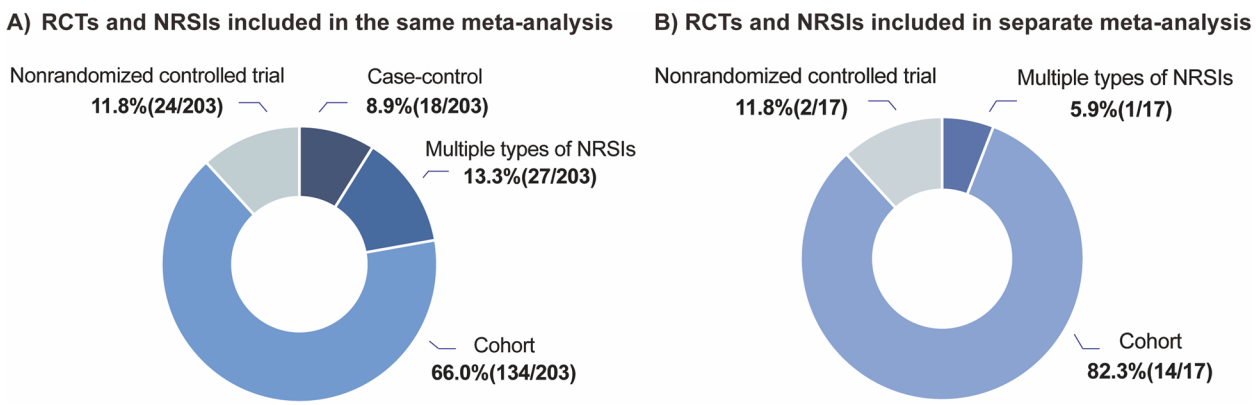

Fig. 3 Consideration of different types of NRSIs in the process of meta-analysis

Table 2 Methods for conducting systematic reviews

Review characteristics		Systematic reviews, No. (%)			*P* value
		Overall (*N*=220)	Core journal (*n*=110)	Non-core journal (*n*=110)	
Protocol	**Is the protocol/registration available**				1
	Yes	84 (38.2)	42 (38.2)	42 (38.2)	
	Where the protocol could be cited				
	Registry platform	71 (32.3)	39 (35.5)	32 (29.1)	
	Published Journal articles	8 (3.6)	2 (1.8)	6 (5.5)	
	Supplementary	5 (2.3)	1 (0.9)	4 (3.6)	
	Did the eligible study design features of NRSI in the protocol stated				
	Specific one study design (e.g., cohort)	20 (9.1)	9 (8.2)	11 (10.0)	
	Specific multiple study design (e.g., cohort and/or case-control)	28 (12.7)	15 (13.6)	13 (11.8)	
	Unclear definition (e.g., retrospective study orcomparative study)	22 (10.0)	8 (7.3)	14 (12.7)	
	Not reported/No restriction	14 (6.4)	10 (9.1)	4 (3.6)	
	No	136 (61.8)	68 (61.8)	68 (61.8)	
Eligibility criteria	**Did the study design features of NRSI in the systematic review eligibility criteria stated**				0.690
	Specific one study design (e.g., cohort)	65 (29.5)	33 (30.0)	32 (29.1)	
	Specific multiple study design (e.g., cohort and/or case-control)	46 (20.9)	26 (23.6)	20 (18.2)	
	Unclear definition (e.g., retrospective study or comparative study)	68 (30.9)	33 (30.0)	35 (31.8)	
	Not reported/No restriction	41 (18.6)	18 (16.4)	23 (20.9)	
Search strategies	**Specific terms for identifying NRSI were used (e.g., "cohort studies")**	16 (7.3)	8 (7.3)	8 (7.3)	1
Certainty of evidence	**Certainty of evidence assessed**				0.734
	Yes	43 (19.5)	20 (18.2)	23 (20.9)	
	Which evidence grading system were used				
	GRADE approach				
	RCTs only	1 (2.6)	0 (0.0)	1 (5.0)	
	NRSIs only	1 (2.6)	1 (5.3)	0 (0.0)	
	Pooled only	18 (46.2)	9 (47.4)	9 (45.0)	
	RCTs and NRSIs separately without pooled	16 (41.0)	8 (42.1)	8 (40.0)	
	RCTs and NRSIs separately with pooled	3 (7.7)	1 (5.3)	2 (10.0)	
	Other approach (e.g., AHRQ, OCEBM)	4 (1.8)	2 (1.8)	2 (1.8)	
	Final rating for pooled body of evidence[a]				
	Very low	4 (19.0)	2 (20.0)	2 (18.2)	
	Low	9 (42.9)	3 (30.0)	6 (54.5)	
	Moderate	7 (33.3)	5 (50.0)	2 (18.2)	
	High	1 (4.8)	0 (0.0)	1 (9.1)	
	No	177 (80.5)	90 (81.8)	87 (79.1)	

GRADE Grading of Recommendations, Assessment, Development, and Evaluations, *AHRQ* Agency for Healthcare Research and Quality, *OCEBM* Oxford Centre for Evidence-Based Medicine

[a] Only studies provided pooled certainty of evidence was collected here, so the overall sample size for final grade was 21

conclusions in Table 5. The distribution of risk of bias assessment tools is shown in Fig. 2. Figure 3 shows the consideration of different types of NRSIs in the meta-analysis process.

Planning and identifying for the inclusion of NRSIs

Only 38.2% ($n=84$) of the datasets had an available registration or protocol, and the majority ($n=71$, 32.3%) were registered on the registry platform. Of those with protocols, 12.7% ($n=28$) preferred to claim to include specific multiple study designs of NRSI in advance, and 10.0% ($n=22$) did not have a clear statement of study design, such as "*retrospective study*" (no details) or "*comparative study*" (no details). In the eligibility criteria of the reviews, almost a-third of the studies ($n=68$, 30.9%) did

not clearly define the types of NRSI. Few studies ($n=16$, 7.3%) used specific terms to identify NRSI in their search strategies (Table 2).

