BMJ Open

Tools for assessing quality of studies investigating health interventions using real-world data: a literature review and content analysis

Li Jiu ,[1] Michiel Hartog,[1] Junfeng Wang,[1] Rick A Vreman,[1] Olaf H Klungel,[1] Aukje K Mantel-Teeuwisse ,[1] Wim G Goettsch[1,2]

To cite: Jiu L, Hartog M, Wang J, *et al*. Tools for assessing quality of studies investigating health interventions using real-world data: a literature review and content analysis. *BMJ Open* 2024;**14**:e075173. doi:10.1136/bmjopen-2023-075173

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2023-075173).

Received 28 April 2023
Accepted 22 January 2024

[1]Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands
[2]National Health Care Institute, Diemen, Netherlands

Correspondence to
Dr Wim G Goettsch; w.g.goettsch@uu.nl, Dr Junfeng Wang; j.wang5@uu.nl and Dr Junfeng Wang; j.wang5@uu.nl

ABSTRACT

Objectives We aimed to identify existing appraisal tools for non-randomised studies of interventions (NRSIs) and to compare the criteria that the tools provide at the quality-item level.

Design Literature review through three approaches: systematic search of journal articles, snowballing search of reviews on appraisal tools and grey literature search on websites of health technology assessment (HTA) agencies.

Data sources Systematic search: Medline; Snowballing: starting from three articles (D'Andrea *et al*, Quigley *et al* and Faria *et al*); Grey literature: websites of European HTA agencies listed by the International Network of Agencies for Health Technology Assessment. Appraisal tools were searched through April 2022.

Eligibility criteria for selecting studies We included a tool, if it addressed quality concerns of NRSIs and was published in English (unless from grey literature). A tool was excluded, if it was only for diagnostic, prognostic, qualitative or secondary studies.

Data extraction and synthesis Two independent researchers searched, screened and reviewed all included studies and tools, summarised quality items and scored whether and to what extent a quality item was described by a tool, for either methodological quality or reporting.

Results Forty-nine tools met inclusion criteria and were included for the content analysis. Concerns regarding the quality of NRSI were categorised into 4 domains and 26 items. The Research Triangle Institute Item Bank (RTI Item Bank) and STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) were the most comprehensive tools for methodological quality and reporting, respectively, as they addressed (n=20; 17) and sufficiently described (n=18; 13) the highest number of items. However, none of the tools covered all items.

Conclusion Most of the tools have their own strengths, but none of them could address all quality concerns relevant to NRSIs. Even the most comprehensive tools can be complemented by several items. We suggest decision-makers, researchers and tool developers consider the quality-item level heterogeneity, when selecting a tool or identifying a research gap.

OSF registration number OSF registration DOI (https://doi.org/10.17605/OSF.IO/KCSGX).

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This literature review identified 49 appraisal tools for non-randomised studies of interventions, through both the systematic approach (ie, database search) and the non-systematic approaches (ie, snowballing and grey literature search).
⇒ Our study compared sufficient descriptions of appraisal tools at quality-item levels, for either methodological quality or reporting.
⇒ We only searched health technology assessment agencies for grey literature, so some tools only mentioned by clinical guideline or regulatory organisations might have been overlooked.
⇒ Usefulness of categorising a quality item as 'sufficient' or 'brief' for each tool, based on whether an explanation was provided for the criteria, has not been tested by previous studies.

INTRODUCTION

Real-world data (RWD) generally refer to data collected during routine clinical practice, but their definitions could vary in settings.[1] According to Makady *et al*, one of the RWD definitions is data collected without interference with treatment assignment.[1] RWD that fit this definition are normally analysed in non-randomised studies of interventions (NRSIs), which estimate effectiveness of a health intervention without randomising intervention groups.[2 3]

NRSIs provide evidence on clinical and cost-effectiveness of health interventions for decision-making, in clinical and health technology assessment (HTA) settings.[4–9] For example, NRSIs could inform clinicians on what diagnosis or treatment strategies to adopt.[4 5] Also, with NRSIs, HTA agencies could gain more certainty on validity of evidence from randomised controlled trials (RCTs), when deciding on which health intervention to reimburse and on which pricing strategy to adopt.[6 7] Also, HTA stakeholders

could exploit NRSIs to evaluate highly innovative or complex interventions, for which RCTs may be considered infeasible or unethical.[8 9] Generally speaking, NRSIs have become increasingly useful, as they complement and sometimes replace RCTs, when RCTs are scarce or even infeasible to conduct.[2 10]

However, the usefulness of NRSIs is often questioned due to quality concerns, in terms of risk of bias (RoB) and reporting. According to the Cochrane Handbook, NRSIs have higher RoB than RCTs and are vulnerable to various types of bias, such as confounding, selection and information bias.[11] Also, the Professional Society for Health Economics and Outcomes Research (ISPOR) published a report in 2020, which stated that insufficient reporting on how an NRSI was generated was a major barrier for decision-makers to adopt NRSIs.[12]

To address NRSI's quality concerns and to build decision-makers' confidence, NRSIs need to be rigorously appraised, and this rationalises the development and use of appraisal tools. According to systematic reviews of appraisal tools for NRSIs, tens of tools have been developed in the past five decades.[13–15] The growing number of tools has then brought a new challenge to users: how to select the best tool. To address this challenge, previous reviews have summarised quality items (ie, a group of criteria or signalling questions for methodological quality or reporting) and compared whether existing tools addressed these items.[13–15] Some example items include 'measurement of outcomes', 'loss to follow-up bias', 'inclusion and exclusion criteria of target population', 'sampling strategies to correct selection bias', etc.[13] In addition, these reviews provided some general recommendations on tool selection, such as referring to multiple tools for quality appraisal.[14] However, information is still lacking on to what extent the tools address each quality item and the heterogeneity of tools at the quality-item level. To take outcome measurement as an example, the Academy of Nutrition and Dietetics Quality Criteria (ANDQ) checklist mentions that outcomes should be measured with 'standard, valid and reliable data collection instruments, tests and procedures' and 'at an appropriate level of precision'.[16] In contrast, the Good ReseArch for Comparative Effectiveness (GRACE) checklist considers the 'valid and reliable' measurement as 'objective rather than subject to clinical judgement'[17]; while the Risk Of Bias In Non-randomised Studies—of Interventions (ROBINS-I) checklist interprets the 'standard' way as 'comparable across study groups' and 'valid and reliable' as low detection bias without 'systematic errors' in outcome measurement.[18] In summary, the heterogeneity in level of detail with which a tool addresses a quality item and the heterogeneity in content and format of signalling questions can pose a challenge when tools are selected, or even merged.

Hence, our study aimed to summarise and compare signalling questions or criteria in the tools provided at the quality-item level, through a content analysis. This research was performed as part of the HTx project.[19] The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825162.

METHOD

Protocol

To ensure credibility of the review and the content analysis, we registered a study protocol in the OSF registry (registration DOI: https://doi.org/10.17605/OSF.IO/KCSGX) on 30 June 2022. The OSF registry is an online repository that accepts registration of all types of research projects, including reviews and content analyses.[20]

Patient and public involvement

Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Scope

In our study, appraisal tools refer to tools, guidelines, instruments or standards that provide guidance on how to report or assess any quality concern of NRSIs. NRSIs, according to the Cochrane Handbook, refer to any quantitative study estimating the effectiveness of an intervention without randomisation to allocate patients to intervention groups.[2] According to Makady *et al*, data collected in such NRSIs belong to the second category of RWD, that is, those collected without interference with treatment assignment, patient monitoring or follow-up, or selection of study population.[1]

Search strategy

To identify appraisal tools for NRSIs from various potential sources, we adopted three approaches. A diagram illustrating how the three approaches complemented each other is shown in online supplemental appendix 1.

Database search

In the first approach, we conducted a systematic review to identify articles on appraisal tools, through a database search using Medline. Since D'Andrea *et al* have already conducted a systematic review to identify appraisal tools for all types of non-randomised studies published before November 2019,[13] we updated their review by searching for articles published between November 2019 and April 2022, with their strings.

Snowballing

In the second approach, we searched for published reviews on appraisal tools for NRSIs. To identify all published reviews, we adopted a snowballing approach described by Wohlin.[21] Snowballing refers to using the citations of articles to identify additional articles, and it is considered a good extension of a database search.[21] To implement the snowballing approach, three researchers (LJ, MH and JW) first conducted a pilot search of articles using Google Scholar, reviewed full-text, judged eligibility through a group discussion, then identified

three reviews (ie, those by D'Andrea *et al*,[13] Quigley *et al*[14] and Faria *et al*[15]). Next, the three reviews were used as a starting set and were uploaded to the website Connected Papers, which provides an online tool for snowballing.[22] With each uploaded review, Connected Papers analysed approximately 50 000 articles and finally returned 40 articles with the highest level of similarity, based on factors such as overlapping citations. After judging eligibility of the returned articles, eligible articles were uploaded to the website Connected Papers for a second round of snowballing.

Grey literature

In the third approach, we searched for grey literature on the websites of European HTA agencies. Our rationale was that some appraisal tools may exist in the format of grey literature, such as agency reports and technical support documents. The list of European HTA agencies was derived from the International Network of Agencies for Health Technology Assessment.[23] On each agency website, two researchers (MH and LJ) independently searched for grey literature with four concepts, respectively, 'quality', 'RoB', 'appraisal' and 'methodology'. For each concept, only the first 10 hits sorted by relevance, if optional, were included (ie, a maximum of 40 hits for each website).

Eligibility criteria for articles and grey literature to identify relevant tools

An article or grey literature document was included if it described one or more appraisal tools. It was excluded if it only described tools for RCTs or only described tools for diagnostic, prognostic, qualitative or secondary studies (eg, systematic reviews and cost-effectiveness analyses). We only included articles identified through the database search and snowballing if published in English, while included grey literature could be published in all languages, as many HTA agencies tend to only use languages of their nations. Relevant documents obtained through this approach were translated using Google Translate.

The process of identifying studies and appraisal tools

Two researchers (MH and LJ) independently scanned all titles and abstract of the identified hits, then reviewed the full-text with Rayyan[24] and Excel. After identifying the eligible studies, one researcher (MH) extracted the name of the tools and downloaded them by tracking study citations. A pilot search with Google was conducted to ensure we downloaded the most up-to-date version. Next, two researchers (MH and LJ) independently reviewed full-text and judged eligibility of the tools. An appraisal tool was included if it (1) was designed for non-randomised studies, (2) was used for assessing either methodological quality or reporting and (3) was developed or updated after 2002. A tool was excluded if it was designed for non-randomised studies of exposures which were not controlled by investigators (eg, diets). All discrepancies were solved through discussion among the three researchers (MH, LJ and JW).

One researcher (MH) extracted tool characteristics using a prespecified Excel form. The data items included publication year, tool format (eg, checklist or rating scale), targeted study design (eg, all NRSIs, cohort studies, etc), target interventions (eg, all or surgical interventions), originality (ie, whether a tool was developed based on an existing tool) and scope. The scope referred to whether the tools were designed for assessing methodological quality (eg, RoB and external validity) and/or for ensuring adequate reporting of research details that could be used for assessing methodological quality.[25]

For the content analysis, we adopted both deductive and inductive coding techniques.[26] First, we derived a list of candidate quality items from the three reviews, the starting set for the snowballing.[13–15] Then, in a pilot coding process, we reviewed all identified appraisal tools and judged whether a candidate quality item was described. After the pilot coding, we summarised signalling questions or criteria that were not covered by the candidate items and coded them as new items. After updating the list of candidate items, three researchers (JW, LJ and MH) finalised the items in four group meetings. During the meetings, we merged items with overlapping content, split items containing too much content and renamed items so they could be self-explanatory.

To score whether and to what extent a quality item was described by a tool, we again reviewed all identified tools. If an item was described by a tool in one or several signalling questions, we judged whether the question(s) was related to methodological quality, reporting or both, independently of what original studies claimed to be. Additionally, we judged whether an item was described sufficiently or briefly. A description was scored as 'brief', if the corresponding signalling question(s) did not explain how to improve or assess methodological quality or specify elements needed for reporting. For example, 'outcomes should be measured appropriately' or 'outcome measurement should be adequately described' are 'brief' descriptions, if no additional explanations were provided. The scoring process was independently conducted by two researchers (LJ and MH) using NVivo V.12, and all discrepancies were solved through discussion between the two.

RESULT

Tool selection

As shown in figure 1, we identified 1738 articles after removing duplicates and excluded 1645 articles after subsequently reviewing titles, abstracts and full-text. From the remaining 27 eligible studies, we identified 417 appraisal tools. After removing duplicates and reviewing full-texts, we included 49 tools which met our criteria. References of the included studies and appraisal tools are shown in online supplemental appendix 2 and 3, respectively.

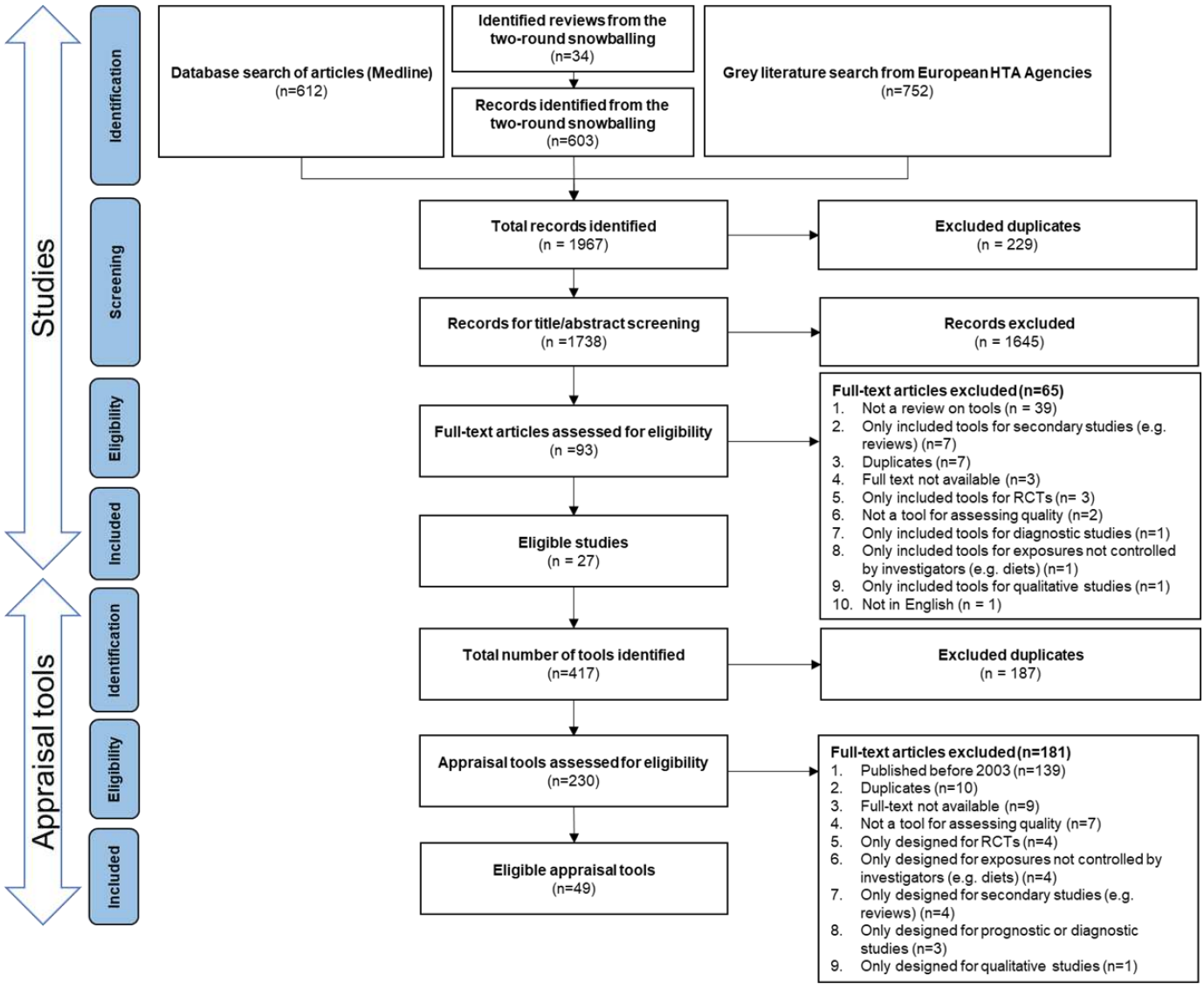


Figure 1 Flow chart for the inclusion and exclusion of appraisal tools for non-randomised studies of interventions

Characteristics of appraisal tools

As shown in table 1, 18 (37%) tools were published between 2002 and 2010, while 31 (63%) tools were published thereafter. Among these, 30 (61%), 6 (12%) and 5 (10%) tools were designed for addressing methodological quality, reporting and both, respectively, while 7 (14%) tools did not report intended use of the tools. About three quarters of the tools were designed for all types of NRSIs, while others were designed for one or several NRSI types, such as cohort (16%) and case–control studies (16%). Regarding sources, 44 (90%) tools were described in articles that developed a tool, in grey literature (eg, online checklist or report), or in both, while the other five tools were extended from existing tools, when researchers conducted systematic reviews on non-randomised studies. Finally, 9 (18%) tools were designed for specific interventions or diseases while all other tools were generic in nature.