Assessment of risk of bias and grading of the strength of evidence

The Cochrane Risk of Bias tool for RCTs ($n=173$, 78.6%) and the Newcastle–Ottawa Scale (NOS) for NRSIs ($n=137$, 62.3%) were the most commonly used tools for assessing RoB. Inappropriate use of these tools was common, with 1.4% ($n=3$) of reviews assessing the risk of bias in RCTs using NOS and 7.3% ($n=16$) of reviews assessing risk of bias in NRSIs using the Cochrane Risk of Bias tool. NOS was most preferred when more than one type of NRSI was included, especially when cohort and case–control studies were included ($n=17$, 70.8%) (Fig. 2). Forty-three (19.5%) reviews attempted to rate the certainty of the evidence using GRADE or other approaches, and 16 (41.0%) of these rated the certainty of the evidence from RCTs and NRSIs separately. For 21 studies that were assessed together (3 studies assessed RCTs and NRSIs separately with pooled, 18 studies assessed pooled only), the certainty of evidence was rated as very low ($n=4$, 19.0%), low ($n=9$, 42.9%), moderate ($n=7$, 33.3%), and high ($n=1$, 4.8%) (Table 2).

Synthesis of results from RCTs and NRSIs

Assessment before conducting meta-analysis

None of the studies assessed the similarity between RCTs and NRSIs, and only 18.2% ($n=40$) considered the RoB of NRSIs in the meta-analysis. For the pooled estimates analyzed, we found that 169 studies (76.8%) used crude estimates of NRSIs when combining RCTs and NRSIs in a meta-analysis, and only 2 studies (0.9%) clearly reported adjustment for prespecified important confounders (Table 3).

The manner of NRSIs integrated into a meta-analysis

Of the 220 systematic reviews, 129 (58.6%) combined RCTs and NRSIs in the same meta-analysis without subgroup analysis, 74 (33.6%) combined RCTs and NRSIs together with a subgroup analysis, and 17 (7.7%) analyzed RCTs and NRSIs separately (one meta-analysis for NRSIs and one for RCTs). About half ($n=101$, 45.9%) of them considered the discrepancy between RCTs and NRSIs. When meta-analyses included multiple NRSI designs, 12.3% ($n=25$) of the studies directly combined the results without considering the type of NRSI (Table 3). The detailed information on the risk of bias assessment tools and certainty of evidence is provided in Additional file 1: Table S3. Among the NRSI designs analyzed, cohort studies were most often combined with RCTs, whether integrated into the same meta-analysis ($n=134$, 66.0%) or separately ($n=14$, 82.3%) (Fig. 3).

The most commonly used effect measure was OR in 114 (51.6%) meta-analyses, followed by RR ($n=93$, 42.1%). Common statistical models used for meta-analysis were the random-effects model ($n=139$, 63.2%) and the fixed-effects model ($n=71$, 32.3%). The most frequently used meta-analysis method was the MH method ($n=147$, 66.8%), followed by the IV method ($n=29$, 13.2%) (Table 3).

Additional analysis

Sources of heterogeneity were investigated by subgroup analysis in 74 studies (33.6%) and by meta-regression in 18 studies (8.2%). Of the studies that explored heterogeneity, only 2 studies (0.9%) that performed subgroup analysis and 3 studies (1.4%) that performed meta-regression addressed the risk of bias. Subgroup analyses were planned in 90 studies (40.9%) and post hoc analyses in 35 studies (15.9%). One fifth studies ($n=49$, 22.3%) only presented subgroup analyses according to study design. Sensitivity analyses were performed by the study design in 5 studies (2.3%) and by risk of bias in 13 studies (5.9%). More than a third of the meta-analyses ($n=99$, 45.0%) did not assess publication bias at all. Of the 121 studies that assessed publication bias, 120 (99.2%) assessed publication bias without distinguishing between RCTs and NRSIs, and 89 (72.3%) assessed asymmetry in funnel plots only (Table 4).

Concordance between effect estimates from NRSIs and RCTs

Two hundred seventeen meta-analyses were used to quantify the concordance of the evidence from NRSIs and RCTs, after excluding three meta-analyses due to insufficient original data for each included study. More than half of the estimates from NRSIs ($n=140$, 64.5%) were "qualitatively agree" with those from RCTs. Of the remaining 77 (35.5%) studies with "qualitatively disagree", 66 (85.7%) were inconsistent in statistical significance but consistent in direction of effect, and 11 (14.3%) of RCTs and NRSIs had treatment effects that were inconsistent in both direction of effect and statistical significance (Table 5). In 68 out of 77 studies (88.3%), the RCTs did not reject the null hypothesis, whereas the NRSI rejected the null hypothesis (Additional file 2: Table S4).

In 101 of 217 studies (46.5%), the estimates of the NRSI were "important different" from those of the RCTs. Of these studies with larger differences, only 12 (11.9%) pooled adjusted estimates of NRSIs, and none of the included studies accounted for discrepancies in the type of NRSI. The percentage of "important different" estimates was 75.2% ($n=76$) for estimates from RCTs and cohort studies, and 8.9% ($n=9$) for estimates from RCTs and multiple NRSI types.

Table 3 Methods for meta-analysis

Meta-analysis characteristics		Systematic reviews, No. (%)			*P* value
		Overall (*N*=220)	Core journal (*n*=110)	Non-core journal (*n*=110)	
Assessment before undertaking a meta-analysis	**Risk of bias of NRSIs considered in meta-analysis**				0.116
	Yes (Methods for addressing risk of bias)	40 (18.2)	15 (13.6)	25 (22.7)	
	Only including studies with a low risk of bias in the primary analysis	1 (0.5)	1 (0.9)	0 (0.0)	
	Considering the risk of bias in additional analysis (e.g., subgroup analysis, sensitivity analysis)	36 (16.4)	13 (11.8)	23 (20.9)	
	Both	3 (1.4)	1 (0.9)	2 (1.8)	
	No	180 (81.8)	95 (86.4)	85 (77.3)	
	Whether adjusted effect estimates of NRSIs used				0.864 [a]
	Yes (Methods for adjusting effect estimates of NRSIs)	21 (9.5)	9 (8.2)	12 (10.9)	
	Adjusted important confounders prespecified	2 (0.9)	2 (0.9)	0 (0.0)	
	Not reported	19 (8.6)	7 (6.4)	12 (10.9)	
	No (Crude estimate)	169 (76.8)	86 (78.2)	83 (75.5)	
	Both (Some provide adjusted estimate, the other provide crude estimate)	5 (2.3)	3 (2.7)	2 (1.8)	
	Not specified	25 (11.4)	12 (10.9)	13 (11.8)	
The manner of NRSIs integrated into a meta-analysis	**Method of data synthesis reported**				0.691
	RCTs and NRSIs combined in the same meta-analysis	203 (92.3)	101 (91.8)	102 (92.7)	
	Without subgroup analysis	129 (58.6)	67 (60.9)	62 (56.4)	
	With subgroup analysis	74 (33.6)	34 (30.9)	40 (36.4)	
	RCTs and NRSIs combined separately (one for NRSIs and one for RCTs)	17 (7.7)	9 (8.2)	8 (7.3)	
	Discrepancy between RCTs and NRSIs considered	101 (45.9)	51 (46.4)	50 (45.5)	1
	Whether combined RCTs and multiple types of NRSIs[b]				0.312
	Yes	28 (13.8)	11 (10.9)	17 (16.7)	
	Directly combined without considering the study type of NRSIs	25 (12.3)	10 (9.9)	15 (14.7)	
	Combined cohorts and case-control studies	22 (10.8)	9 (8.9)	13 (12.7)	
	Combined cohorts and nonrandomized controlled trials	3 (1.5)	1 (1.0)	2 (2.0)	
	Results in each type of NRSIs were synthesized separately	2 (1.0)	0 (0.0)	2 (2.0)	
	Both synthesized separately and directly	1 (0.5)	1 (1.0)	0 (0.0)	
	No	175 (86.2)	90 (89.1)	85 (83.3)	
	Effect measures used[c]				0.245 [a]
	Risk ratio	93 (42.1)	42 (37.8)	51 (46.4)	
	Odds ratio	114 (51.6)	59 (53.2)	55 (50.0)	
	Hazard ratio	13 (5.9)	9 (8.1)	4 (3.6)	
	Risk difference	1 (0.5)	1 (0.9)	0 (0.0)	
	Analysis model used				0.627[a]
	Fixed effect model	71 (32.3)	36 (32.7)	35 (31.8)	
	Random effect model	139 (63.2)	67 (60.9)	72 (65.5)	
	Both fixed and random effects	7 (3.2)	5 (4.5)	2 (1.8)	
	Not reported	3 (1.4)	2 (1.8)	1 (0.9)	
	Statistical method used				0.189
	Mantel-Haenszel	147 (66.8)	76 (69.1)	71 (64.5)	
	Inverse variance	29 (13.2)	10 (9.1)	19 (17.3)	
	Other	44 (20.0)	24 (21.8)	20 (18.2)	

[a] Fisher exact test

[b] The integration of RCT and multiple types of NRSIs only existed in the 203 articles in which RCT and NRSI were combined in a same meta-analysis

[c] This includes one study in which the effect measure for RCTs was the risk ratio, and the effect measure for NRSIs was the odds ratio, so the overall sample size for effect estimates was 221

The impact of combining RCTs and NRSIs on estimates

Two hundred meta-analyses that pooled RCTs and NRSIs in the same meta-analysis without subgroups were used to quantify changes in the effect estimates after including NRSIs, of which 72 (36.0%) showed "qualitatively disagree" results. In 63 of 72 studies (87.5%), the estimates of RCTs changed from including the null effect to excluding the null effect after pooling RCTs and NRSIs (Additional file2: Table S5).

Interpretation and conclusions

In 41 studies that pooled RCTs and NRSIs in the same meta-analysis with subgroup analysis or analyzed separately with inconsistent results, we observed that most review authors ($n=38$, 92.7%) reported only positive results in their conclusion (Additional file3: Fig. S4). Twelve out of 38 estimates (31.6%) were in the opposite direction of effect (Table 5).

Discussion

Summary of findings

This study comprehensively outlined the characteristics of NRSI inclusion in systematic reviews based on a large-scale empirical dataset. Our findings identified the main justifications for including NRSIs in the systematic reviews of RCTs, including adverse outcomes, long-term outcomes, and generalizability. Methodological issues related to design, conduct, analysis, and interpretation are widespread. For example, 154 (70.0%) did not provide a rationale for the inclusion of NRSI, 68 (30.9%) did not clearly define the design type of NRSI in the eligibility criteria, 169 (76.8%) combined crude estimates of NRSIs with RCTs, and 129 (58.6%) combined RCTs and NRSIs in the same meta-analysis without distinction, 38 (92.7%) of the authors likely reported positive results, all of which exacerbated the gap in synthesizing multiple sources of evidence.

Our study summarized the clinical scenarios in which NRSIs were included in the meta-analysis of RCTs into four classifications according to the GRADE and AHRQ guidelines [4, 7]. NRSIs were often considered when existing RCTs answered questions about adverse outcomes (63.6%), long-term outcomes (16.4%), and the generalizability of RCT results to broader populations (5.0%) in empirical analyses based on reporting or reclassification. Adverse outcomes accounted for most of the included studies, regardless of whether different interventions or types of NRSI designs were evaluated. An important reason for this may be that RCTs are typically underpowered to detect adverse effects due to insufficient sample size or follow-up, and that patient groups at high risk of adverse effects, such as the elderly, pregnant women, and people with comorbidities, may go undetected in RCTs [11, 29]. In comparison, NRSIs can usually serve as a complement, with a larger sample size, longer follow-up duration, and a more representative population [11]. This classification provides a clear insight into when evidence from NRSIs can be considered, and a rationale for methodologists to explore quantitative methods for combining NRSIs and RCTs in different settings.