Quality domains and items

We identified 44 criteria to describe study quality from three previous reviews.[13–15] After merging criteria with similar content (eg, 'Follow-up' and 'Loss to follow-up') and incorporating items into those with wider meanings (eg, 'Loss to follow-up bias' into 'Loss to follow-up'), we obtained a list of 18 items. After the pilot coding, we summarised criteria of appraisal tools not covered by the 18 items into another 8 items. According to the general order of conducting an NRSI (eg, study design and data analysis, etc.), these 26 items were categorised into four domains: Study design, Data quality, Data analysis and Results presentation. As shown in figure 2 and table 2, all domains and most items were addressed by existing tools, but for each item, the number of tools with sufficient descriptions was relatively small. For three items in methodology and nine items in reporting, less than five tools addressed them, and none of the tools sufficiently described them.

Figure 2 illustrates whether and to what extent the identified tools addressed quality items in terms of methodological quality or reporting. The 26 columns represent the 26 quality items as shown in table 2. The

Table 1 Characteristics of the 49 included appraisal tools for non-randomised studies of interventions

No	Appraisal tools	Abbreviation	Publication year	Scope	Study design	Publication format	What intervention?
1	Critical Appraisal Skills Programme Tool	CASP	2022	M	Cohort, case–control	GL, TD article	NA
2	Joanna Briggs Institute's Critical Appraisal Tools	JBI	2020	M	Cohort, case–control	GL, TD article	NA
3	REal Life EVidence AssessmeNt Tool	RELEVANT	2019	M & R	All	TD article	NA
4	STrengthening the Reporting of OBservational studies in Epidemiology Checklists	STROBE	2019	R	All	GL, TD article	NA
5	Basque Office for Health Technology Assessment Tool	OSTEBA	2019	NR	All	GL	NA
6	NA	Kennedy *et al*[42]	2019	R	All	TD article	NA
7	Mixed Methods Appraisal Tool	MMAT	2018	M	All	TD article	NA
8	Critical Appraisal Tools of Specialist Unit for Review Evidence	SURE	2018	M	Cohort, cross-sectional, case–control, case-series	GL, TD article	NA
9	NA	Viswanathan *et al*[50]	2018	M	All	TD article	NA
10	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance Guide on Methodological Standards in Pharmacoepidemiology	ENCePP	2018	R	All	GL, TD article	NA
11	Risk Of Bias In Non-randomised Studies—of Interventions	ROBINS-I	2017	M & R	All	TD article	NA
12	NA	Faillie *et al*[29]	2017	M	All	GL	NA
13	Joint Task Force between the International Society for Pharmacoepidemiology and the International Society for Pharmacoeconomics and Outcomes Research	ISPE-ISPOR	2017	R	All	TD article	NA
14	Appraisal tool for Cross-Sectional Studies	AXIS	2016	M	Cross-sectional	TD article	NA
15	NA	Handu *et al*[32]	2016	M	All	TD article	NA

Continued

Table 1 Continued

No	Appraisal tools	Abbreviation	Publication year	Scope	Study design	Publication format	What intervention?
16	International Society of Pharmacoepidemiology Guidelines for Good Pharmacoepidemiology Practice	ISPE	2016	NR	All	GL, TD article	NA
17	REporting of studies Conducted using Observational Routinely-collected Data Checklist	RECORD	2015	M	All	TD article	NA
18	Comparative Effectiveness Research Collaborative Initiative Questionnaire	CER-CI	2014	M	All	TD article	NA
19	Good ReseArch for Comparative Effectiveness Checklist	GRACE	2014	NR	All	GL, TD article	NA
20	Scottish Intercollegiate Guidelines Network Checklists	SIGN	2014	M	Cohort, case–control	GL, TD article	NA
21	A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions	ACROBAT-NRSI	2014	M	All	GL	NA
22	Interventional Pain Management Techniques—Quality Appraisal of Reliability and Risk of Bias Assessment for Non-randomized Studies	IPM-QRBNR	2014	M	All	TD article	Interventional Pain Management
23	Quality Assessment Tool of National Heart, Lung, and Blood Institute	NIH	2013	NR	Cohort, cross-sectional, case–control, case-series	GL	NA
24	Guidelines manual of National Institute for Health and Care Excellence: Appendices D–E	NICE	2013	M	Cohort, case–control	GL	NA
25	CAse REport (CARE) Guidelines Checklist	CARE	2013	M	Case-report	TD article	NA
26	Institute of Health Economic Quality Appraisal Tool for Case-Series Studies	IHE	2012	M & R	Case-series	GL, TD article	NA

Continued

Table 1 Continued

No	Appraisal tools	Abbreviation	Publication year	Scope	Study design	Publication format	What intervention?
27	Agency for Healthcare Research and Quality Methodology Checklist	AHRQ	2012	M	All	GL	NA
28	Risk of Bias Assessment tool for Non-randomized Studies Tool	RoBANS	2011	M	All	GL, TD article	NA
29	Research Triangle Institute Item Bank	RTI Item Bank	2011	M	All	GL, TD article	NA
30	The Montreal Critical Appraisal Worksheet	Montreal	2011	R	All	GL	NA
31	Grades of Recommendation, Assessment, Development and Evaluation Guideline	GRADE	2011	M & R	All	GL, TD article	NA
32	NA	Blagojevic *et al*[51]	2010	M	Cohort, case–control	Modified for review	Knee osteoarthritis
33	Academy of Nutrition and Dietetics Quality Criteria Checklist (Primary Research)	ANDQ	2010	M	All	GL	Diabetes
34	NA	Bishop *et al* 2010[66]	2009	NR	All	Modified for review	Complementary medicine in paediatric cancer
35	Harm Critical Appraisal Worksheet	Harm	2009	M	All	GL	NA
36	Newcastle-Ottawa Scale	NOS	2009	NR	Cohort, case–control	GL, TD article	NA
37	NA	Pluye *et al* 2009[67]	2009	M	All	TD article	NA
38	NA	Young *et al* 2009[68]	2009	M	All	TD article	NA
39	NA	Atluri *et al* 2008[69]	2008	M	All	Modified for review	Thoracic facet joint interventions
40	NA	Tseng *et al*[38]	2008	M	All	Modified for review	Surgical interventions
41	NA	Heller *et al* 2008[70]	2008	R	All	TD article	NA
42	NA	Genaidy *et al*[35]	2007	M	All	TD article	NA
43	Graphic Appraisal Tool for Epidemiological Studies	GATE	2006	M	All	TD article	NA
44	NA	Weightman *et al* 2004[71]	2004	M	All	GL	NA

Continued

Table 1 Continued

No	Appraisal tools	Abbreviation	Publication year	Scope	Study design	Publication format	What intervention?
45	Transparent Reporting of Evaluations with Non-randomized Designs	TREND	2004	M	All	GL, TD article	NA
46	NA	Thomas *et al* 2004[72]	2004	M	All	Modified for review	Nursing interventions
47	NHS Wales Questions to Assist with the Critical Appraisal of a Cross-Sectional Study	NHS Wales	2004	M	Cross-sectional	GL	NA
48	Methodology Index for Non-randomized Studies	MINORS	2003	M & R	All	TD article	Surgical research
49	NA	Rangel *et al* 2003[73]	2003	NR	All	TD article	Paediatric surgery

Modified for review: an appraisal tool modified from existing appraisal tools during a review of primary studies in a certain disease field or for a certain health intervention.
GL, grey literature; M, methodological quality; NA, not applicable; NR, not reported; R, reporting; TD, tool development.

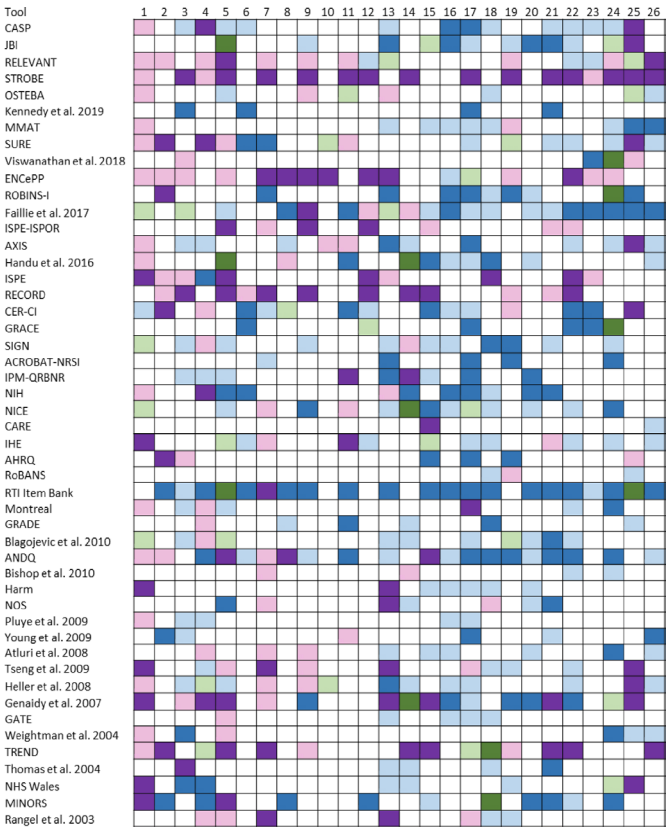


Figure 2 The extent to which the appraisal tools addressed quality items on methodological quality or reporting.

ranking of appraisal tools based on the number of items addressed or sufficiently described, either general or segmented by quality domains, is shown in online supplemental appendix 4–6. Regarding methodological quality, Research Triangle Institute Item Bank (RTI Item Bank)[27] addressed (n=20) and sufficiently described (n=18) the highest number of items. In addition, the tools that ranked both top 10, based on number of items addressed or sufficiently described, included Methodology Index for Non-randomized Studies (MINORS),[28], Faillie *et al*,[29] ROBINS-I,[18] ANDQ,[16] Comparative Effectiveness Research Collaborative Initiative Questionnaire (CER-CI)[30] and Joanna Briggs Institute's Critical Appraisal Tool (JBI).[31] These tools addressed at least 10 items and sufficiently described at least 5 items. In the study-design domain, RTI Item Bank[27] sufficiently described the most items (n=7), while in the Data quality domain, RTI Item Bank[27] and MINOR[28] ranked the top two, which sufficiently described at least 5 of the 10 items. In the Data analysis domain, only Faillie *et al*[29] and Handu *et al*[32] sufficiently described all the three included items. In the Results presentation domain, the relevant two items sufficiently described by Faillie *et al*[29] and Handu *et al*,[32] and ANDQ.[16] Regarding reporting, STrengthening the Reporting of OBservational studies in Epidemiology (STROBE)[33] addressed (n=17) and sufficiently described (n=14) the highest number of items. Also, the tools that ranked both top 10, based on the two criteria, included Transparent Reporting of Evaluations

Table 2 Overview of the 4 domains and 26 quality items, with numbers and proportions of appraisal tools that addressed or sufficiently described them

Domains	Items	Number (%) of appraisal tools that addressed or sufficiently described a quality item			
		Methodology (n=49)		Reporting (n=49)	
		Addressed	Sufficiently described	Addressed	Sufficiently described
1. Study design	1. Study objective	5 (10)	0	27 (55)	7 (14)
	2. Protocol	3 (6)	3 (6)	10 (20)	5 (10)
	3. Selection of study design	15 (31)	3 (6)	9 (18)	3 (6)
	4. Sample size/power calculation	12 (24)	5 (10)	15 (31)	4 (8)
	5. Eligibility criteria	16 (37)	5 (10)	20 (41)	12 (24)
	6. Intervention selection	9 (18)	6 (12)	2 (4)	0
	7. Intervention definition	4 (8)	2 (4)	19 (39)	7 (14)
	8. Outcome selection	6 (12)	3 (6)	4 (8)	2 (4)
	9. Outcome definition	6 (12)	3 (6)	11 (22)	5 (10)
	10. Ethical approval	2 (4)	0	4 (8)	1 (2)
	11. Conflict of interest	7 (14)	6 (12)	9 (18)	3 (6)
2. Data quality	12. Data source	5 (10)	1 (2)	7 (14)	5 (10)
	13. Patient recruitment	19 (39)	7 (14)	12 (24)	6 (12)
	14. Participation rate	11 (22)	4 (8)	10 (20)	7 (14)
	15. Baseline characteristics	14 (29)	5 (10)	9 (18)	5 (10)
	16. Intervention measurement	19 (39)	7 (14)	1 (2)	0
	17. Outcome measurement	28 (57)	12 (24)	7 (14)	2 (4)
	18. Blinding of outcome	21 (43)	7 (14)	4 (8)	3 (6)
	19. Missing data	12 (24)	6 (12)	11 (22)	1 (2)
	20. Length of follow-up	15 (31)	6 (12)	0	0
	21. Loss to follow-up	14 (29)	9 (18)	6 (12)	3 (6)
3. Data analysis	22. Description	19 (39)	6 (12)	6 (12)	5 (10)
	23. Sensitivity analysis	7 (14)	4 (8)	3 (6)	0
	24. Bias adjustment	22 (45)	12 (24)	9 (18)	4 (8)
4. Results presentation	25. Are all the results presented	10 (20)	4 (8)	15 (31)	11 (22)
	26. Reasonable conclusions from results	14 (29)	4 (8)	3 (6)	3 (6)

The judgement on whether criteria or signalling questions of an appraisal tool were relevant to methodological quality or reporting was made by authors, independently of what original studies claimed to be.

with Non-randomized Designs (TREND),[34] the tool by Genaidy *et al*,[35] REporting of studies Conducted using Observational Routinely-collected Data (RECORD),[36] European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP),[36] International Society of Pharmacoepidemiology (ISPE),[37] the tool by Tseng *et al*[38] and Joint Task Force between the International Society for Pharmacoepidemiology and the International Society for Pharmacoeconomics and Outcomes Research (ISPE-ISPOR).[39] These tools at least addressed and sufficiently described seven and three quality items, respectively. In all the four quality domains, STROBE[32] sufficiently described the (equally) most items, compared with other tools. Besides, in the Study design domain, ENCePP[36] and RECORD[40] sufficiently described at least 4 of the 11 items, while in the

Data quality domain, TREND[34] and Genaidy *et al*[35] sufficiently described at least 4 of the 10 items. In the Data analysis and Results presentation domain, STROBE was the only tool that sufficiently described two of the thee items, while 7 and 12 other tools sufficiently described only one item, respectively.

Methodological quality

Among the four domains, the Study design domain was the most ignored domain by appraisal tools, as only 4 of the 11 relevant items were described with sufficient details by more than four tools. More specifically, no tool described methodological quality on Ethical approval or Study objective with sufficient detail. For example, the guidelines manual of the National Institute for Health and Care Excellence (NICE) stated that: "The study addresses an appropriate and clearly focused question".[41] The tool did not explain the standard of appropriateness and clearness.

In addition, although one-third of tools discussed what a good study design was, only three tools defined the goodness.[42–44] For example, the NHS Wales Questions to Assist with the Critical Appraisal of a Cross-Sectional Study (NHS Wales) stated that the choice of study design should be appropriate to the research question and ensure the reliability of study results.[44] Outcome selection was also ignored by most tools, as only three tools (ie, RTI Item Bank,[27] MINORS[28] and the tool by Faillie *et al*[29]) sufficiently described them. Similarly, only RTI Item Bank,[27] the tool by Genaidy *et al*[35] and NICE[41] sufficiently described the item Outcome definition. For example, Genaidy *et al*[35] stated that a definition was clear only if 'definitions of all outcome variables were clearly described', and was partially clear if not all variables were clearly described, but 'sufficient information was provided for the reader to understand the intent'.[35] Other items that were rarely addressed or insufficiently described included Intervention definition and Data source. The respective tools with sufficient descriptions included SURE,[45] ROBINS-I,[18] MINORS,[28] CER-CI,[30] GRACE[17] and the tools described by Faillie *et al*.[29]

Reporting

The Data quality domain was ignored by most tools, as 4 of the 10 relevant items were sufficiently addressed by less than three tools. In particular, the item Intervention measurement and Length of follow-up were sufficiently addressed by none of the tools, JBI was the only tool stating that method of measuring interventions should be clearly reported,[31] while 19 tools addressing Intervention measurement only focused on methodological quality. Some other items that were rarely addressed or insufficiently addressed included Outcome blinding and Loss to follow-up. Regarding Outcome blinding, only three tools provided sufficient descriptions, that is, MINORS, TREND and ISPE.[28 34 37]. Similarly, only the tool by Genaidy *et al*,[35] TREND and STROBE sufficiently described Loss to follow-up.[32 35 36]

DISCUSSION

We conducted a review of appraisal tools for NRSIs and assessed whether and how sufficiently these tools addressed quality concerns, in terms of methodological quality or reporting, in 4 quality domains and across 26 items. Our study identified 49 tools and showed that the RTI Item Bank and STROBE were most comprehensive, with the highest number of items addressed and sufficiently described, respectively, on methodological quality and reporting. However, none of the tools addressed concerns in all items, not even briefly. The items least addressed for methodological quality included Outcome selection, Outcome definition and Ethical approval, and for reporting included Intervention selection, Intervention measurement and Length of follow-up.