We identified methodological issues related to the planning and conduct of the inclusion of NRSIs in a systematic review of RCTs. Protocol/registration is available for only 84 (38.2%) studies. Twenty-two (10.0%) and 68 (30.9%) systematic reviews did not clearly define the study characteristics of NRSIs in the protocol and the systematic review eligibility criteria respectively, such as "*observational studies*" or "*comparative* studies". Various tools were used to assess the RoB of RCTs and NRSIs, the most commonly used being the Cochrane Risk of Bias tool and the NOS. Notably, although the Cochrane Risk of Bias tool was explicitly designed to assess the risk of bias in RCTs [30], several reviews have inappropriately applied it to NRSIs. The problem is also serious for the quality assessment of RCTs. Inadequate RoB assessment can directly influence which studies are included in the evidence synthesis and substantially affect the results of the reviews [31].

When data from NRSIs were including in a meta-analysis of RCTs, only 0.9% assessed the RoB of NRSIs in the primary analysis, and less than 1.0% adjusted NRSIs for presumed important confounders. An important caveat to this finding is that combining effect estimates across studies is rarely justified, as estimated effects for NRSIs with different study design characteristics may be influenced by different sources of bias [23, 32]. For example, an NRSI study with poor methodological quality but a large sample size may dominate the overall estimates, further reducing the certainty of the evidence [16, 33, 34]. However, a substantial proportion of studies directly combined the estimates from RCTs and NRSIs in the same meta-analysis, or combined multiple types of NRSIs without distinction. Another concern is that the treatment effect from NRSIs was rarely interpreted. About 1.4% considered study design characteristics when identifying sources of heterogeneity by performing meta-regression, and 99.2% performed publication bias tests without distinguishing between the two types of evidence, even though they were considered heterogeneous and influential [12, 35].

There are fundamental differences between RCTs and NRSIs in design, conduct, data collection, analysis, etc. [4, 27, 35]. These differences can raise questions about potential bias and conflicting evidence between studies [16]. When analyzing the concordance of the estimates from NRSIs and RCTs, 35.5% showed "qualitative

Table 4 Additional analysis

Additional analysis characteristics		Systematic reviews, No. (%)			*P* value
		Overall (*N*=220)	Core journal (*n*=110)	Non-core journal (*n*=110)	
Heterogeneity test	**Heterogeneity test conducted**				0.127
	Yes (Identification of sources of heterogeneity)	84 (38.2)	36 (32.7)	48 (43.6)	
	Subgroup analyses conducted	74 (33.6)	32 (29.1)	42 (38.2)	
	Based on study design (and other)	38 (17.3)	15 (13.6)	23 (20.9)	
	Based on risk of bias (and other)	2 (0.9)	0 (0.0)	2 (1.8)	
	Based on both study design and risk of bias (and other)	4 (1.8)	1 (0.9)	3 (2.7)	
	Others	30 (13.6)	16 (14.5)	14 (12.7)	
	Meta-regression conducted	18 (8.2)	8 (7.3)	10 (9.1)	
	Consider study designs as an explanatory variable	3 (1.4)	2 (1.8)	1 (0.9)	
	Consider risk of bias as an explanatory variable	3 (1.4)	1 (0.9)	2 (1.8)	
	Others	12 (5.5)	5 (4.5)	7 (6.4)	
	No	136 (61.8)	74 (67.3)	62 (56.4)	
Subgroup analyses	**Subgroup analyses conducted**				0.891
	Yes	132 (60.0)	67 (60.9)	65 (59.1)	
	Subgroup analyses prespecified				
	All predefined	90 (40.9)	46 (41.8)	44 (40.0)	
	All post-hoc	35 (15.9)	18 (16.4)	17 (15.5)	
	Combined predefined and post-hoc	7 (3.2)	3 (2.7)	4 (3.6)	
	Type of subgroup analyses				
	Based on study design only	49 (22.3)	24 (21.8)	25 (22.7)	
	Based on type of intervention/comparison only	29 (13.2)	11 (10.0)	18 (16.4)	
	Based on multiple subgroup analyses (e.g., study design and risk of bias)	43 (19.5)	25 (22.7)	18 (16.4)	
	Others	11 (5.0)	7 (6.4)	4 (3.6)	
	No	88 (40.0)	43 (39.1)	45 (40.9)	
Sensitivity analyses	**Sensitivity analyses conducted**				0.057
	Yes (Type of sensitivity analyses)	97 (44.1)	56 (50.9)	41 (37.3)	
	Based on study design	5 (2.3)	1 (0.9)	4 (3.6)	
	Based on risk of bias	13 (5.9)	5 (4.5)	8 (7.3)	
	Based on study design and risk of bias	3 (1.4)	1 (0.9)	2 (1.8)	
	Others	76 (34.5)	49 (44.5)	27 (24.5)	
	No	123 (55.9)	54 (49.1)	69 (62.7)	
Publication bias	**Publication bias assessed**				0.104
	Yes	121 (55.0)	67 (60.9)	54 (49.1)	
	Methods used to assess publication bias				
	RCTs and NRSIs assessed together	120 (99.2)	67 (1.0)	53 (98.1)	
	RCTs and NRSIs assessed separately but different types of NRSIs assessed together	1 (0.8)	0 (0.0)	1 (1.9)	
	Method of quantifying publication bias[a]				
	Standard funnel plot	89 (72.3)	50 (72.5)	39 (70.9)	
	Begg's test	21 (17.1)	11 (16.2)	10 (18.2)	
	Egger's test	13 (10.6)	7 (10.3)	6 (10.9)	
	No	99 (45.0)	43 (39.1)	56 (50.9)	

[a] This includes two study in which both Begg's test and Egger's test were performed, so the overall sample size was 123

Table 5 Interpretation and conclusions

Characteristics	Systematic reviews, No. (%)			*P* value
	Overall (*N*=220)	Core journal (*n*=110)	Non-core journal (*n*=110)	
The effect estimates between RCTs and NRSIs were qualitatively agree[a,b]				0.142
Yes (qualitatively agree)	140 (64.5)	76 (69.7)	64 (59.3)	
Increased, significantly	17 (12.1)	8 (10.5)	9 (14.1)	
Decreased, significantly	38 (27.1)	24 (31.6)	14 (21.9)	
Not significantly	85 (60.7)	44 (57.9)	41 (64.1)	
No (qualitatively disagree)	77 (35.5)	33 (30.3)	44 (40.7)	
Opposite statistically significant, with concordant direction of effect	66 (85.7)	28 (84.8)	38 (86.4)	
Opposite statistically significant, with opposite direction of effect	11 (14.3)	5 (15.2)	6 (13.6)	
Magnitude of the difference between effect estimates from RCTs and NRSIs[b]				0.242
Important difference (0.7<RoR or RoR>1.43)	101 (46.5)	45 (42.1)	56 (50.9)	
Estimates of NRSIs adjusted	12 (11.9)	5 (11.1)	7 (12.5)	
Type of NRSIs included				
Nonrandomized controlled studies only	9 (8.9)	5 (11.1)	4 (7.1)	
Cohort studies only	76 (75.2)	34 (75.6)	42 (75.0)	
Case-control studies only	7 (6.9)	3 (6.7)	4 (7.1)	
Multiple NRSI type	9 (8.9)	3 (6.7)	6 (10.7)	
Discrepancy in NRSIs type were considered	0 (0.0)	0 (0.0)	0 (0.0)	
Not important difference (0.7≤RoR≤1.43)	116 (53.5)	62 (57.9)	54 (49.1)	
Estimates of NRSIs adjusted	14 (12.1)	7 (11.3)	7 (12.9)	
Type of NRSIs included				
Nonrandomized controlled studies only	16 (13.7)	8 (12.9)	8 (14.8)	
Cohort studies only	70 (60.3)	40 (64.5)	30 (55.6)	
Case-control studies only	11 (9.5)	6 (9.7)	5 (9.3)	
Multiple NRSI type	19 (16.4)	8 (12.9)	11 (20.4)	
Discrepancy in NRSIs type were considered	3 (15.8)	1 (12.5)	2 (18.2)	
The effect estimates between RCTs and pooled RCTs and NRSIs were qualitatively agree[b,c]				0.185
Yes (qualitatively agree)	128 (64.0)	69 (69.0)	59 (59.0)	
Increased, significantly	15 (11.7)	7 (10.1)	8 (13.6)	
Decreased, significantly	35 (27.3)	21 (30.4)	14 (23.7)	
Not significantly	78 (60.9)	41 (59.4)	37 (62.7)	
No (qualitatively disagree)	72 (36.0)	31 (31.0)	41 (41.0)	
Change in statistically significant, with concordant direction of effect	64 (88.9)	28 (90.3)	36 (87.8)	
Change in statistically significant, with opposite direction of effect	8 (11.1)	3 (9.7)	5 (12.2)	
Positive results were reported in conclusion when RCT and NRSI results were inconsistent[b,d]	38 (92.7)	14 (100.0)	24 (88.9)	0.507
Concordant direction of effect	26 (68.4)	11 (78.6)	15 (62.5)	
Opposite direction of effect	12 (31.6)	3 (21.4)	9 (37.5)	
The impact of the inclusion of NRSIs discussed	101 (45.9)	50 (45.5)	51 (46.4)	1

RoR Ratio of odds ratio

[a] The results will be said to qualitatively agree if RCTs and NRSIs identify the same direction of effects, namely a statistically significant increase, a statistically significant decrease, or no statistically significant difference. Statistically significant base on $P<0.05$, not statistically significant based on $P\geq0.05$

[b] Three meta-analyses were removed, which did not provide original data for each included study

[c] The qualitatively agreement of estimates between RCTs and pooled RCTs and NRSIs was assessed in studies that combined RCTs and NRSIs in the same meta-analysis with or without subgroups, namely, 203 articles. The results will be said to qualitatively agree if RCTs and pooled RCTs and NRSIs identify the same direction of effects, namely a statistically significant increase, a statistically significant decrease, or no statistically significant difference

[d] Whether the authors tended to report positive results was assessed in studies that combined RCTs and NRSIs in the same meta-analysis with subgroup analysis or analyzed separately with inconsistent results, namely, 41 articles

disagree", and almost half of the studies found "important difference" between the different evidence. The integration of NRSIs changed the qualitative direction of the estimates from RCTs in 36.0% of the studies. The evidence syntheses of RCTs and NRSIs also did not address inappropriate reporting of results, with 38 of 41 studies (92.7%) were more likely to report positive results when the results of RCTs and NRSIs were inconsistent. This practice may be due to a lack of practical guidance on when and how to integrate evidence from RCTs and NRSIs, which strongly influences the validity of the evidence synthesis [35]. Although studies published in Core journals are generally considered to be better designed and conducted, we found no significant differences between core and non-core journals in almost all aspects, highlighting methodological areas for improvement in the integration of RCTs and NRSIs.

Comparison with other studies

Several previous methodological surveys have examined various issues in the evidence synthesis of RCTs and NRSIs [17–19, 31]. Regarding the reasons for including NRSIs, one review identified 202 Cochrane reviews of interventions and found that 56% of the reports did not specify the reasons for including NRSIs [31]. Two meta-epidemiological reviews compared the estimates from RCTs only with those from pooled RCTs and observational studies, and found a substantial change (i.e., 27–71%) in conclusions after including observational studies in evidence pairs [17, 36].

Compared with previous studies, our study included a wide range of systematic reviews and provided a complete picture of the key considerations in evidence synthesis. First, we thoroughly explored different clinical scenarios for the practical application of NRSIs in systematic reviews and categorized them into four classifications with multiple examples from the included studies. Second, we systematically identified the methodological issues in studies included in NRSI and RCT reporting from the perspective of research design, conduct, analysis, and interpretation of results. We also emphasize the importance of interpreting the overall results after incorporating data from NRSIs into a meta-analysis of RCTs through heterogeneity analysis, sensitivity analysis, publication bias analysis, etc. Our study provides a new perspective and will help researchers as a reference and improve the generation of best evidence.