To our knowledge, this is the first study that compared level of sufficient descriptions of appraisal tools at quality-item levels. Previous reviews also compared appraisal tools but from different perspectives. D'Andrea *et al* identified 44 tools evaluating the comparative safety and effectiveness of medications, and only assessed whether or not these tools addressed methodological quality in eight domains.[13] In another review, Ma *et al* elaborated for what types of study design a tool was suited.[46] For example, for cohort studies, they encouraged using five tools, while discouraged the use of another two. However, they did not clarify why some tools were more suitable than the others. Quigley *et al* identified 48 tools for appraising quality of systematic reviews of non-randomised studies, listed the five most commonly used tools and assessed whether they addressed the 12 quality domains, such as 'appropriate design' and 'appropriate statistical analysis'.[14] Although the tools were compared using different criteria, some results were consistent among all studies. For example, both D'Andrea *et al*[13] and our study found that intervention measurement, outcome measurement and confounding were frequently addressed by existing tools. Also, Ma *et al*[46] and Quigley *et al*[14] both recommended ROBINS-I, MINORS and JBI, and all these tools ranked top 10 for addressing and sufficiently describing methodological quality in our study. With detailed information on level of sufficient descriptions of appraisal tools at the quality-item level, we add value to previous reviews by listing quality concerns that such commonly recommended tools could not adequately address.

We also found some discrepancies in the tools identified or recommended. For example, of the 44 tools identified by D'Andrea *et al*,[13] 27 were published between 2003 and 2019; while in our study, 47 were identified as published between 2003 and 2019. This discrepancy could be explained by additional tools identified through other reviews, tools from grey literature and differences in eligibility criteria (eg, exclusion of non-pharmacological interventions or assessing only one or a few specific types of bias). Another discrepancy was that some tools that ranked top in our study were less recommended by previous reviews, such as RTI Item Bank[27] and the tool by Faillie *et al*[29] for methodological quality and by Genaidy *et*

al[35] for reporting. This might be explained by the novel criteria (ie, how sufficiently quality items were addressed) we used to evaluate these tools.

We discovered that, with information on how sufficiently a tool described a quality item, tool users might broaden their horizons on quality concerns of non-randomised studies to be considered. For example, if ROBINS-I[18] is used for assessing methodological quality, the quality concerns known to users will be RoB in eight domains (eg, confounding and selection bias). However, as shown in figure 2, quality concerns in 16 items (eg, Intervention selection and Outcome definition) may not be sufficiently described in ROBINS-I but in other tools, such as RTI Item Bank,[27] the NICE checklist[41] and the tool by NHS Wales.[44] Similarly, if users check the ENCePP[36] and ISPE tools,[37] in addition to STROBE, for reporting quality concerns, they may more comprehensively understand concerns on Ethical approval, Outcome definition, Study objective and Data source. Tool users who may benefit from such information are not only researchers who conduct non-randomised studies and decision-makers who assess study quality, but also tool developers who may identify a research gap.

While the needs of tool users may vary, they could all be somewhat satisfied by our research. For example, it is important for researchers to ensure sufficient reporting of the strengths and weaknesses of an NRSI, as such information will be ultimately used for determining the eligibility of their studies for a decision-making.[32 47] For HTA agencies, NRSIs can be used to extrapolate long-term drug effectiveness and to identify drug-related costs, and a deep and consistent understanding of how to assess NRSI quality among the agencies is important for promoting the use of RWD.[48] For regulators, a comprehensive understanding of how to evaluate NRSI quality may promote a structured pattern of using RWD to support drug regulation.[49] While researchers focus more on reporting, and decision-makers (eg, HTA agencies) have emphasis on methodological quality, we suggest all users pay attention to the linkage between methodology and reporting for each quality item, as illustrated in our research, as it could help understand the necessity of investigating each item.

Another finding of our research was that whether and to what extent a quality concern was addressed by a tool partly depended on the tool purpose. For example, the GRACE checklist was designed as a 'screening tool' to exclude studies that did not meet basic quality requirements,[17] and ROBINS-I focused on RoB, rather than all methodological quality issues, such as appropriateness of study objectives or statistical analyses for patient matching.[18] Some tools, such as JBI Cohort,[31] were specific to a type of study design. While they addressed less than half of quality items defined in our research, they were proven robust in many studies.[14] Additionally, for several quality items we found some heterogeneity in content of signalling questions or criteria among the tools with sufficient description. For example, to assess methodological quality of sensitivity analysis, CER-CI[30] stated that key assumptions or definitions of outcomes should be tested, while the tool by Viswanathan *et al*[50] emphasised the importance of reducing uncertainty in individual judgements. Given the heterogeneity of tools, we suggest users following a two-step approach when selecting a tool. First, users may narrow down the scope of tools based on their own needs, for example, excluding tools for a different study design. This step could be achieved by referring to synthesised results and recommendations from existing reviews.[13 14] Second, users could use the overview we provide (figure 2) to see which tool(s) could provide complementary insights the tool of their first choice is lacking.

Furthermore, we found that appraisal tools designed for specific interventions had potential to be transferred for general interventions. In our research, the tools described by Tseng *et al*[38] and Blagojevic *et al*[51] and ANDQ[16] were originally designed for a surgical intervention, knee osteoarthritis and for the field of diabetes, respectively. All these tools ranked top 15 in our study for addressing either methodological quality or reporting (online supplemental file appendix 4–6), and many of their criteria could be generalisable. For example, Tseng *et al*[38] stated that interventions could be adequately described with specifically referenced articles (online supplemental file appendix 7).[38] Though such tools could be transferred, they often used disease-or-intervention-specific concepts in their criteria, which might be adjusted before being applied more widely.

Moreover, we noticed that, some quality items were less frequently addressed, such as Study objective, Ethical approval or Sensitivity analysis, compared with other items. This might be explained by the fact that, some items were more related to a certain need of users than the others. For example, a tool addressing concerns on RoB may focus less on Study objective, which is relatively more difficult to be directly linked to a well-defined type of bias. Still, since these quality items are related to NRSI quality, and they are rarely sufficiently described, particular efforts investigating these quality items may be needed in future tool development. In contrast, while some quality items have been frequently addressed, such as Length of follow-up and Intervention measurement, they are not necessarily relevant to all types of user needs. For example, as shown in table 2 and online supplemental file appendix 7, 14 tools highlighted that the follow-up should be sufficiently substantial for detecting an association between intervention and outcome, but none of these tools linked Length of follow-up to RoB. Therefore, we recommend tool developers to clarify not only the purpose of their tools but also the relevance of their signalling questions to any user needs (eg, RoB assessment). We also advise that in future research the relationships between quality items and user needs will be investigated in more detail.

Our study has a number of limitations. One limitation is that, some tools identified by our study were originally developed for purposes beyond assessing methodological

quality of reporting of NRSIs, so our study could not cover all potentials of these tools. For example, the GRADE framework was mainly designed for addressing certainty of evidence, such as indirectness (ie, whether interventions were compared directly), and for making relevant clinical practice recommendations. While it mentions RoB (eg, publication bias), its main purpose is to illustrate how to grade quality of evidence, rather than to function as an exact quality appraisal tool. In other words, the GRADE allows users to use any additional tools to assess NRSI quality.[52] Also, the GRADE checklist was designed for both RCTs and NRSIs, so some criteria might be relatively brief, compared with specifically designed tools, such as RTI Item Bank.[27] Finally, GRADE can be used to estimate and score the quality of evidence for the full body of evidence and not only for individual primary studies. Therefore, tool users who assess NRSIs beyond methodological quality or reporting should consider criteria in addition to those mentioned in our study, for selecting a tool. Another limitation is that, some tools were predecessors of others, but we did not exclude them if they met the inclusion criteria. For example, the ROBINS-I tool was developed from the Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions (ACROBAT-NRSI),[53] and some of their signalling questions differed. Such information on tool linkage may also be considered for tool selection, if available from the tools. Another limitation is that we only searched HTA agencies for grey literature, and the returned hits on the snowballing approach depended on the starting-set articles, so some tools only mentioned by clinical guideline or regulatory organisations, or tools missed by the previous reviews might have been overlooked. Also, only one researcher (MH) traced versions of tools, by following reference lists of the identified studies and by visiting websites of the online tools. Consequently, the most up-to-date version of a tool might be missing, and the extent to which a quality item was described by a tool might be underestimated. As existing appraisal tools are improved continuously and new tools are being developed (eg, the HARmonized Protocol Template to Enhance Reproducibility (HARPER) and Authentic Transparent Relevant Accurate Track-Record (ATRAcTR)),[54 55] an online platform that automatically identifies appraisal tools and summarises tool information is promising. Such platforms have already been established for tools for assessing observational studies for exposures that were not controlled by investigators (eg, dietary patterns).[56] Another limitation is that we categorised criteria of a quality item as 'sufficient' or 'brief' for each tool, based on whether an explanation was provided for the criteria. Though consensus was reached among authors, and all tool criteria were independently reviewed by two researchers, tool users might question the feasibility of such categorisation when selecting a tool. Additionally, as we categorised quality items based on the order of conducting an NRSI (ie, from study design to results presentation), we did not provide specific suggestions on how to select tools based on bias categories. For example, motivational bias, which would occur when judgements are influenced by the desirability or undesirability of events or outcomes, may affect reporting and measurement of patient outcomes and adherence to healthcare interventions.[57 58] Although the items Conflict of interest and Outcome measurement are relevant to motivational bias, we did not investigate their relationships. Hence, we recommend for future research to bridge our quality items to all potential categories of bias, then test whether a tool selected based on such categorisation, together with recommendations from previous reviews, can really satisfy tool users. It is also worth noting that, the target audience of this review and content analysis could be decision-makers who assess the general quality of an NRSI, NRSI performers who may report quality of their studies, or developers of relevant appraisal tools. However, when users focus on a specific type of concern (eg, causal effect or data quality), some methodological guidance investigating the specific issue or tools beyond the healthcare field (eg, social science) really exist[59 60] and may be referred to by users. In addition, the tools for diagnosis studies, prognosis studies and secondary studies were beyond the scope of our study, and relevant users may refer to other studies, such as Quigley *et al*[14], for further information. Moreover, some frameworks specifically designed for assessing data quality, for example, in terms of data structures and completeness, have been published, and some of their instructions may also be considered as criteria for assessing NRSI quality.[61–65] While evaluating these frameworks is beyond the scope of this study, we recommend tool developers to refer to these frameworks when they define relevant criteria or signalling questions in the future.

CONCLUSION

Most of the appraisal tools for NRSIs have their own strengths, but none of them could address all quality concerns relevant to these studies. Even the most comprehensive tools could be complemented with items from other tools. With information on how sufficiently a tool describes a quality item, tool users might broaden their horizons on quality concerns of non-randomised studies to be considered and might select a tool that more completely satisfies their needs. We suggest decision-makers, researchers and tool developers consider the quality-item level heterogeneity when selecting a tool or identifying a research gap.

Acknowledgements This research was once published as an abstract in 2022-11, ISPOR Europe 2022, Vienna, Austria. The citation was: Jiu L, Hartog MK, Wang J, *et al*. OP18 applicability of appraisal tools of real-world evidence in health technology assessment: a literature review and content analysis. *Value Health*. 2022 Dec 1;25(12):S389.

Contributors LJ designed the study protocol, identified appraisal tools, conducted the content analysis and wrote the manuscript; MH identified appraisal tools, collected data on appraisal tools and conducted the content analysis; JW designed the study protocol, solved the discrepancies on identification of appraisal tools and edited the manuscript; RAV designed the study protocol and edited the manuscript; OK provided assistance on coding of quality items and edited the manuscript; AM-T

edited the manuscript; WG edited the manuscript, and was responsible for the overall content as the guarantor.

Funding This research was performed as part of the HTx project. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825162.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iDs
Li Jiu http://orcid.org/0000-0001-6546-0778
Aukje K Mantel-Teeuwisse http://orcid.org/0000-0002-8782-0698

REFERENCES

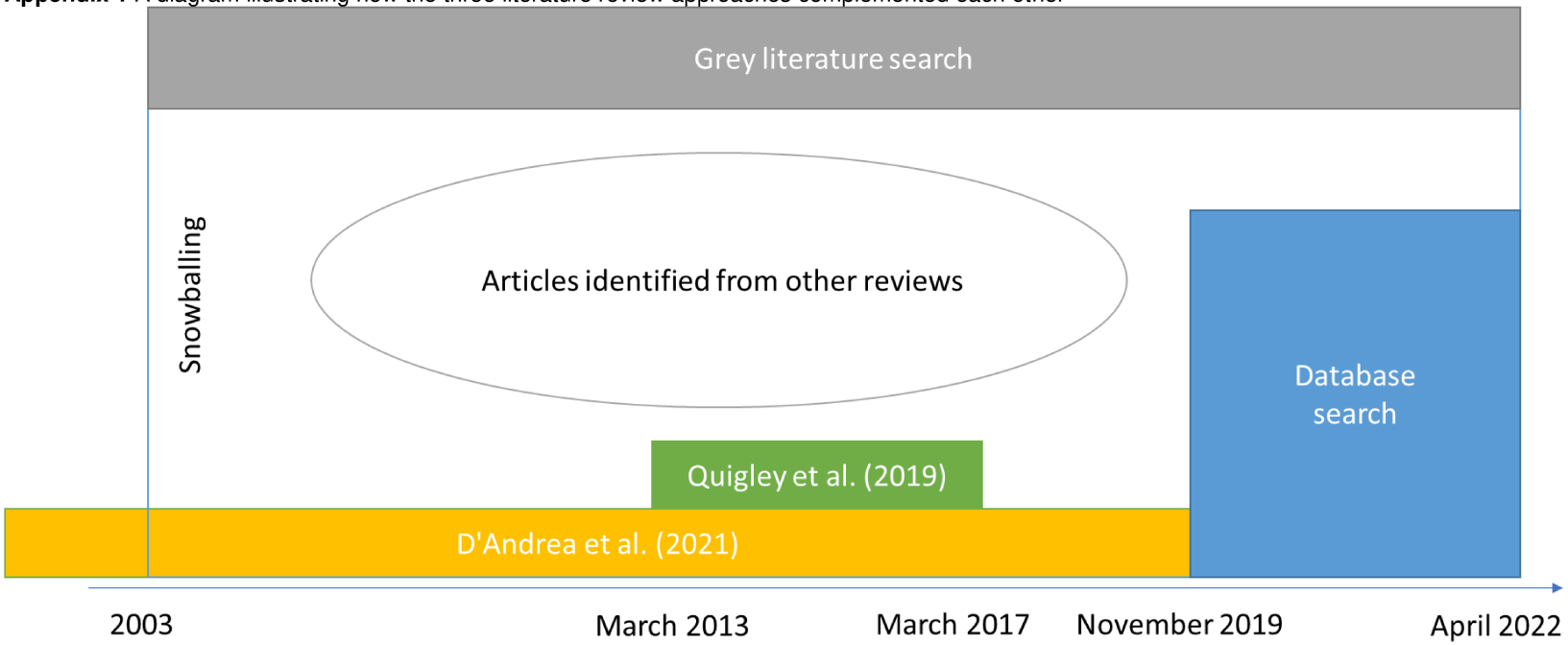
1 Makady A, de Boer A, Hillege H, *et al*. What is real-world data? a review of definitions based on literature and stakeholder interviews. *Value Health* 2017;20:858–65. 10.1016/j.jval.2017.03.008 Available: mid: https://www.valueinhealthjournal.com/article/S1098-3015(17)30171-7/fulltext?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1098301517301717%3Fshowall%3Dtrue
2 Reeves BC, Deeks JJ, Higgins JPT, *et al*. Including non-randomized studies on intervention effects. In: Higgins JPT, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions version* . 2022: 6. 3. Available: https://training.cochrane.org/handbook/current/chapter24#:~:text=NRSI%20are%20defined%20here%20as,of%20individuals)%20to%20intervention%20groups
3 Higgins J, Morgan R, Rooney A, *et al*. Risk of bias in non-randomized studies - of exposure (ROBINS-E). 2022 Available: https://www.riskofbias.info/welcome/robins-e-tool
4 Katkade VB, Sanders KN, Zou KH. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. Multidiscip Healthc. 2018. Available: https://www.tandfonline.com/doi/full/10.2147/JMDH.S160029
5 Baumfeld Andre E, Carrington N, Siami FS, *et al*. The Current Landscape and Emerging Applications for Real-World Data in Diagnostics and Clinical Decision Support and its Impact on Regulatory Decision Making. *Clin Pharmacol Ther* 2022;112:1172–82. 10.1002/cpt.2565 Available: https://ascpt.onlinelibrary.wiley.com/doi/full/10.1002/cpt.2565
6 Makady A, van Veelen A, Jonsson P, *et al*. Using Real-World Data in Health Technology Assessment (HTA) Practice: A Comparative Study of Five HTA Agencies. *Pharmacoeconomics* 2018;36:359–68. 10.1007/s40273-017-0596-z Available: mid:https://link.springer.com/article/10.1007/s40273-017-0596-z
7 Kent S, Salcher-Konrad M, Boccia S, *et al*. The use of nonrandomized evidence to estimate treatment effects in health technology assessment. *J Comp Eff Res* 2021;10:1035–43. 10.2217/cer-2021-0108 Available: https://becarispublishing.com/doi/10.2217/cer-2021-0108?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0pubmed
8 Facey KM, Rannanheimo P, Batchelor L, *et al*. Real-world evidence to support Payer/HTA decisions about highly innovative technologies in the EU-actions for stakeholders. *Int J Technol Assess Health Care* 2020:1–10. 10.1017/S026646232000063X Available: https://www.cambridge.org/core/journals/international-journal-of-technology-assessment-in-health-care/article/realworld-evidence-to-support-payerhta-decisions-about-highly-innovative-technologies-in-the-euactions-for-stakeholders/4256A23FBFCFE5E80D80BC379953D1E6
9 Hogervorst MA, Pontén J, Vreman RA, *et al*. Real world data in health technology assessment of complex health technologies. *Front Pharmacol* 2022;13:837302. 10.3389/fphar.2022.837302 Available: mid: https://www.frontiersin.org/articles/10.3389/fphar.2022.837302/full
10 Schünemann HJ, Tugwell P, Reeves BC, *et al*. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Res Synth Methods* 2013;4:49–62. 10.1002/jrsm.1078 Available: mid: https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1078
11 Sterne JA, Hernán MA, McAleenan A, *et al*. Assessing risk of bias in a non-randomized study. In: Higgins JPT, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions version* . Cochrane;2022 [cited 2023 Feb 8]. Chapter 25, 2022: 6. 3. Available: https://training.cochrane.org/handbook/current/chapter-25
12 Orsini LS, Berger M, Crown W, *et al*. Improving transparency to build trust in real-world secondary data studies for hypothesis testing-why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Value Health* 2020;23:1128–36. 10.1016/j.jval.2020.04.002 Available: https://www.valueinhealthjournal.com/article/S1098-3015(20)30190-X/fulltext?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS109830152030190X%3Fshowall%3Dtrue
13 D'Andrea E, Vinals L, Patorno E, *et al*. How well can we assess the validity of non-randomised studies of medications? a systematic review of assessment tools. *BMJ Open* 2021;11:e043961. 10.1136/bmjopen-2020-043961 Available: https://bmjopen.bmj.com/content/11/3/e043961.long
14 Quigley JM, Thompson JC, Halfpenny NJ, *et al*. Critical appraisal of nonrandomized studies-A review of recommended and commonly used tools. *J Eval Clin Pract* 2019;25:44–52. 10.1111/jep.12889 Available: https://onlinelibrary.wiley.com/doi/10.1111/jep.12889
15 University of Sheffield. The use of observational data to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data. 2015. Available: https://www.sheffield.ac.uk/nice-dsu/tsds/full-list
16 Evidence Analysis. Quality criteria checklist: primary research. 2023. Available: https://www.andeal.org/vault/2440/web/files/QCC_3.pdf
17 Dreyer NA, Velentgas P, Westrich K, *et al*. The GRACE Checklist for Rating the Quality of Observational Studies of Comparative Effectiveness: A Tale of Hope and Caution. *JMCP* 2014;20:301–8. 10.18553/jmcp.2014.20.3.301 Available: mid: https://www.jmcp.org/doi/10.18553/jmcp.2014.20.3.301?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed
18 Sterne JA, Hernán MA, Reeves BC, *et al*. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919. 10.1136/bmj.i4919 Available: mid: https://www.bmj.com/content/355/bmj.i4919.long
19 HTx. About Htx Project.[Cited. 2022. Available: https://www.htx-h2020.eu/about-htx-project
20 Open science framework (OSF) Registry. 2022. Available: https://osf.io/dashboard
21 Wohlin C. Second-generation systematic literature studies using Snowballing. EASE '16; Limerick Ireland.New York, NY, USA, June 2016:1–6
22 Connected papers. 2022. Available: https://www.connectedpapers.com/about
23 The International network of agencies for health technology assessment members. 2022. Available: https://www.inahta.org/members/members_list/
24 Ouzzani M, Hammady H, Fedorowicz Z, *et al*. Rayyan-a web and mobile app for systematic reviews. *Syst Rev* 2016;5:210:210.:. 10.1186/s13643-016-0384-4 Available: mid: https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-016-0384-4
25 Whiting P, Wolff R, Mallett S, *et al*. A proposed framework for developing quality assessment tools. *Syst Rev* 2017;6:204:204.:. 10.1186/s13643-017-0604-6 Available: https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-017-0604-6
26 Nowell LS, Norris JM, White DE, *et al*. Thematic analysis: Striving to meet the trustworthiness criteria. *Int J Qual Methods* 2017;16. Available: https://journals.sagepub.com/doi/pdf/10.1177/1609406917733847

27 Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. *J Clin Epidemiol* 2012;65:163–78. 10.1016/j.jclinepi.2011.05.008 Available: mid: https://www.jclinepi.com/article/S0895-4356(11)00177-6/fulltext
28 Slim K, Nini E, Forestier D, *et al*. Methodological index for non-randomized studies (MINORS): development and validation of a new instrument. *ANZ J Surg* 2003;73:712–6. 10.1046/j.1445-2197.2003.02748.x Available: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1445-2197.2003.02748.x?sid=nlm%3Apubmed
29 Faillie J-L, Ferrer P, Gouverneur A, *et al*. A new risk of bias checklist applicable to randomized trials, observational studies, and systematic reviews was developed and validated to be used for systematic reviews focusing on drug adverse events. *J Clin Epidemiol* 2017;86:168–75. 10.1016/j.jclinepi.2017.04.023 Available: https://www.jclinepi.com/article/S0895-4356(16)30582-0/fulltext
30 Berger ML, Martin BC, Husereau D, *et al*. A questionnaire to assess the relevance and credibility of observational studies to inform health care decision making: an ISPOR-AMCP-NPC Good practice task force report. *Value Health* 2014;17:143–56. 10.1016/j.jval.2013.12.011 Available: https://www.valueinhealthjournal.com/article/S1098-3015(14)00009-6/fulltext?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1098301514000096%3Fshowall%3Dtrue
31 Joanna Briggs Institute. Critical appraisal tools. 2023. Available: https://jbi.global/critical-appraisal-tools
32 Handu D, Moloney L, Wolfram T, *et al*. Academy of nutrition and dietetics methodology for conducting systematic reviews for the evidence analysis library. *J Acad Nutr Diet* 2016;116:311–8. 10.1016/j.jand.2015.11.008 Available: https://www.jandonline.org/article/S2212-2672(15)01705-0/fulltext
33 Vandenbroucke JP, von Elm E, Altman DG, *et al*. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Ann Intern Med* 2007;147:W163–94. 10.7326/0003-4819-147-8-200710160-00010-w1 Available: https://www.acpjournals.org/doi/full/10.7326/0003-4819-147-8-200710160-00010-w1?rfr_dat=cr_pub++0pubmed&url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org
34 Des Jarlais DC, Lyles C, Crepaz N, *et al*. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *Am J Public Health* 2004;94:361–6. 10.2105/ajph.94.3.361 Available: https://ajph.aphapublications.org/doi/10.2105/ajph.94.3.361?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0pubmed
35 Genaidy AM, Lemasters GK, Lockey J, *et al*. An epidemiological appraisal instrument - a tool for evaluation of epidemiological studies. *Ergonomics* 2007;50:920–60. 10.1080/00140130701237667 Available: mid: https://www.tandfonline.com/doi/abs/10.1080/00140130701237667?journalCode=terg20
36 ENCePP Guide on Methodological Standards in Pharmacoepidemiology (Revision 10). European network of centres for pharmacoepidemiology and pharmacovigilance guide on methodological standards in pharmacoepidemiology (ENCePP), . 2022Available: https://www.encepp.eu/standards_and_guidances/methodologicalGuide.shtml
37 Public Policy Committee. International society of pharmacoepidemiology. guidelines for good pharmacoepidemiology practice (GPP). *Pharmacoepidemiol Drug Saf* 2016:2–10. 10.1002/pds.3891.pmid Available: https://onlinelibrary.wiley.com/doi/10.1002/pds.3891
38 Tseng TY, Breau RH, Fesperman SF, *et al*. Evaluating the evidence: the methodological and reporting quality of comparative observational studies of surgical interventions in urological publications. *BJU Int* 2009;103:1026–31. 10.1111/j.1464-410X.2008.08155.x Available: https://bjui-journals.onlinelibrary.wiley.com/doi/10.1111/j.1464-410X.2008.08155.x
39 Wang SV, Schneeweiss S, Berger ML, *et al*. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1.0. *Value Health* 2017;20:1009–22.
40 Benchimol EI, Smeeth L, Guttmann A, *et al*. The REporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLoS Med* 2015;12:e1001885. 10.1371/journal.pmed.1001885 Available: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001885
41 National institute for health and care excellence (NICE)]. The guidelines manual: Appendices B–I. 2012. Available: https://www.nice.org.uk/process/pmg6/resources/the-guidelines-manual-appendices-bi-pdf-3304416006853
42 Kennedy CE, Fonner VA, Armstrong KA, *et al*. The Evidence Project risk of bias tool: assessing study rigor for both randomized and non-randomized intervention studies. *Syst Rev* 2019;8:3:3.:. 10.1186/s13643-018-0925-0 Available: mid: https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-018-0925-0
43 NHS Wales. A systematic approach to identifying the evidence. project methodology 5. Cardiff: information services UWCM. 2004. Available: https://www2.nphs.wales.nhs.uk/VulnerableAdultsDocs.nsf/0/3811E6F969F2D3FC8025783E005B59AB/$file/housingrelatedsupport_descriptive_evidencereview_final_200111.doc?OpenElement
44 NHS Wales. Questions to assist with the critical appraisal of a cross-sectional study (type IV evidence). 2022. Available: https://www2.nphs.wales.nhs.uk/PubHObservatoryProjDocs.nsf/($All)/E7B0C80995DC1BA380257DB80037C699/$File/Cross%20sectional%20study%20checklist.docx?OpenElement
45 Critical appraisal tools. 2022. Available: https://www.cardiff.ac.uk/specialist-unit-for-review-evidence/resources/critical-appraisal-checklists
46 Ma L-L, Wang Y-Y, Yang Z-H, *et al*. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? *Mil Med Res* 2020;7. 10.1186/s40779-020-00238-8 Available: https://mmrjournal.biomedcentral.com/articles/10.1186/s40779-020-00238-8
47 Juni P. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 2001;323:42–6. 10.1136/bmj.323.7303.42 Available: https://pubmed.ncbi.nlm.nih.gov/11440947
48 Makady A, van Veelen A, Jonsson P, *et al*. Using Real-World Data in Health Technology Assessment (HTA) Practice: A Comparative Study of Five HTA Agencies. *Pharmacoeconomics* 2018;36:359–68. 10.1007/s40273-017-0596-z Available: https://pubmed.ncbi.nlm.nih.gov/29214389
49 Franklin JM, Glynn RJ, Martin D, *et al*. Evaluating the Use of Nonrandomized Real-World Data Analyses for Regulatory Decision Making. *Clin Pharma and Therapeutics* 2019;105:867–77. 10.1002/cpt.1351 Available: https://ascpt.onlinelibrary.wiley.com/toc/15326535/105/4
50 Viswanathan M, Patnode CD, Berkman ND, *et al*. Recommendations for assessing the risk of bias in systematic reviews of health-care interventions. *J Clin Epidemiol* 2018;97:26–34. 10.1016/j.jclinepi.2017.12.004 Available: https://www.jclinepi.com/article/S0895-4356(17)31066-1/fulltext
51 Blagojevic M, Jinks C, Jeffery A, *et al*. Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis. *Osteoarthritis Cartilage* 2010;18:24–33. 10.1016/j.joca.2009.08.010 Available: https://www.oarsijournal.com/article/S1063-4584(09)00225-8/fulltext
52 Guyatt GH, Oxman AD, Vist G, *et al*. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *J Clin Epidemiol* 2011;64:407–15. 10.1016/j.jclinepi.2010.07.017 Available: https://www.jclinepi.com/article/S0895-4356(10)00413-0/fulltext
53 University of Bristol [Internet. Archived tool: a Cochrane risk of bias assessment tool for non-randomized studies of interventions (ACROBAT-NRSI). 2023. Available: https://www.bristol.ac.uk/population-health-sciences/centres/cresyda/barr/riskofbias/robins-i/acrobat-nrsi
54 Wang SV, Pottegård A, Crown W, *et al*. HARmonized protocol template to enhance reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: a good practices report of a joint ISPE/ISPOR task force. *Value Health* 2022;25:1663–72.
55 Berger ML, Crown WH, Li JZ, *et al*. ATRAcTR (Authentic Transparent Relevant Accurate Track-Record): a screening tool to assess the potential for real-world data sources to support creation of credible real-world evidence for regulatory decision-making. *Health Serv Outcomes Res Method* 2023;29:1–8. 10.1007/s10742-023-00319-w Available: https://link.springer.com/article/10.1007/s10742-023-00319-w
56 Wang Z, Taylor K, Allman-Farinelli M, *et al*. A systematic review: tools for assessing methodological quality of human observational studies. *MetaArXiv* [Preprint].
57 Arjmand EM, Shapiro JA, Shah RK, *et al*. Human Error and Patient Safety: Managing Cognitive and Motivational Bias in Medical Decision Making. *Otolaryngol--Head Neck Surg* 2014;151. 10.1177/0194599814538403a2 Available: https://aao-hnsfjournals.onlinelibrary.wiley.com/doi/10.1177/0194599814538403a2
58 Montibeller G, von Winterfeldt D. Cognitive and Motivational Biases in Decision and Risk Analysis. *Risk Anal* 2015;35:1230–51. 10.1111/risa.12360 Available: https://pubmed.ncbi.nlm.nih.gov/25873355
59 WILEY online library. How to appraise the studies: an introduction to assessing study quality. 2023. Available: https://onlinelibrary.wiley.com/doi/10.1002/9780470754887.ch5
60 Sharma Waddington H, Cairncross S. PROTOCOL: water, sanitation and hygiene for reducing childhood mortality in Low- and middle-

income countries. *Campbell Syst Rev* 2021;17:e1135. 10.1002/cl2.1135 Available: https://pubmed.ncbi.nlm.nih.gov/37050969
61 Food and Drug Administration (FDA). Framework for FDA's real-world evidence program. 2023. Available: https://www.fda.gov/media/120060/download
62 European Medicines Agency (EMA). Data quality framework for EU medicines regulation. 2023. Available: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf
63 Duke Margolis Center for Health Policy. Determining real-world data's fitness for use and the role of reliability. n.d. Available: https://healthpolicy.duke.edu/sites/default/files/2019-11/rwd_reliability.pdf
64 Kahn MG, Callahan TJ, Barnard J, *et al*. A Harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS* 2016;4:1244.
65 Schmidt CO, Struckmann S, Enzenbach C, *et al*. Facilitating Harmonized data quality assessments. A data quality framework for observational health research data collections with software Implementations in R. *BMC Med Res Methodol* 2021;21:63.
66 Bishop FL, Prescott P, Chan YK, *et al*. Prevalence of complementary medicine use in pediatric cancer: a systematic review. *Pediatrics* 2010;125:768–76. 10.1542/peds.2009-1775 Available: https://pubmed.ncbi.nlm.nih.gov/20308209
67 Pluye P, Gagnon M-P, Griffiths F, *et al*. A scoring system for appraising mixed methods research, and Concomitantly appraising qualitative, quantitative and mixed methods primary studies in mixed studies reviews. *International Journal of Nursing Studies* 2009;46:529–46. 10.1016/j.ijnurstu.2009.01.009 Available: https://pubmed.ncbi.nlm.nih.gov/19233357
68 Young JM, Solomon MJ. How to critically appraise an article. *Nature Clinical Practice Gastroenterology & Hepatology* 2009;6:82–91. 10.1038/ncpgasthep1331 Available: https://pubmed.ncbi.nlm.nih.gov/19153565
69 Atluri S, Datta S, Falco FJE, *et al*. Systematic review of diagnostic utility and therapeutic effectiveness of Thoracic facet joint interventions. *Pain Physician* 2008;11:611–29. Available: https://pubmed.ncbi.nlm.nih.gov/18850026
70 Heller RF, Verma A, Gemmell I, *et al*. Critical appraisal for public health: a new checklist. *Public Health* 2008;122:92–8. 10.1016/j.puhe.2007.04.012 Available: https://pubmed.ncbi.nlm.nih.gov/17765937
71 Weightman AL, Mann MK, Sander L, Turley RL. Health evidence bulletins Wales: A systematic approach to identifying the evidence. 2004. Available: http://www.hebw.cf.ac.uk/projectmethod/title.htm
72 Thomas BH, Ciliska D, Dobbins M, *et al*. A process for systematically reviewing the literature: providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing* 2004;1:176–84. 10.1111/j.1524-475X.2004.04006.x Available: https://pubmed.ncbi.nlm.nih.gov/17163895
73 Rangel SJ, Kelsey J, Colby CE, *et al*. Development of a quality assessment scale for retrospective clinical studies in pediatric surgery. *Journal of Pediatric Surgery* 2003;38:390–6. 10.1053/jpsu.2003.50114 Available: https://pubmed.ncbi.nlm.nih.gov/12632355