Implications for the broader research field

Clarifying clinical scenarios and methodological issues, and assessing agreement between RCTs and NRSIs, have important implications for the design, conduct, analysis, and interpretation of evidence syntheses of RCTs and NRSIs. Our meta-epidemiological review found that discrepancies in effect estimates between RCTs and NRSIs are often ignored in empirical studies, with almost a quarter showing inconsistencies in the statistical significance or direction of effects. Our study also provides valuable insights into evidence syntheses that include RCTs and NRSIs in public health, occupational health, environmental health, or toxicology, where the definitions of RCTs and NRSIs are consistent, although they have different purposes and use different tools. In particular, the methodological weaknesses identified in our review also apply to these areas.

We proposed four recommendations to guide when and how to synthesize RCTs and NRSIs in the following steps (Fig. 4):

- Step 1. Decide at the outset of a systematic review whether to include NRSIs. We encourage authors to specify the questions of interest, be transparent about the rationale for including NRSIs, and discuss the potential implications of this action.
- Step 2. When deciding to include NRSIs in a systematic review, first assess the compatibility of the different types of evidence [27, 32, 35, 37]. If they are not compatible, caution should be warranted when combining RCTs and NRSIs in the same meta-analysis. Adequately addressing the bias of NRSIs is necessary, such as adjusting for important confounding of NRSIs and excluding NRSIs with high risk of bias from the analysis [34].
- Step 3. Evidence from RCTs and NRSIs may be presented as narrative syntheses, quantitative analyses, or a combination. If it is possible to combine RCTs and NRSIs in the same meta-analysis, we recommend that advanced statistical approaches that allow for bias corrections are the preferred method for quantitative analysis, rather than traditional meta-analysis [16, 33, 38–40].
- Step 4. We encouraged authors to explore the sources of heterogeneity, highlight differences between included RCTs and NRSIs, and discuss their impact on the direction and magnitude of pooled estimates.

Strengths and limitations

This article is the first, to our knowledge, to systematically explore when and how to integrate evidence from RCTs and NRSIs, and to identify methodological gaps in key considerations in the process of evidence synthesis. We utilized rigorous systematic survey methods, including explicit eligibility criteria, standardized screening procedures, and pilot-tested forms for study screening and data extraction. We did not restrict the specific types of NRSI and randomly selected systematic reviews from both Core and non-Core clinical journals, thus

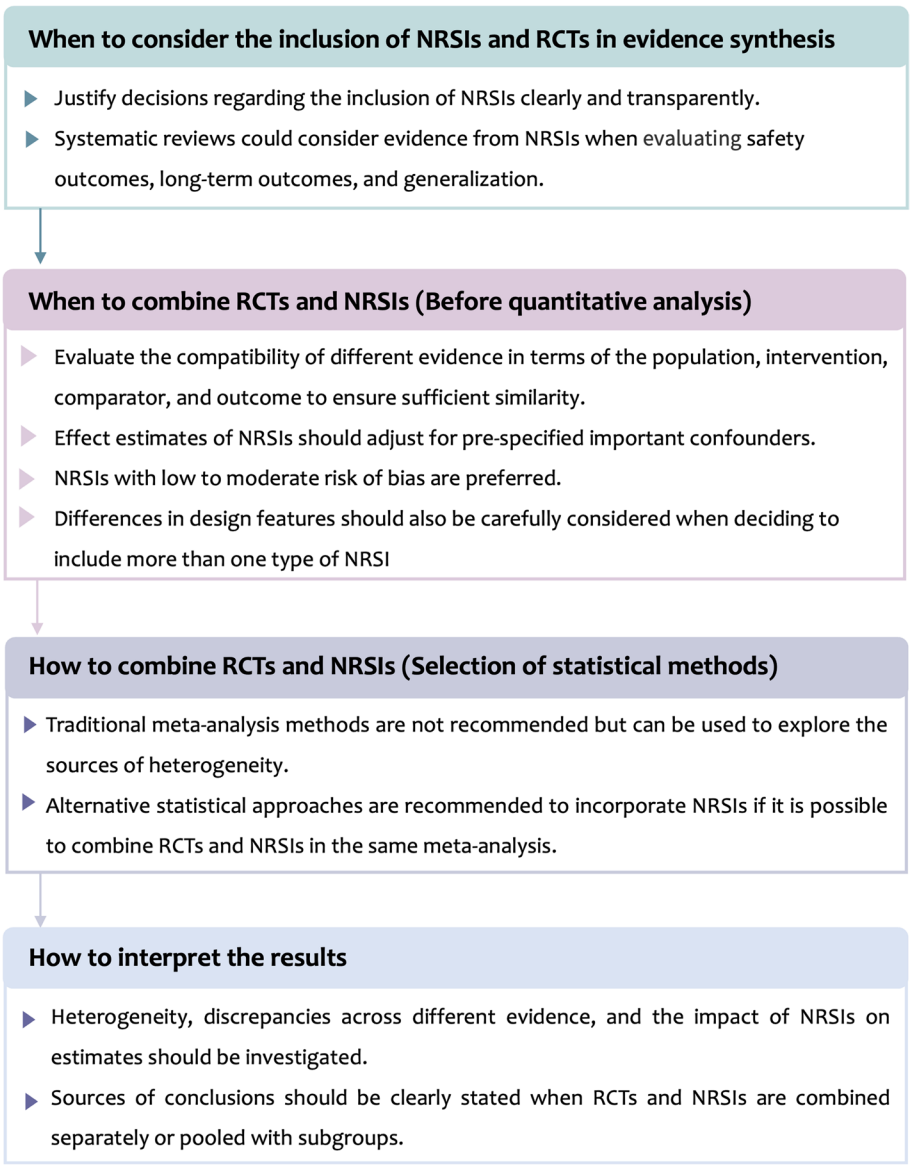

Fig. 4 Framework of recommendations for incorporating NRSIs into a meta-analysis of RCTs

enhancing the generalizability of our findings. Second, our study was based on a wide range of methods, including the rationale for including NRSIs, planning and identifying NRSIs for inclusion, assessing the risk of bias and grading the strength of evidence, considerations before including data from NRSIs, methods of conducting meta-analyses, and discussion of the conclusions.