Appendix

Appendix 1 A diagram illustrating how the three literature review approaches complemented each other



Appendix 2 Reference list of the included studies reviewing or describing appraisal tools for non-randomized studies of interventions

1. D'Andrea E, Vinals L, Patorno E, Franklin JM, Bennett D, Largent JA, Moga DC, Yuan H, Wen X, Zullo AR, Debray TP. How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools. BMJ open. 2021 Mar 1;11(3):e043961.
2. Sanderson S, Tatt ID, Higgins J. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. International journal of epidemiology. 2007 Jun 1;36(3):666-76.Methodological quality assessment tools of non-experimental studies: a systematic review evaluating non-randomised intervention studies.
3. Scott HS. Systems to rate the strength of scientific evidence.
4. Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. International Journal for Quality in Health Care. 2004 Feb 1;16(1):9-18.
5. Ma LL, Wang YY, Yang ZH, Huang D, Weng H, Zeng XT. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better?. Military Medical Research. 2020 Dec;7:1-1.
6. Zeng X, Zhang Y, Kwong JS, Zhang C, Li S, Sun F, Niu Y, Du L. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. Journal of evidence-based medicine. 2015 Feb;8(1):2-10.
7. Farrah K, Young K, Tunis MC, Zhao L. Risk of bias tools in systematic reviews of health interventions: an analysis of PROSPERO-registered protocols. Systematic reviews. 2019 Dec;8(1):1-9.
8. Quigley JM, Thompson JC, Halfpenny NJ, Scott DA. Critical appraisal of nonrandomized studies—a review of recommended and commonly used tools. Journal of Evaluation in Clinical Practice. 2019 Feb;25(1):44-52.
9. Patole S. Systematic Reviews and Meta-Analyses of Non-randomised Studies. InPrinciples and Practice of Systematic Reviews and Meta-Analysis 2021 Jun 27 (pp. 139-146). Cham: Springer International Publishing.
10. Zeng X, Zhang Y, Kwong JS, Zhang C, Li S, Sun F, Niu Y, Du L. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. Journal of evidence-based medicine. 2015 Feb;8(1):2-10.
11. Page MJ, McKenzie JE, Higgins JP. Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review. BMJ open. 2018 Mar 1;8(3):e019703.
12. Waddington H, Aloe AM, Becker BJ, Djimeu EW, Hombrados JG, Tugwell P, NOS G, Reeves B. Quasi-experimental study designs series—paper 6: risk of bias assessment. Journal of Clinical Epidemiology. 2017 Sep 1;89:43-52.
13. Saunders LD, Soomro GM, Buckingham J, Jamtvedt G, Raina P. Assessing the methodological quality of nonrandomized intervention studies. Western Journal of Nursing Research. 2003 Mar;25(2):223-37.
14. Brand J, Hardy R, Monroe E. Research pearls: Checklists and flowcharts to improve research quality. Arthroscopy: The Journal of Arthroscopic & Related Surgery. 2020 Jul 1;36(7):2030-8.
15. Lundh A, Rasmussen K, Østengaard L, Boutron I, Stewart LA, Hróbjartsson A. Systematic review finds that appraisal tools for medical research studies address conflicts of interest superficially. Journal of Clinical Epidemiology. 2020 Apr 1;120:104-15.
16. Yao X, Florez ID, Zhang P, Zhang C, Zhang Y, Wang C, Liu X, Nie X, Wei B, Ghert MA. Clinical research methods for treatment, diagnosis, prognosis, etiology, screening, and prevention: a narrative review. Journal of Evidence-Based Medicine. 2020 May;13(2):130-6.
17. Liebherz S, Schmidt N, Rabung S. How to assess the quality of psychotherapy outcome studies: A systematic review of quality assessment criteria. Psychotherapy Research. 2016 Sep 2;26(5):573-89.
18. Tate RL, Douglas J. Use of reporting guidelines in scientific writing: PRISMA, CONSORT, STROBE, STARD and other resources. Brain Impairment. 2011 May;12(1):1-21.
19. Viswanathan et al. 2018 M, Berkman ND, Dryden DM, Hartling L. Assessing risk of bias and confounding in observational studies of interventions or exposures: further development of the RTI item bank.

20. Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B, Davies P, Kleijnen J, Churchill R. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. Journal of clinical epidemiology. 2016 Jan 1;69:225-34.
21. Real-world studies for the assessment of medicinal products and medical devices. https://www.has-sante.fr/upload/docs/application/pdf/2021-06/real-world_studies_for_the_assessment_of_medicinal_products_and_medical_devices.pdf.
22. Quality Appraisal Checklist for Case Series Studies and Instructions for Use . https://www.ihe.ca/publications/ihe-quality-appraisal-checklist-for-case-series-studies.
23. Kmet LM, Cook LS, Lee RC. Standard quality assessment criteria for evaluating primary research papers from a variety of fields.
24. NICE real-world evidence framework. https://www.nice.org.uk/corporate/ecd9/chapter/overview.
25. Lewin S, Booth A, Glenton C, Munthe-Kaas H, Rashidian A, Wainwright M, Bohren MA, Tunçalp Ö, Colvin CJ, Garside R, Carlsen B. Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series. Implementation Science. 2018 Jan;13(1):1-0.
26. Development of a quality appraisal tool for case series studies using a modified Delphi technique. https://cobe.paginas.ufsc.br/files/2014/10/MOGA.Case-series.pdf.
27. A Guide to Real-World Evaluations of Primary Care Interventions: Some Practical Advice. https://www.ahrq.gov/ncepcr/tools/pcmh/implement/evaluation-guide.html.

Appendix 3 Reference list of appraisal tools for non-randomized studies of interventions

No.	Full name	Abbreviation	Reference
1	REal Life EVidence AssessmeNt Tool	RELEVANT	Campbell JD, Perry R, Papadopoulos NG, Krishnan J, Brusselle G, Chisholm A, Bjermer L, Thomas M, Van Ganse E, Van Den Berge M, Quint J. The REal Life EVidence AssessmeNt Tool (RELEVANT): development of a novel quality assurance asset to rate observational comparative effectiveness research studies. Clinical and translational allergy. 2019;9(1):21.
2	Graphic Appraisal Tool for Epidemiological Studies	GATE	Jackson R, Ameratunga S, Broad J, Connor J, Lethaby A, Robb G, NOS S, Glasziou P, Heneghan C. The GATE frame: critical appraisal with pictures. BMJ Evidence-Based Medicine. 2006 Apr 1;11(2):35-8.
3	Mixed Methods Appraisal Tool	MMAT	Hong QN, Gonzalez-Reyes A, Pluye P. Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). Journal of evaluation in clinical practice. 2018 Jun;24(3):459-67.
4	Critical Appraisal Skills Programme Tool	CASP	Long, H. A., French, D. P., & Brooks, J. M. (2020). Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. Research Methods in Medicine & Health Sciences, 1(1), 31-42.
5	Critical Appraisal Tools of Specialist Unit for Review Evidence	SURE	Available from : https://www.cardiff.ac.uk/specialist-unit-for-review-evidence/resources/critical-appraisal-checklists.
6	Joanna Briggs Institute's Critical Appraisal Tools	JBI	Available from : https://jbi.global/critical-appraisal-tools.
7	Risk Of Bias In Non-randomised Studies - of Interventions	ROBINS-I	Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. bmj. 2016 Oct 12;355.

8	Comparative Effectiveness Research Collaborative Initiative Questionnaire	CER-CI	Berger ML, Martin BC, Husereau D, Worley K, Allen JD, Yang W, Quon NC, Mullins CD, Kahler KH, Crown W. A questionnaire to assess the relevance and credibility of observational studies to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value in health. 2014 Mar 1;17(2):143-56.
9	Good ReseArch for Comparative Effectiveness Checklist	GRACE	Dreyer NA, Velentgas P, Westrich K, Dubois R. The GRACE checklist for rating the quality of observational studies of comparative effectiveness: a tale of hope and caution. Journal of Managed Care Pharmacy. 2014 Mar;20(3):301-8.
10	Quality Assessment Tool of National Heart, Lung, and Blood Institute	NIH	Available from : https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools.
11	NA	Weightman et al. 2004	Weightman AL, Mann MK, Sander L, Turley RL. Health evidence bulletins Wales: A systematic approach to identifying the evidence. Project methodology 5. Cardiff: Information Services UWCM. 2004.
12	Risk of Bias Assessment tool for Non-randomized Studies Tool	RoBANS	Available from : https://abstracts.cochrane.org/2011-madrid/risk-bias-assessment-tool-non-randomized-studies-robans-development-and-validation-new.
13	Research Triangle Institute Item Bank	RTI Item Bank	Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. Journal of clinical epidemiology. 2012 Feb 1;65(2):163-78.
14	Scottish Intercollegiate Guidelines Network Checklists	SIGN	Available from : https://www.sign.ac.uk/what-we-do/methodology/checklists.
15	The Montreal Critical Appraisal Worksheet	Montreal	Available from : https://guides.bib.umontreal.ca/ckfinder/ckeditor_assets/attachments/critical-appraisal-worksheet.pdf.

16	STrengthening the Reporting of OBservational studies in Epidemiology Checklists	STROBE	Available from : https://www.strobe-statement.org.
17	Transparent Reporting of Evaluations with Nonrandomized Designs	TREND	Des Jarlais DC, Lyles C, Crepaz N, Trend Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. American journal of public health. 2004 Mar;94(3):361-6.
18	A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions	ACROBAT-NRSI	Sterne JA, Higgins J, Reeves B. A Cochrane risk of bias assessment tool: for non-randomized studies of interventions (ACROBAT-NRSI). Version. 2014 Sep;1(0):24.
19	Methodology Index for Non-randomized Studies	MINORS	Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (MINORS): development and validation of a new instrument. ANZ journal of surgery. 2003 Sep;73(9):712-6.
20	Grades of Recommendation, Assessment, Development and Evaluation Guideline	GRADE	Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). Journal of clinical epidemiology. 2011 Apr 1;64(4):407-15.
21	NA	Rangel et al. 2003	Rangel SJ, Kelsey J, Colby CE, Anderson J, Moss RL. Development of a quality assessment scale for retrospective clinical studies in pediatric surgery. Journal of pediatric surgery. 2003 Mar 1;38(3):390-6.
22	NA	Thomas et al. 2004	Thomas BH, Ciliska D, Dobbins M, Micucci S. A process for systematically reviewing the literature: providing the research evidence for public health nursing interventions. Worldviews on Evidence-Based Nursing. 2004 Sep;1(3):176-84.
23	NA	Atluri et al. 2008	Atluri S, Datta S, Falco F, Lee M. Systematic review of diagnostic utility and therapeutic effectiveness of thoracic facet joint interventions. Pain Physician. 2008;11(5):611.

24	NA	Bishop et al. 2010	Bishop FL, Prescott P, Chan YK, Saville J, von Elm E, Lewith GT. Prevalence of complementary medicine use in pediatric cancer: a systematic review. Pediatrics. 2010 Apr;125(4):768-76.
25	NA	Blagojevic et al. 2010	Blagojevic M, Jinks C, Jeffery A, Jordan 1. Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis. Osteoarthritis and cartilage. 2010 Jan 1;18(1):24-33.
26	NA	Genaidy et al. 2007	Genaidy AM, Lemasters GK, Lockey J, Succop P, Deddens J, Sobeih T, Dunning K. An epidemiological appraisal instrument–a tool for evaluation of epidemiological studies. Ergonomics. 2007 Jun 1;50(6):920-60.
27	Harm Critical Appraisal Worksheet	Harm	Available from : https://www.colleaga.org/tools/harm-critical-appraisal-worksheet.
28	NA	Tseng et al. 2009	Tseng TY, Breau RH, Fesperman SF, Vieweg J, Dahm P. Evaluating the evidence: the methodological and reporting quality of comparative observational studies of surgical interventions in urological publications. BJU international. 2009 Apr;103(8):1026-31.
29	NHS Wales Questions to Assist with the Critical Appraisal of a Cross-Sectional Study	NHS Wales	Available from : https://www2.nphs.wales.nhs.uk/PubHObservatoryProjDocs.nsf/($All)/E7B0C80995DC1BA380257DB80037C699/$File/Cross%20sectional%20study%20checklist.docx?OpenElement.
30	Newcastle-Ottawa Scale	NOS	Available from : https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.
31	Academy of Nutrition and Dietetics ANDQ (Primary Research)	ANDQ	Available from : https://www.andeal.org/vault/2440/web/files/QCC_3.pdf.
32	Guidelines manual of National Institute for Health and Care Excellence: Appendices D-E	NICE	Available from : https://www.nice.org.uk/process/pmg6/resources/the-guidelines-manual-appendices-bi-pdf-3304416006853.

33	Institute of Health Economic Quality Appraisal Tool for Case-Series Studies	IHE	Available from : https://www.ihe.ca/advanced-search/development-of-a-quality-appraisal-tool-for-case-series-studies-using-a-modified-delphi-technique.
34	Appraisal tool for Cross-Sectional Studies	AXIS	Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ open. 2016 Dec 1;6(12):e011458.
35	Agency for Healthcare Research and Quality Methodology Checklist	AHRQ	Viswanathan M, Patnode CD, Berkman ND, Bass EB, Chang S, Hartling L, Murad MH, Treadwell JR, Kane RL. Assessing the risk of bias in systematic reviews of health care interventions. Methods guide for effectiveness and comparative effectiveness reviews [Internet]. 2017 Dec 13.
36	NA	Pluye et al. 2009	Pluye P, Gagnon MP, Griffiths F, Johnson-Lafleur J. A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in mixed studies reviews. International journal of nursing studies. 2009 Apr 1;46(4):529-46.
37	NA	Heller et al. 2008	Heller RF, Verma A, Gemmell I, Harrison R, Hart J, Edwards R. Critical appraisal for public health: a new checklist. Public health. 2008 Jan 1;122(1):92-8.
38	CAse REport (CARE) Guidelines Checklist	CARE	CARE JJ, Kienle G, Altman DG, Moher D, Sox H, Riley D. The CARE guidelines: consensus-based clinical case reporting guideline development. Journal of medical case reports. 2013 Dec;7(1):1-6.
39	NA	Faillie et al. 2017	Faillie JL, Ferrer P, Gouverneur A, Driot D, Berkemeyer S, Vidal X, Martínez-Zapata MJ, Huerta C, Castells X, Rottenkolber M, Schmiedl S. A new risk of bias checklist applicable to randomized trials, observational studies, and systematic reviews was developed and validated to be used for systematic reviews focusing on drug adverse events. Journal of Clinical Epidemiology. 2017 Jun 1;86:168-75.

40	Interventional Pain Management Techniques – Quality Appraisal of Reliability and Risk of Bias Assessment for Nonrandomized Studies	IPM-QRBNR	Manchikanti L, Hirsch JA, Heavner JE, Cohen SP, Benyamin RM, Sehgal N, Falco F, Vallejo R, Onyewu CO, Zhu J, Kaye AD. Development of an interventional pain management specific instrument for methodologic quality assessment of nonrandomized studies of interventional techniques. Pain physician. 2014;17(3):E291.
41	NA	Handu et al. 2016	Handu D, Moloney L, Wolfram T, Ziegler P, Acosta A, Steiber A. Academy of Nutrition and Dietetics methodology for conducting systematic reviews for the Evidence Analysis Library. Journal of the Academy of Nutrition and Dietetics. 2016 Feb;116(2):311-8.
42	NA	Viswanathan et al. 2018	Viswanathan M, Patnode CD, Berkman ND, Bass EB, Chang S, Hartling L, Murad MH, Treadwell JR, Kane RL. Recommendations for assessing the risk of bias in systematic reviews of health-care interventions. Journal of clinical epidemiology. 2018 May 1;97:26-34.
43	NA	Young et al. 2009	Young JM, Solomon MJ. How to critically appraise an article. Nature Clinical Practice Gastroenterology & Hepatology. 2009 Feb;6(2):82-91.
44	International Society of Pharmacoepidemiology Guidelines for Good Pharmacoepidemiology Practice	ISPE	Public Policy Committee, International Society of Pharmacoepidemiology. Guidelines for good pharmacoepidemiology practice (GPP). pharmacoepidemiology and drug safety. 2016 Jan;25(1):2-10.
45	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance Guide on Methodological Standards in	ENCePP	Available from: https://www.encepp.eu/standards_and_guidances/methodologicalGuide.shtml.

	Pharmacoepidemiology		
46	Joint Task Force between the International Society for Pharmacoepidemiology and the International Society for Pharmacoeconomics and Outcomes Research	ISPE-ISPOR	Wang SV, Schneeweiss S, Berger ML, Brown J, de Vries F, Douglas I, Gagne JJ, Gini R, Klungel O, Mullins CD, Nguyen MD. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1. 0. Value in health. 2017 Sep 1;20(8):1009-22.
47	Basque Office for Health Technology Assessment Tool	OSTEBA	Available from : http://www.lecturacritica.com/en/plataforma-flc_para-que-sirve-la-plataforma-web.php.
48	NA	Kennedy et al. 2019	Kennedy CE, Fonner VA, Armstrong KA, Denison JA, Yeh PT, O'Reilly KR, Sweat MD. The Evidence Project risk of bias tool: assessing study rigor for both randomized and non-randomized intervention studies. Systematic reviews. 2019 Dec;8(1):1-0.
49	REporting of studies Conducted using Observational Routinely-collected Data Checklist	RECORD	Available from: https://www.record-statement.org/checklist.php.

Appendix 4 Ranking of appraisal tools based on the number of quality items addressed or sufficiently described

Ranking	Sufficiently described?		Addressed by the tool?	
	Appraisal tool	**Number of items(%)**	**Appraisal tool**	**Number of items(%)**
Methodology				
1	RTI Item Bank	18 (69)	RTI Item Bank	20 (77)
2	Faillie et al. 2017	8 (31)	Faillie et al. 2017	17 (65)
3	MINORS	8 (31)	ANDQ	14 (54)
4	ANDQ	8 (31)	NICE	12 (46)
5	ROBINS-I	7 (27)	CER-CI	11 (42)
6	NIH	7 (27)	CASP	11 (42)
7	Genaidy et al. 2007	6 (23)	SIGN	11 (42)
8	CER-CI	5 (19)	JBI	11 (42)
9	Handu et al. 2016	5 (19)	Heller et al. 2008	11 (42)
10	JBI	5 (19)	MINORS	10 (38)
11	GRACE	5 (19)	Blagojevic et al. 2010	10 (38)
12	NICE	4 (15)	ROBINS-I	10 (38)
13	Kennedy et al. 2019	4 (15)	Handu et al. 2016	9 (35)
14	ACROBAT-NRSI	4 (15)	AXIS	9 (35)
15	IPM-QRBNR	3 (12)	IHE	9 (35)
Reporting				
1	STROBE	14 (54)	STROBE	17 (65)
2	TREND	9 (35)	ENCePP	15 (58)
3	RECORD	8 (31)	TREND	14 (54)
4	Genaidy et al. 2007	8 (31)	Genaidy et al. 2007	12 (46)
5	ENCePP	7 (27)	RECORD	12 (46)
6	ISPE	5 (19)	RELEVANT	11 (42)
7	Tseng et al. 2009	4 (15)	ISPE	9 (35)
8	ISPE-ISPOR	3 (12)	SURE	9 (35)

9	ANDQ	3 (12)	Tseng et al. 2009	7 (27)
10	RTI Item Bank	3 (12)	ISPE-ISPOR	7 (27)
11	MINORS	3 (12)	ANDQ	6 (23)
12	SURE	2 (8)	IHE	6 (23)
13	IHE	2 (8)	NICE	6 (23)
14	CER-CI	2 (8)	Heller et al. 2008	5 (19)
15	Rangel et al. 2003	2 (8)	CER-CI	5 (19)

Appendix 5 Ranking of appraisal tools based on the number of quality items on methodology, which were addressed or sufficiently described, segmented by quality domains

Ranking	Appraisal tool	Number of items(%)	
		Sufficiently described?	Addressed?
Domain 1_Study design			
1	RTI Item Bank	7	8
2	MINORS	3	3
3	Faillie et al. 2017	2	5
4	CER-CI	2	5
5	ANDQ	2	4
6	SURE	2	4
7	NIH	2	2
8	NHS Wales	2	2
9	Kennedy et al. 2019	2	2
10	Handu et al. 2016	2	2
11	NICE	1	3
12	Young et al. 2009	1	2
13	JBI	1	2
14	GRADE	1	2
15	ROBINS-I	1	1
16	Genaidy et al. 2007	1	1
17	ISPE	1	1
18	GRACE	1	1
19	NOS	1	1
20	Weightman et al. 2004	1	1
Domain 2_Data quality			
1	RTI Item Bank	7	7
2	MINORS	5	6
3	Faillie et al. 2017	4	7
4	ROBINS-I	4	7
5	ANDQ	4	6

6	Handu et al. 2016	4	5
7	AHRQ	4	5
8	Atluri et al. 2008	3	6
9	NIH	3	4
10	JBI	3	3
11	CASP	3	3
12	Blagojevic et al. 2010	2	7
13	NICE	2	6
14	CER-CI	2	5
15	NOS	2	3
16	GRACE	2	2
17	GRADE	1	7
18	MMAT	1	6
19	Genaidy et al. 2007	1	4
20	Harm	1	4
21	GATE	1	4
22	IHE	1	3
23	SIGN	1	2
24	ACROBAT-NRSI	1	2
25	CARE	1	2
26	Weightman et al. 2004	1	2
Domain 3_Data analysis			
1	ANDQ	3	3
2	Handu et al. 2016	3	3
3	Faillie et al. 2017	2	3
4	ROBINS-I	2	2
5	Montreal	2	2
6	Harm	2	2
7	CER-CI	1	2

8	MINORS	1	2
9	AXIS	1	2
10	STROBE	1	2
11	MMAT	1	1
12	GRACE	1	1
13	CARE	1	1
14	ENCePP	1	1
Domain 4_Results presentation			
1	Faillie et al. 2017	2	2
2	ENCePP	2	2
3	NICE	2	2
4	Handu et al. 2016	1	1
5	JBI	1	1

Appendix 6 Ranking of appraisal tools based on the number of quality items on reporting, which were addressed or sufficiently described, segmented by quality domains

Ranking	Appraisal tool	Number of items(%)	
		Sufficiently described?	Addressed?
Domain 1_Study design			
1	STROBE	5	7
2	ENCePP	4	8
3	RECORD	4	6
4	TREND	3	6
5	Genaidy et al. 2007	3	5
6	SURE	2	8
7	ANDQ	2	5
8	ISPE	2	4
9	Tseng et al. 2009	2	4
10	IHE	2	4
11	ISPE-ISPOR	2	3
12	MINORS	2	2
13	RTI Item Bank	2	2
14	RELEVANT	1	7
15	Rangel et al. 2003	1	3
16	CER-CI	1	3
17	Faillie et al. 2017	1	3
18	Handu et al. 2016	1	3
19	NIH	1	2
20	CASP	1	2
21	AHRQ	1	2
22	ROBINS-I	1	2
23	Harm	1	1
24	NHS Wales	1	1
25	JBI	1	1
26	IPM-QRBNR	1	1

27	Thomas et al. 2004	1	1
Domain 2_Data quality			
1	STROBE	5	5
2	Genaidy et al. 2007	4	6
3	TREND	4	5
4	ENCePP	3	5
5	RECORD	2	4
6	ISPE	2	3
7	IPM-QRBNR	1	3
8	Montreal	1	3
9	ANDQ	1	2
10	ISPE-ISPOR	1	2
11	NOS	1	2
12	CARE	1	1
13	RELEVANT	1	1
14	Faillie et al. 2017	1	1
15	CER-CI	1	1
16	AHRQ	1	1
17	Weightman et al. 2004	1	1
18	Thomas et al. 2004	1	1
Domain 3_Data analysis			
1	STROBE	2	3
2	RELEVANT	1	3
3	SURE	1	2
4	JBI	1	1
5	CER-CI	1	1
6	RoBANS	1	1
7	Young et al. 2009	1	1
8	Atluri et al. 2008	1	1
Domain 4_Results presentation			

1	STROBE	2	2
2	SIGN	1	2
3	IHE	1	1
4	AXIS	1	1
5	NICE	1	1
6	Kennedy et al. 2019	1	1
7	ROBINS-I	1	1
8	Faillie et al. 2017	1	1
9	CER-CI	1	1
10	GRACE	1	1
11	ACROBAT-NRSI	1	1
12	AHRQ	1	1
13	Tseng et al. 2009	1	1

Appendix 7 Coding of appraisal tools for non-randomized studies of interventions
M indicates a criterion is related to apprising methodological quality; R indicate a criterion is related to reporting; The level "1" indicates a quality item was described briefly; the level "2" indicates a quality item was described with sufficient details.

Item 1- Study objective (Study design)

	Tool	Content	Level
1	RELEVANT	Clearly stated research question?	R1
3	MMAT	Are there clear research questions?	R1
4	CASP	Did the study address a clearly focused issue?	R1
5	SURE	Does the study address a clearly focused question/hypothesis?	R1
8	CER-CI	Were the study hypotheses or goals prespecified a priori?	M1
10	NIH	Was the research question or objective in this paper clearly stated?	R1
11	Weightman et al. 2004	Does the paper address a clearly focused issue?	R1
14	SIGN	The study addresses an appropriate and clearly focused question.	R1 & M1
15	Montreal	What is the research question?	R1
16	STROBE	State specific objectives, including any prespecified hypotheses.	R1
17	TREND	Specific objectives and hypotheses.	R1
19	MINORS	A clearly stated aim: the question addressed should be precise and relevant in the light of available literature	R2
25	Blagojevic et al. 2010	Clearly defined and appropriate study objective.	R1 & M1
26	Genaidy et al. 2007	Is the hypothesis/aim/objective of the study clearly described?	R2
27	Harm	Is there a clearly focused question? Consider patients, exposure, and outcome.	R2
28	Tseng et al. 2009	Specific objectives or hypotheses stated (i.e. broadly outlined method for comparison indicated)?	R2
29	NHS Wales	Does the paper address a clearly focused issue, in terms of aims of the investigation, setting (location and dates), the population studied, and the variables measured?	R2
31	ANDQ	Was the research question clearly stated?	R1
32	NICE	The study addresses an appropriate and clearly focused question.	R1 & M1
33	IHE quality appraisal	Is the hypothesis/aim/objective of the study clearly stated in the abstract, introduction or methods section?	R2
34	AXIS	Were the aims/objectives of the study clear?	R1
36	Pluye et al. 2009	Qualitative objective or question.	R1
37	Heller et al. 2008	Is the research question and/or hypothesis stated clearly?	R1
39	Faillie et al. 2017	Are study objectives clearly specified and appropriate?	R1 & M1
41	Handu et al. 2016	Was the research question clearly stated?	R1
44	ISPE	A statement of research objectives, specific aims, and rationale.	R2

		Research objectives describe the knowledge or information to be gained from the study. Specific aims list key exposures and outcomes of interest, and any hypotheses to be evaluated. The protocol should distinguish between a limited number of a priori research hypotheses and hypotheses that are generated based on knowledge of the source data. The rationale explains how achievement of the specific aims will further the research objectives. The research question may be phrased by using the PICOT template; population, intervention, comparator, outcome, and timing.	
45	ENCePP	The objective(s) of the study? Which hypothesis(-es) is (are) to be tested?	R1
47	OSTEBA	Describe the objectives of the study. Is the study based on a clearly defined research question?	R1

Item 2 - Protocol (Study design)

	Tool	**Content**	**Level**
1	RELEVANT	Evidence of a priori design, e.g. protocol registration in a dedicated website.	R1
5	SURE	Was a trial protocol published? Was a protocol published in a journal or clinical trial registry before participants were recruited? If a protocol is available, are the outcomes reported in the paper listed in the protocol?	R2
7	ROBINS-I	Specify the review question: Participants; Experimental intervention; Comparator; Outcomes. List the confounding domains relevant to all or most studies. List co-interventions that could be different between intervention groups and that could impact on outcomes.	R2
8	CER-CI	Was there evidence that a formal study protocol including an analysis plan was specified before executing the study? (Refer to the original tool for more details)	R2
13	RTI-Item Bank	Did execution of the study vary from the intervention protocol proposed by the investigators and therefore compromise the conclusions of the study? Consider intensity, duration, frequency, route, setting, and timing of intervention/exposures.	M2
17	TREND	Description of protocol deviations from study as planned, along with reasons.	R2
19	MINORS	Prospective collection of data: data were collected according to a protocol established before the beginning of the study.	M2
31	ANDQ	Were protocols described for all regimens studied?	R1
35	AHRQ	Develop protocol: • Specify risk-of-bias categories (including sources of potential confounding for nonrandomized studies) and criteria and explain their inclusion; • Select and justify choice of specific risk-of-bias rating tool(s), including validity of selected tools (use risk-of-bias assessment tools that can identify potential risk-of-bias categories specific to the content area and study design) ;	R2

		• Explain how individual risk-of-bias categories (or items from a tool) will be presented or summarized (e.g., individually in tables, incorporated in sensitivity analysis, combined in an algorithm to obtain low, moderate, high, or unclear risk of bias for individual outcomes) ; • Explain how inconsistencies between pairs of risk-of-bias reviewers will be.	
43	Young et al. 2009	Deviations from the planned protocol can affect the validity or relevance of a study. (Refer to the original tool for more details)	M2
44	ISPE	Each study should have a written protocol. A protocol should be drafted as one of the first steps in any research project, and the protocol should be amended or updated as needed throughout the course of the study. (Refer to the original tool for more details)	R1
45	ENCEPP	The study protocol should also explain how the results will be interpreted, avoiding misuse of p-values and statistical significance.	R1
49	RECORD	Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.	R1

Item 3 – Selection of study design (Study design)

	Tool	**Content**	**Level**
4	CASP	Did the authors use an appropriate method to answer their question?	M1
11	Weightman et al. 2004	Has an acceptable method been chosen (e.g. before-and after study)? Is the choice of study method appropriate?	M2
13	RTI Item Bank	Is the study design prospective, retrospective, or mixed? Prospective design requires that the outcome has not occurred at the time the study is initiated and information is collected over time to assess relationships with the outcome (and includes nested Case-control studies). Mixed design includes Case-control or cohort studies in which one group is studied prospectively and the other retrospectively. A retrospective design analyzes data from past records. The question is not applicable to cross-sectional studies.	M1
14	SIGN	Is the paper really a Case-control study? If in doubt, check the study design algorithm available from SIGN and make sure you have the correct checklist.	M1
15	Montreal	What is the study type? Is the study type appropriate to the research question? If not, how useful are the results produces by this type of study?	M1
16	STROBE	Indicate the study's design with a commonly used term in the title or the abstract. Present key elements of study design early in the paper.	R2
22	Thomas et al. 2004	Indicate the study design 1 Randomized controlled trial 2 Controlled clinical trial 3 Cohort analytic (two group pre + post) 4 Case-control 5 Cohort (one group pre + post (before and after))	R2 & M1

		6 Interrupted time series 7 Other specify _____ 8 Can't tell (Refer to the original tool for more details)	
25	Blagojevic et al. 2010	Prospective study design.	M1
26	Genaidy et al. 2007	Is the study design clearly described?	R1
29	NHS Wales	Is the choice of study method appropriate to the study question? Is the study design and/or execution flawed to the extent that the results are unreliable?	M2
34	AXIS	Was the study design appropriate for the stated aim(s)?	M1
35	AHRQ	Determine study design of each (individual) study.	R1
36	Pluye et al. 2009	Appropriate qualitative approach or design or method.	M1
37	Heller et al. 2008	What is the study type? Is the study type appropriate for the research question? Is there a comparison group?	M1
39	Faillie et al. 2017	Is study design clearly specified and appropriate?	R1 & M1
40	IPM-QRBNR	Ranking different study designs on their strengths (points): Case report (0); Retrospective cohort (1); Prospective cohort (2); Prospective Case control (3); Prospective controlled, nonrandomized (4).	M1
42	Viswanathan et al. 2018	Determine study design of each (individual) study.	R1
43	Young et al. 2009	Was the study design appropriate for the research question?	M1
44	ISPE	The overall research design and reasons for choosing the proposed study design. Research designs include, for example, case–control, cohort, cross-sectional, nested case–control, self-controlled, randomized trials or hybrid designs. Any feasibility or pilot work that informed the choice of design should be described here.	R1
45	ENCEPP	Is the study design described (e.g. cohort, Case-control, cross-sectional, other design) ?	R1
48	Kennedy et al. 2019	If the study includes a cohort that was followed over time and included multiple assessments with the same people, this criterion is met. If the study did not conduct multiple assessments with a cohort of individuals over time, this criterion is not met. For example, a study that used a serial cross-sectional design with different individuals (even if they are from the same population) completing the assessments would not be considering as having a cohort design. Pre-post intervention outcome data is included in the risk of bias assessment, as it is common for studies to only assess outcome measures in the post-intervention catchments, especially for post hoc analyses and secondary study aims.	M2