However, some limitations are still present. First, we only included pairwise meta-analyses, and the outcomes of these meta-analyses are binary. The findings from our study may not be generally applicable to other types of reviews. Second, we only accepted information and data as reported by the authors of the included systematic reviews or meta-analyses, which makes the results vulnerable to underreporting or selective reporting. Third, although risk of bias may be an important driver of important differences between RCTs and NRSIs, we were unable to assess the effect of risk of bias on differences due to the high heterogeneity of the tools used by systematic review authors. Fourth, we did not take into account systematic reviews that did not specify any type of NRSI included. Although guidelines emphasize the importance of clearly reporting study design characteristics[23], almost 50% of studies excluded from the literature screening process did not report this, which is an important methodological issue for current research. Fifth, we did not assess factors

influencing disagreement between RCTs and NRSIs, such as lack of statistical power, clinically meaningful differences. Although we restricted the evidence from RCTs and NRSIs to the same outcome, there may be differences in PI/ECO (Population, Intervention/ Exposure, Comparison, Outcome) characteristics.

Conclusions

Systematic reviews typically included NRSIs in the context of assessing adverse or long-term outcomes, and the applicability of RCT results to broader populations. The inclusion of NRSIs in a meta-analysis of RCTs has a significant impact on estimates, with more than a third of studies changing their quantitative direction. Our findings highlight areas for improvement in the synthesis of evidence from RCTs and NRSIs, in particular that discrepancies between RCTs and NRSIs on the magnitude and direction of effects are significant but often ignored. We recommend careful consideration of when and how to integrate evidence from RCTs and NRSIs.

Abbreviations

ACR	Assumed control risk
AHRQ	Agency for Healthcare Research and Quality
GRADE	Grading of Recommendations, Assessment, Development and Evaluation
HR	Hazard ratio
IQR	Interquartile range
IV	Inverse variance
MH	Mantel-Haenszel
NOS	Newcastle-Ottawa Scale
NRSI	Non-randomized studies of intervention
OR	Odds ratio
PI/ECO	Population, Intervention/Exposure, Comparison, Outcome
RCT	Randomized controlled trial
RD	Risk difference
RoB	Risk of bias
RR	Risk ratio
RoR	Ratio of odds ratio
WHO	World Health Organization

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12916-024-03778-1.

Additional file 1: Supplementary Methods and Supplementary Results. Supplementary Methods: Search Strategy. Supplementary Results: Lists of 220 systematic reviews included.

Additional file 2: Tables S1-S5. Table S1. Scenario classification for the inclusion of NRSIs and real-life examples. Table S2. Reclassification of justification for including NRSIs. Table S3. Description of Intervention/Exposure (I/E) and Outcome (O), risk of bias assessment tools, and evidence rating system for all included studies. Table S4. Consistency of effect estimates and direction between RCTs and NRSIs. Table S5. Impact of combining RCTs and NRSIs on estimates.

Additional file 3: Fig. S1-S4. Fig. S1. Literature screening process. The list of included studies is available in Additional file 1: Supplementary Results. Fig. S2. Distribution of the description of NRSI types in included studies. Fig. S3. Distribution of justification for including NRSIs with different interventions and different study design characteristics. Fig. S4. Proportion of selective reporting of conclusions.

Acknowledgements

Not applicable.

Authors' contributions

XS, LL, MY, and FM conceived and designed the study. FM and YW collected the data. FM, YW, JH, MY, YM, and GL screened the literature and extracted the data. FM and MY analyzed the data and drafted the manuscript. XS, LL, MY, KZ, and GL critically revised the manuscript. All authors read and approved the final manuscript.

Funding

This study was supported by National Natural Science Foundation of China (Grant No. 72204173, 82274368, and 71904134), National Science Fund for Distinguished Young Scholars (Grant No. 82225049), special fund for traditional Chinese medicine of Sichuan Provincial Administration of Traditional Chinese Medicine (Grant No. 2024zd023), and 1.3.5 project for disciplines of excellence, West China Hospital, Sichuan University (Grant No. ZYGD23004).

Data availability

Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Institute of Integrated Traditional Chinese and Western Medicine, Chinese Evidence-Based Medicine Center, Cochrane China and MAGIC China Center, West China Hospital, Sichuan University, Chengdu, China. [2]NMPA Key Laboratory for Real World Data Research and Evaluation in Hainan, Chengdu, China. [3]Sichuan Center of Technology Innovation for Real World Data, Chengdu, China. [4]Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Ontario, Canada. [5]Center for Clinical Epidemiology and Methodology, Guangdong Second Provincial General Hospital, Guangzhou, China. [6]Biostatistics Unit, Research Institute at St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada. [7]Department of Epidemiology and Biostatistics, School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China.

Received: 13 June 2024 Accepted: 14 November 2024
Published online: 02 December 2024