		If the study presents data from both before (baseline) and after the intervention, this criterion is met. If data are only presented post-intervention, this criterion is not met.	
49	RECORD	Present key elements of study design early in the paper. Include details of the specific study design (and its features) and report the use of multiple designs if used. The use of a diagram(s) is recommended to illustrate key aspects of the study design(s), including exposure, washout, lag and observation periods, and covariate definitions as relevant.	R2

Item 4 - Sample size/Power calculation (Study design)

	Tool	**Content**	**Level**
1	RELEVANT	Sample size/Power pre-specified.	R1
4	CASP	Was there a power calculation? Was there a sufficient number of cases selected? Was there a sufficient number of controls selected?	R2
5	SURE	Was the sample size sufficient? Were there enough participants? Was there a power calculation? If yes, for which outcome? Were there sufficient participants?	R2&M1
8	CER-CI	Were sample size and statistical power to detect difference addressed?	R1
10	NIH	Was a sample size justification, power description, or variance and effect estimates provided?	R2
13	RTI item bank	Was the sample size sufficiently large to detect a clinically significant difference of 5% or more between groups in at least one primary outcome measure? Specify a different percent, if clinically relevant for each outcome of interest. Reviewers whose evaluation of quality is limited to considerations of systematic error or risk of bias (not random error/precision) need not include this question. Reviewers who include both precision and systematic error in their evaluation of quality but rely on meta-analysis for pooled estimates need not include this question. Who choose to include considerations of precision in their assessment may include the question, but should be aware of the need for collaboration between clinical and statistical expertise in determining the threshold for a clinically adequate sample size.	M2
14	SIGN	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	R1
15	Montreal	Was the sample size adequate to detect a clinically/socially significant result?	R1
16	STROBE	Explain how the study size was arrived at; Cross-sectional study—If applicable, describe analytical methods taking account of sampling strategy	R1
17	TREND	How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules.	R1 & M1

19	MINORS	Prospective calculation of the study size: information of the size of detectable difference of interest with a calculation of 95% confidence interval, according to the expected incidence of the outcome event, and information about the level for statistical significance and estimates of power when comparing the outcomes.	M2
20	GRADE	What is the magnitude of the median sample size? • High (e.g. 300 participants); • Intermediate (e.g. 100-300 participants); • Low (e.g. <100 participants).	R1
21	Rangel et al. 2003	Can the number of surgeons who participated in the study be determined?	R1
23	Atluri et al. 2008	Sample size justification; Power calculation provided.	R1
25	Blagojevic et al. 2010	Sample size calculation given or about 20 subjects per variable included in multivariate analysis.	R1
26	Genaidy et al. 2007	Are sample size calculations performed and reported? Yes - Calculations are performed, and, all details are reported for effect size, type I or II errors and number of confounders; Partial - Somewhat described. Calculations are performed, and, not all details are reported; No - Not described N . No mention of any calculations.	R2
28	Tseng et al. 2009	Calculation to justify sample size?	M1
29	NHS Wales	Is the population studied appropriate? • Was the sample representative of its target population? • How was the sample selected – random, stratified? If appropriate, was a power calculation made?	M2
31	ANDQ	If negative findings, was a power calculation reported to address type 2 error?	M2
34	AXIS	Was the sample size justified?	M1
36	Pluye et al. 2009	Appropriate sampling and sample.	M1
37	Heller et al. 2008	Was sample size/power calculated and appropriate?	R1 & M1
40	IPM-QRBNR	Less than 100 participants without appropriate sample size determination (0); At least 100 participants in the study without appropriate sample size determination (1); Sample size calculation with less than 50 patients in each group (2); Appropriate sample size calculation with at least 50 patients in each group (3); Appropriate sample size calculation with 100 patients in each group (4).	M1
44	ISPE	Some justification should be given to support that the necessary study size is actually attainable from the given data source or design. (Refer to the original tool for more details)	M2

Item 5 - Eligibility criteria (Study design)

	Tool	Content	Level
1	RELEVANT	Population justified;	R2 & M1

		Flow chart explaining all exclusions and individuals screened or selected at each stage of defining the final sample.	
2	GATE	Eligible population recruitment process.	R1
4	CASP	Were the cases recruited in an acceptable way? Were the controls selected in an acceptable way? Was the cohort recruited in an acceptable way?	M1
5	SURE	Population/Problem? Can you identify the setting & eligibility criteria?	R1
6	JBI	Were there clear criteria for inclusion in the case series? Did the case series have complete inclusion of participants? Were the same criteria used for identification of cases and controls? Were the two groups similar and recruited from the same population? (Refer to the original tool for more details)	R2
10	NIH	Were all the subjects selected or requited from the same or similar populations (including the same time period)? Were inclusion and exclusion criteria for being in the study prespecified and applied uniformly to all participants?	M2
11	Weightman et al. 2004	Are the inclusion/exclusion criteria given?	R1
13	RTI-item bank	Are critical inclusion/exclusion criteria clearly stated (does not require the reader to infer)? Provide direction to abstractors by listing individual criteria of a priori significance and minimal requirements for criteria to be considered “clearly stated.” Include this question to identify specific inclusion/exclusion criteria that should be consistently recorded across studies Use “Partially” if only some criteria are stated or if some criteria are not clearly stated. Note that studies may describe inclusion criteria alone (i.e., include x), exclusion criteria (i.e., do not include x), or a combination of inclusion and exclusion criteria. Are the inclusion/exclusion criteria measured using valid and reliable measures? Separately specify each criterion that abstractors should consider based on its relevance to study bias. It is unlikely that all criteria will need to be evaluated in relation to this question. Provide direction to abstractors on valid and reliable measurement of each criterion that is to be considered. For example, prior exposure or disease status is a frequent inclusion/exclusion criterion, particularly in inception cohorts. Subjective measures based on self-report tend to have lower reliability and validity than objective measures such as clinical reports and lab findings. Replicate question to evaluate each individual inclusion/exclusion criterion. Did the study apply inclusion/exclusion criteria uniformly to all comparison groups/arms of the study? Drop question if not relevant to entire body of evidence (e.g., all case-series, singlearm studies).	R2&M2
14	SIGN	The same exclusion criteria are used for both cases and controls.	M2

15	Montreal	What are the sampling frame and sampling method? Is there selection bias? Does this selection bias threaten the external validity of the study?	M2
16	STROBE	Cohort study—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up. Case-control study—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls. Cross-sectional study—Give the eligibility criteria, and the sources and methods of selection of participants. Cohort study—For matched studies, give matching criteria and number of exposed and unexposed Case-control study—For matched studies, give matching criteria and the number of controls per case	R2
17	TREND	Eligibility criteria for participants, including criteria at different levels in recruitment/sampling plan (e.g., cities, clinics, subjects).	R2
19	MINORS	Inclusion of consecutive patients : all patients potentially fit for inclusion (satisfying the criteria for inclusion) have been included in the study during the study period (no exclusion or details about the reasons for exclusion).	R2
21	Rangel et al. 2003	Are selection and/or exclusion criteria for cases clearly stated?	R1
25	Blagojevic et al. 2010	Inclusion and exclusion criteria are clear and appropriate. Representative sample e.g., general population sample should not exclude subgroups.	R1 & M1
26	Genaidy et al. 2007	Are the eligibility criteria for subject selection clearly described? Yes – Clearly described: Cohort, Intervention, and Cross-sectional designs: ~ Inclusion and/or exclusion criteria are clearly described in few sentences; Case-control designs: ~ A case-definition is clearly described in few sentences; Proportional designs: ~ Inclusion and/or exclusion criteria or case definitions are clearly described in few sentences. Partial – Somewhat described . Criteria are not clearly described. No – Not described.	R2
28	Tseng et al. 2009	Are selection and/or exclusion criteria for cases clearly stated?	R1
30	NOS	Selection 1) Representativeness of the exposed cohort a) truly representative of the average _____ (describe) in the community ¯ b) somewhat representative of the average _____ in the community ¯ c) selected group of users eg nurses, volunteers d) no description of the derivation of the cohort. 2) Selection of the non exposed cohort a) drawn from the same community as the exposed cohort ¯ b) drawn from a different source c) no description of the derivation of the non exposed cohort. 3) Ascertainment of exposure a) secure record (e.g. surgical records) ¯ b) structured interview ¯ c) written self report d) no description.	M2

		4) Demonstration that outcome of interest was not present at start of study. Comparability 1) Comparability of cohorts on the basis of the design or analysis a) study controls for _____ (select the most important factor) ¯ b) study controls for any additional factor ¯ (This criteria could be modified to indicate specific control for a second important factor.)	
31	ANDQ	Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? Were criteria applied equally to all study groups? Were health, demographics, and other characteristics of subjects described? Were the subjects/patients a representative sample of the relevant population?	R2
32	NICE	The same exclusion criteria are used for both cases and controls.	M1
33	IHE quality appraisal	Are the eligibility criteria (inclusion and exclusion criteria) to entry the study explicit and appropriate? Description of the eligibility criteria (inclusion and exclusion criteria).	R1 & M1
37	Heller et al. 2008	Are exclusion criteria appropriate?	M1
39	Faillie et al. 2017	Were inclusion and exclusion criteria implemented uniformly across study groups?	M1
40	IPM-QRBNR	A study's population is clinically relevant to assessing methodological quality: (1) studies including ≥ 200 patients with a large sample size (2) clearly identified mixed population (3) studies examining a specific disorder that has well defined limitations.	M1
41	Handu et al. 2016	Was the selection of study subjects/patients free from bias? Were inclusion/exclusion criteria specified (e.g. risk, point in disease progression, and diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? Were criteria applied equally to all study groups? Were the subjects/patients a representative sample of the relevant population?	R2 & M2
44	ISPE	The rationale for the inclusion and exclusion criteria and their impact on the number of subjects available for analysis should be described, if known.	R2
45	ENCePP	Does the protocol define how the study population will be sampled from the source population (e.g. event or inclusion/exclusion criteria) ?	R1
46	ISPE-ISPOR	Reporting on inclusion/exclusion criteria should include: Study entry date (SED), Person or episode level study entry, Sequencing of exclusions, Enrollment window (EW, Enrollment gap, Inclusion/Exclusion definition window, Codes, Frequency and temporality of codes,etc.	R2
47	OSTEBA	Was the participant selection method suitable?	M1

49	RECORD	Describe the study entry criteria and the order in which these criteria were applied to identify the study population. Specify whether only users with a specific indication were included and whether patients were allowed to enter the study population once or if multiple entries were permitted. (Refer to the original tool for more details)	R2

Item 6 – Intervention selection (Study design)

	Tool	Content	Level
4	CASP	Were the cases recruited in an acceptable way? Were the controls selected in an acceptable way?	M1
5	SURE	Were interventions (and comparisons) well described and appropriate? Aside from the intervention, were the groups treated equally? Was exposure to intervention and comparison adequate? Was contamination acceptably low?	R1 & M2
8	CER-CI	Are any relevant interventions missing? This question addresses whether the interventions analysed in the study include ones of interest to the decision maker and whether all relevant comparators have been considered. (Refer to the original tool for more details)	M2
9	GRACE	Was the study (or analysis) population restricted to new initiators of treatment or those starting a new course of treatment?	M2
10	NIH	For exposure that can vary in amount or level did the study examine different levels of exposure as related to the outcome (e.g., categories of exposure, exposure measured as continuous variable)?	M2
13	RTI Item Bank	Is the selection of the comparison group appropriate, after taking into account feasibility and ethical considerations? Provide instruction to the abstractor based on the type of study. Interventions with community components are likely to have contamination if all groups are drawn from the same community. Interventions without community components should select groups from the same source (e.g., community or hospital) to reduce baseline differences across groups. For Case-control studies, controls should represent the population from which cases arose; that is, controls should have met the case definition if they had the outcome.	M2
31	ANDQ	Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect?	M1
33	IHE	Were additional interventions (cointerventions) clearly reported in the study?	M1
48	Kennedy et al. 2019	If the study included a control and/or comparison arm in addition to the intervention arm, this criterion is met. If the study only had an intervention arm, this criterion is not met. Comparison group sociodemographic matching is assessed in multi-arm studies to determine if there are statistically significant differences in sociodemographic measures across arms at baseline	M2

		If the study arms are equivalent on sociodemographic characteristics, this criterion is met. If there are significant differences between one or more of the study arms on socio-demographic characteristics, this criterion is not met.	
47	RECORD	Use of any comparator groups should be outlined and justified.	R1

Item 7 - Intervention definition (Study design)

	Tool	**Content**	**Level**
1	RELEVANT	(If relevant), exposure (e.g. treatment) is clearly defined.	R1
5	SURE	Were interventions (and comparisons) well described and appropriate? Aside from the intervention, were the groups treated equally? Was exposure to intervention and comparison adequate? Was contamination acceptably low?	R1&M2
7	ROBINS-I	Were intervention groups clearly defined? Was the information used to define intervention groups recorded at the start of the intervention?	R1&M2
8	CER-CI	Was exposure defined and measured in a valid way?	M1
13	RTI Item Bank	What is the level of detail in describing the intervention or exposure? Specify which details need to be stated (e.g., intensity, duration, frequency, route, setting, and timing of intervention/exposure). For Case-control studies, consider whether the condition, timing, frequency, and setting of symptoms are provided in the case definition.	R2
16	STROBE	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	R2
17	TREND	Details of the interventions intended for each study condition and how and when they were actually administered, specifically including: Content: what was given? Exposure quantity and duration: how many sessions or episodes or events were intended to be delivered? How long were they intended to last?	R2
18	ACROBAT-NRSI	Is intervention status well defined?	M1
21	Rangel et al. 2003	Description of the intervention: Is the surgical technique adequately described? Is there any mention of an attempt to standardize operative technique? Is there any mention of an attempt to standardize perioperative care?	R2
23	Atluri et al. 2008	Clear definition of exposure.	R1
24	Bishop et al. 2010	A definition of CAM and/or a list of specific CAM therapies is provided to participants.	R1

26	Genaidy et al. 2007	Are all the exposure variables/intervention(s) clearly described?	R1
28	Tseng et al. 2009	Is the surgical technique/intervention adequately described (e.g. specifically referenced article)?	R2
30	NOS	Is the case definition adequate? Same method of ascertainment for cases and controls	R1
31	ANDQ	In RCT or other intervention trial, were protocols described for all regimens studied? In observational study, were interventions, study settings, and clinicians/provider described?	R1
32	NICE	Cases are clearly defined and differentiated from controls.	R1
33	IHE	Was the intervention clearly described in the study?	R1
37	Heller et al. 2008	Intervention features (for an intervention study): is the intervention described adequately?	R1
45	ENCePP	Exposure definitions can include simple dichotomous variables (e.g., ever vs. never exposed) or be more granular, including estimates of duration, exposure windows (e.g., current vs. past exposure) also referred to as risk periods, or dosage (e.g., current dosage, cumulative dosage over time).	R2
46	ISPE-ISPOR	The type of exposure that is captured or measured, e.g. drug versus procedure, new use, incident, prevalent, cumulative, time-varying.	R1
49	RECORD	RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.	R2

Item 8 - Outcome selection (Study design)

	Tool	**Content**	**Level**
8	CER-CI	Are the outcomes relevant? This question asks what outcomes are assessed in the study and whether the outcomes are meaningful to the patients the decision maker is concerned with.	R1 & M1
13	RTI Item Bank	Are outcomes assessed using valid and reliable measures, implemented consistently across all study participants? Primary outcomes should be identified for abstractors and if there is more than one, they may be listed separately. Also, identify any relevant secondary outcomes and harms. Subjective measures based on self-report tend to have lower reliability and validity than objective measures such as clinical reports and lab findings. Note for Case-control studies: consider whether the ascertainment of cases was independent of exposure.	M2
19	MINORS	Endpoints appropriate to the aim of the study: unambiguous explanation of the criteria used to evaluate the main outcome which should be in accordance with the question addressed by the study. Also, the endpoints should be assessed on an intention-to-treat basis.	M2
20	GRADE	Was an objective outcome used? Was the included outcome not a surrogate outcome?	M1