References

1. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet. 2017;390(10092):415–23.
2. Watson C. How to modernize medical evidence for the misinformation era. Nat Med. 2023;29(10):2383–6.
3. Yao M, Wang Y, Mei F, Zou K, Li L, Sun X. Methods for the Inclusion of Real-World Evidence in a Rare Events Meta-Analysis of Randomized Controlled Trials. J Clin Med. 2023;12(4):1690.
4. 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. J Clin Epidemiol. 2022;142:200–8.
5. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med. 2016;375(23):2293–7.
6. Cuello-Garcia CA, Morgan RL, Brozek J, Santesso N, Verbeek J, Thayer K, et al. A scoping review and survey provides the rationale, perceptions, and preferences for the integration of randomized and nonrandomized

studies in evidence syntheses and GRADE assessments. J Clin Epidemiol. 2018;98:33–40.
7. Norris SL, Atkins D, Bruening W, Fox S, Johnson E, Kane R, et al. Observational studies in systematic reviews of comparative effectiveness: AHRQ and the Effective Health Care Program. J Clin Epidemiol. 2011;64(11):1178–86.
8. Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, et al. Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study. PLoS Med. 2016;13(5): e1002028.
9. Munn Z, Barker TH, Aromataris E, Klugar M, Sears K. Including nonrandomized studies of interventions in systematic reviews: principles and practicalities. J Clin Epidemiol. 2022;152:314–5.
10. Shrier I, Boivin JF, Steele RJ, Platt RW, Furlan A, Kakuma R, et al. Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. Am J Epidemiol. 2007;166(10):1203–9.
11. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. PLoS Med. 2011;8(5):e1001026.
12. Saldanha IJ, Adam GP, Bañez LL, Bass EB, Berliner E, Devine B, et al. Inclusion of nonrandomized studies of interventions in systematic reviews of interventions: updated guidance from the Agency for Health Care Research and Quality Effective Health Care program. J Clin Epidemiol. 2022;152:300–6.
13. Wang SV, Schneeweiss S, Initiative R-D. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses Results of 32 Clinical Trials. JAMA. 2023;329(16):1376–85.
14. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev. 2014;2014(4):Mr000034.
15. Verde PE, Ohmann C. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. Res Synth Methods. 2015;6(1):45–62.
16. Yao M, Wang Y, Ren Y, Jia Y, Zou K, Li L, et al. Comparison of statistical methods for integrating real-world evidence in a rare events meta-analysis of randomized controlled trials. Res Synth Methods. 2023;14(5):689–706.
17. Bun RS, Scheer J, Guillo S, Tubach F, Dechartres A. Meta-analyses frequently pooled different study types together: a meta-epidemiological study. J Clin Epidemiol. 2020;118:18–28.
18. Cheurfa C, Tsokani S, Kontouli KM, Boutron I, Chaimani A. Empirical evaluation of the methods used in systematic reviews including observational studies and randomized trials. J Clin Epidemiol. 2023;158:44–52.
19. Faber T, Ravaud P, Riveros C, Perrodeau E, Dechartres A. Meta-analyses including non-randomized studies of therapeutic interventions: a methodological review. BMC Med Res Methodol. 2016;16:35.
20. Yao M, Wang Y, Busse JW, Briel M, Mei F, Li G, et al. Evaluating the impact of including non-randomised studies of interventions in meta-analysis of randomised controlled trials: a protocol for a meta-epidemiological study. BMJ Open. 2023;13(7): e073232.
21. U.S. National Library of Medicine. Abridged index Medicus (AIM or "core clinical") Journal titles. http://www.nlm.nih.gov/bsd/aim.html. Accessed 22 Oct 2023.
22. Saldanha IJ, Skelly AC, Ley KV, Wang Z, Berliner E, Bass EB, et al. Inclusion of Nonrandomized Studies of Interventions in Systematic Reviews of Intervention Effectiveness: An Update. Rockville (MD): Agency for Healthcare Research and Quality (US). 2022.
23. Reeves BC DJ, Higgins JPT, et al. Chapter 24: Including non-randomized studies on intervention effects. In: Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editor. Cochrane Handbook for Systematic Reviews of Interventions version 64 (updated August 2023). Cochrane. 2023.
24. Zhou Q, He H, Li Q, Zhao J, Wang L, Luo Z, et al. Patient and public involvement in systematic reviews: frequency, determinants, stages, barriers, and dissemination. J Clin Epidemiol. 2024;170: 111356.
25. Grant RL. Converting an odds ratio to a range of plausible relative risks for better communication of research findings. BMJ. 2014;348: f7450.
26. 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 Med. 2021;19(1):307.
27. 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 Med. 2022;20(1):174.
28. Dahabreh IJ, Sheldrick RC, Paulus JK, 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. Eur Heart J. 2012;33(15):1893–901.
29. Zorzela L, Golder S, Liu YL, Pilkington K, Hartling L, Joffe A, et al. Quality of reporting in systematic reviews of adverse events: systematic review. BMJ. 2014;348:f7668.
30. Sterne JAC, Savovic J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:14898.
31. Ijaz S, Verbeek JH, Mischke C, Ruotsalainen J. Inclusion of nonrandomized studies in Cochrane systematic reviews was found to be in need of improvement. J Clin Epidemiol. 2014;67(6):645–53.
32. Higgins JPT, Ramsay C, Reeves BC, Deeks JJ, Shea B, Valentine JC, et al. Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. Res Synth Methods. 2013;4(3):288.
33. Moran JL, Linden A. Problematic meta-analyses: Bayesian and frequentist perspectives on combining randomized controlled trials and non-randomized studies. BMC Med Res Methodol. 2024;24(1):99.
34. Valentine JC, Thompson SG. Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions. Res Synth Methods. 2013;4(1):26–35.
35. 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.
36. Bröckelmann N, Stadelmaier J, Harms L, Kubiak C, Beyerbach J, Wolkewitz M, et al. An empirical evaluation of the impact scenario of pooling bodies of evidence from randomized controlled trials and cohort studies in medical research. BMC Med. 2022;20(1):355.
37. 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 Technol Assess. 2000;4(34):1–154.
38. Efthimiou O, Mavridis D, Debray TP, Samara M, Belger M, Siontis GC, et al. Combining randomized and non-randomized evidence in network meta-analysis. Stat Med. 2017;36(8):1210–26.
39. Zhou Y, Yao M, Mei F, Ma Y, Huan J, Zou K, et al. Integrating randomized controlled trials and non-randomized studies of interventions to assess the effect of rare events: a Bayesian re-analysis of two meta-analyses. BMC Med Res Methodol. 2024;24(1):219.
40. Yao M, Mei F, Zou K, Li L, Sun X. A Bayesian bias-adjusted random-effects model for synthesizing evidence from randomized controlled trials and nonrandomized studies of interventions. J Evid Based Med. 2024;17(3):550–8.

Publisher's Note