31	ANDQ	Were primary and secondary endpoints described and relevant to the question? Were nutrition measures appropriate to question and outcomes of concern? Were other factors accounted for (measured) that could affect outcomes?	R2
34	AXIS	Were the risk factor and outcome variables measured appropriate to the aims of the study?	M1
39	Faillie et al. 2017	Was the method for ascertaining the drug safety outcome adequately constructed and equal for all participants? (Refer to the original tool for more details)	M2
41	Handu et al. 2016	Were primary and secondary endpoints described and relevant to the question?	R1
45	ENCePP	Does the protocol specify the primary and secondary (if applicable) outcome(s) to be investigated? Does the protocol describe specific outcomes relevant for Health Technology Assessment? (e.g. HRQoL, QALYs, DALYS, health care services utilisation, burden of disease or treatment, compliance, disease management)?	R2

Item 9 - Outcome definition (Study design)

	Tool	**Content**	**Level**
1	RELEVANT	Primary outcomes defined?	R1
8	CER-CI	Were the primary outcomes defined and measured in a valid way?	M1
13	RTI Item Bank	Are the important outcomes pre-specified by the researchers? Do not consider harms in answering this question unless they should have been pre-specified. This question can be asked for all outcomes together or replicated for each event. Each adverse event of interest should be specified for abstractors. Relevant source information includes all study data, including what may have been established in relation to an initial randomized controlled trial. Drop question if not relevant (e.g., primary outcome for Case-control studies).	M2
14	SIGN	The outcomes are clearly defined.	M1
16	STROBE	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable (Refer to the original tool for more details)	R2
17	TREND	Clearly defined primary and secondary outcome measures.	R1
23	Atluri et al. 2008	Primary/secondary outcomes clearly defined.	R1
26	Genaidy et al. 2007	Are the main outcomes clearly described? (Refer to the original tool for more details)	M2
28	Tseng et al. 2009	Is there a clearly defined single primary outcome?	R1
31	ANDQ	Were primary and secondary endpoints described and relevant to the question? Were outcomes clearly defined and the measurements valid and reliable	M1
32	NICE	The study used a precise definition of outcome A valid and reliable method was used to determine the outcome	M2

		The outcome under study should be well defined and it should be clear how the investigators determined whether participants experienced, or did not experience, the outcome. The same methods for defining and measuring outcomes should be used for all participants in the study. Often there may be more than one way of measuring an outcome (for example, physical or laboratory tests, questionnaire, reporting of symptoms). The method of measurement should be valid (that is, it measures what it claims to measure) and reliable (that is, it measures something consistently).	
37	Heller et al. 2008	What are the outcome factors?	R1
39	Faillie et al. 2017	Is the definition of the drug safety outcome clearly stated? Clear / standardized definition of the drug safety outcome (e.g. diagnostic codes, clinical and laboratory data).	R2
45	ENCePP	Does the protocol describe how the outcomes are defined and measured? Outcomes? (e.g. clinical records, laboratory markers or values, claims data, self-report, patient interview including scales and questionnaires, vital statistics)	R2
46	ISPE-ISPOR	Reporting on outcome definition should include: date of an event occurrence, codes, frequency and temporality of codes, diagnosis position, care setting validation.	R2
47	OSTEBA	Is the test used for comparison adequately defined? Are the outcomes of interest adequately defined? Please, note it down.	R1
49	RECORD	A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.	R2

Item 10 - Ethical approval (Study design)

	Tool	**Content**	**Level**
5	SURE	Was ethical approval sought and received? Do the authors report this?	R1 & M1
34	AXIS	Was ethical approval or consent of participants attained?	R1
37	Heller et al. 2008	Has the impact on the population been presented? Yes/no Is the study ethical?	R1 & M1
45	ENCePP	Have requirements of Ethics Committee/ Institutional Review Board been described? Has any outcome of an ethical review procedure been addressed?	R2

Item 11 - Conflict of interest (Study design)

	Tool	**Content**	**Level**
1	RELEVANT	Potential conflicts of interest, including study funding, are stated.	R1
5	SURE	Is any sponsorship/conflict of interest reported?	R1
8	CER-CI	Were there any potential conflicts of interest? If there were potential conflicts of interest, were steps taken to address these?	M2

13	RTI Item Bank	Is the source of funding identified? [PI: The relevance of this question will depend upon the topic. This question may be modified to identify particular sources of funding (e.g., industry, government, university, or foundation funding).]	M2
16	STROBE	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based.	R2
20	GRADE	There was no industry influence on studies included in the review?	M2
31	ANDQ	Is bias due to study's funding or sponsorship. Were sources of funding and investigators' affiliations described? 10.2 Was there no apparent conflict of interest? Are biases and study limitations identified and discussed?	M2
32	NICE	How was the study funded?	R1
33	IHE	Are both competing interest and source of support for the study reported?	R2
34	AXIS	Were there any funding sources or conflicts of interest that may affect the authors' interpretation of the results?	R1
39	Faillie et al. 2017	Were the conflict of interest or sources of funding clearly acknowledged? Potential sources of support are acknowledged. Does the study appear free of conflicts of interest susceptible to have influenced design, analysis, or reporting (selective reporting of outcome or analysis)?	M2
40	IPM-QRBNR	Include industry employees with or without proper disclosure.	R2
41	Handu et al. 2016	Is bias due to study's funding or sponsorship unlikely? Were sources of funding and investigators' affiliations described? Was there no apparent conflict of interest?	M2
43	Kennedy et al. 2019	Are there any conflicts of interest?	R1
47	OSTEBA	Is the existence or absence of conflicts of interest properly described? When possible, specify the financial source.	R1&M1

Item 12 - Data source (Data quality)

	Tool	**Content**	**Level**
1	REVELANT	The data source (or database), as described, contains adequate exposures (if relevant) and outcome variables to answer the research question.	M1
8	CER-CI	Were the sources, criteria, and methods for selecting participants appropriate to address the study questions/hypotheses? Were the data sources sufficient to support the study?	M1
9	GRACE	Were the primary outcomes adequately recorded for the study purpose (e.g., available in sufficient detail through data sources)?	R1&M1

16	STROBE	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group.	R2
19	MINORS	Prospective collection of data: data were collected according to a protocol established before the beginning of the study.	M2
33	IHE	Case series collected in more than one centre (multicentre study).	M1
39	Faillie et al. 2017	Secondary databases studies: Are the characteristics of the database clearly described?	R1
44	ISPE	Data sources might include, for example, questionnaires, hospital discharge files, abstracts of primary clinical records, clinical databases, electronic medical records, ad hoc data collection, administrative records such as eligibility files, prescription drug files, biological measurements, exposure/ work history record reviews, or exposure/ disease registries. If the study uses secondary data, the name of the data source should be included (e.g., Medicare, CPRD, and MarketScan). Use validated instruments and measures whenever such exist and describe the validation method and summarize what is known about the completeness and validity of those instruments and measures. If data collection methods or instruments will be tested in a pilot study, plans for the pilot study should be described. Any procedures to be used to validate diagnosis should be described.	R2
45	ENCePP	Does the protocol describe the data source(s) used in the study for the ascertainment of: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included.	R2
46	ISPE-ISPOR	Reporting on data source should include: Data provider *Data source name and name of organization that provided data.* Data extraction date (DED) *The date (or version number) when data were extracted from the dynamic raw transactional data stream (e.g. date that the data were cut for research use by the vendor).* Data sampling *The search/extraction criteria applied if the source data accessible to the researcher is a subset of the data available from the vendor.* Source data range (SDR) *The calendar time range of data used for the*	R2

		study. Note that the implemented study may use only a subset of the available data. Type of data *The domains of information available in the source data, e.g. administrative, electronic health records, inpatient versus outpatient capture, primary vs secondary care, pharmacy, lab, registry.* Data linkage, other supplemental data *Data linkage or supplemental data such as chart reviews or survey data not typically available with license for healthcare database.* Data cleaning *Transformations to the data fields to handle missing, out of range values or logical inconsistencies. This may be at the data source level or the decisions can be made on a project specific basis.* Data model conversion *Format of the data, including description of decisions used to convert data to fit a Common Data Model (CDM).*	
49	RECORD	If linkage between databases was conducted for the study, this should be clearly stated in the title or abstract.Specify the data sources from which drug exposure information for individuals was obtained. State whether the study included person level, institutional level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.	R2

Item 13 - Patient recruitment (Data quality)

	Tool	**Content**	**Level**
1	RELEVANT	Population defined. Population justified.	R1 & M1
2	GATE	Recruitment of participants 'who are the findings applicable to?'	M1
3	MMAT	Are the participants representative of the target population?	M1

4	CASP	Were the cases recruited in an acceptable way? were the controls representative of the defined population (geographically and/or temporally)	M1
6	JBI	Were the groups/participants free of the outcome at the start of the study (or at the moment of exposure)?	M2
7	ROBINS-I	*2.1. Was selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of intervention?* *If N/PN to 2.1: go to 2.4* *Y / PY / PN / N / NI* *2.2. If Y/PY to 2.1: Were the post-intervention variables that influenced selection likely to be associated with intervention?* *2.3 If Y/PY to 2.2: Were the post-intervention variables that influenced selection likely to be influenced by the outcome* *2.4. Do start of follow-up and start of intervention coincide for most participants?* *2.5. If Y/PY to 2.2 and 2.3, or N/PN to 2.4: Were adjustment techniques used that are likely to correct for the presence of selection biases?*	M2
10	NIH	Was the study population clearly specified and defined?	R1
13	RTI Item Bank	Was the strategy for recruiting participants into the study the same across study groups/arms of the study? This question is likely to be more relevant for prospective or mixed designs than retrospective designs. Drop question if not relevant to entire body of evidence (e.g., all studies generally have only one arm).	M2
14	SIGN	The cases and controls are taken from comparable populations.	M1
18	ACROBAT-NRSI	Were the controls sampled from the population that gave rise to the cases, or using another method that avoids selection bias?	M2
21	Rangel et al. 2003	Description and definition of participating surgeons/institutions: Can the number of participating centers be determined? Can the practice type of participating centers be determined? Can the number of surgeons who participated in the study be determined? Can the reader determine where the authors are on the learning curve for the reported procedure? Is the timeline when all cases were performed clearly stated? Was the patient population from which the cases were selected from adequately described?	R2
22	Thomas et al. 2004	Are the individuals selected to participate in the study likely to be representative of the target population?	M1
23	Atluri et al. 2008	Subjects similar to populations in which the test would be used and with a similar spectrum of disease.	M1
25	Blagojevic et al. 2010	Representative sample e.g., general population sample should not exclude subgroups.	M1
26	Genaidy et al. 2007	Is the source of subject population (including sampling frame) clearly described? Yes – Clearly described:	R2

		Details are clearly described in few sentences. This may or may not be supplemented with a flowchart. Example: ~ The study population was workers identified through the 'International Register of Workers to Phenoxy Herbicides and their Contaminant', which was set up by an international and a US group. This consisted of 20 separate cohorts representing different employers, workplace and countries involving in total 18 390 workers (16 683 male, 1527 female) from ten countries. The derivation of the study participants is also demonstrated in a flowchart. Partial – Somewhat described: Details are not clearly described. No – Not described Are newly incident cases taken into account?	
27	Harm	Were there clearly defined groups of patients, similar in all important ways other than exposure to the treatment or other causes?	R2
28	Tseng et al. 2009	Was the patient population from which the cases were selected from adequately described or identified (e.g. geographically?	R2
29	NHS Wales	Is the population studied appropriate? Was the sample representative of its target population? How was the sample selected – random, stratified?	M1
30	NOS	1) Is the case definition adequate? a) yes, with independent validation ¯ b) yes, eg record linkage or based on self reports c) no description 2) Representativeness of the cases. a) consecutive or obviously representative series of cases ¯ b) potential for selection biases or not stated 3) Selection of Controls a) community controls ¯ b) hospital controls c) no description 4) Definition of Controls a) no history of disease (endpoint) ¯ b) no description of source.	R2
31	ANDQ	Was the selection of study subjects/patients free from bias? Were study groups comparable?	M1
32	NICE	- The cases and controls are taken from comparable populations.	M1
34	AXIS	Was the target/reference population clearly defined? (Is it clear who the research was about?) Was the sample frame taken from an appropriate population base so that it closely represented the target/reference population under investigation? Was the selection process likely to select subjects/participants that were representative of the target/reference population under investigation?	R1 & M2
37	Heller et al. 2008	Are the sampling frame and sampling method appropriate? Is the sample representative of the population being studied? Can you generalize from the population being studied (External validity) ? Is this sample relevant to my population?	M2

		In a case–control study, are the controls representative of the source population for the cases, are exposures and population representative of your population of interest?	
39	Faillie et al. 2017	Are all the subjects recruited from the same source population? Is the origin of controls clearly specified?	R1 & M1
40	IPM-QRBNR	Method of assigning patients Case report/case series or selective assignment based on outcomes or retrospective evaluation based on clinical criteria (1) Prospective study with inclusion without specific criteria (2) Retrospective method with inclusion of all participants or random selection of retrospective data (3) Prospective, well-defined assignment of methodology and inclusion criteria (4)	M2
44	ISPE	If any sampling from a defined population is undertaken, description of the population and details of sampling methods should be provided.	R1
45	ENCePP	Is the source population described? Is the planned study population defined in terms of: 4.2.1 Study time period 4.2.2 Age and sex 4.2.3 Country of origin 4.2.4 Disease/indication 4.2.5 Duration of follow-up Does the protocol address selection bias (e.g. healthy user/adherer bias) ?	R2
47	OSTEBA	Describe the location and study period. Is the target population of the study adequately defined? Please, note it down.	R1

Item 14 - Participation rate (Data quality)

	Tool	**Content**	**Level**
10	NIH	Was the participation rate of eligible persons at least 50%?	M2
14	SIGN	What percentage of each group (cases and controls) participated in the study? The study indicates how many of the people asked to take part did so, in each of the groups being studied.	R1
16	STROBE	Report numbers of individuals at each stage of study—e.g. numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed. Give reasons for non-participation at each stage.	R2
17	TREND	Flow of participants through each stage of the study: enrollment, assignment, allocation and intervention exposure, follow-up, analysis (a diagram is strongly recommended). Enrollment: the numbers of participants screened for eligibility, found to be eligible or not eligible, declined to be enrolled, and enrolled in the study; Assignment: the numbers of participants assigned to a study condition;	R2

		Allocation and intervention exposure: the number of participants assigned to each study condition and the number of participants who received each intervention.	
20	GRADE	Were more than 80% of participants enrolled in trials included in the analysis (i.e. no potential reporting bias)?	M1
22	Thomas et al. 2004	What percentage of selected individuals agreed to participate? 1 80 - 100% agreement 2 60 – 79% agreement 3 less than 60% agreement 4 Not applicable 5 Can't tell	M1
24	Bishop et al. 2010	What is the response rate? Number of participants in the study? Number of people invited to take part)?	R1
25	Blagojevic et al. 2010	Baseline response is about 70%?	M1
26	Genaidy et al. 2007	Are the participation rate(s) reported? Are ascertainment's of record availability described? Is the participation rate adequate? Is the ascertainment of record availability adequate? (Refer to the original tool for more details)	R2 & M2
29	NHS Wales	Did the study achieve a good response rate?	M1
30	NOS	Non-Response rate: a) same rate for both groups ¯ b) non respondents described c) rate different and no designation.	M1
32	NICE	What was the participation rate for each group (cases and controls)? Differences between the eligible population and the study participants are important because they may influence the validity of the study. A participation rate can be calculated by dividing the number of study participants by the number of people who are eligible to participate.	R2 & M2
34	AXIS	Does the response rate raise concerns about non-response bias?	M1
37	Heller et al. 2008	In a cross-sectional study, is the item-specific response rate adequate?	M1
39	Faillie et al. 2017	Are the number of participants clearly reported throughout the study?	R1
40	IPM-QRBNR	Description of Drop Out Rate No description despite reporting of incomplete data or more than 30% withdrawal Less than 30% withdrawal in one year in any group Less than 40% withdrawal at 2 years in any group.	R2
41	Handu et al. 2016	Was the number, characteristics of withdrawals (ie, dropouts, lost to follow-up, attrition rate) and/or response rate (cross-sectional studies) described for each group (Follow-up goal for a strong study is 80%)?	R2 & M2
49	RECORD	Describe in detail the selection of the persons included in the study (i.e., study population selection), including filtering based on data quality, data availability, and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram.	R2

Item 15 - Baseline characteristics (Data quality)

	Tool	**Content**	**Level**
3	MMAT	Are the groups comparable at baseline?	M1
6	JBI	Was there clear reporting of the demographics of the participants in the study? Was there clear reporting of clinical information of the participants? Were the two groups similar and recruited from the same population?	R1

8	CER-CI	Were the study groups selected so that comparison groups would be sufficiently similar to each other (e.g., either by restriction or recruitment based on the same indications for treatment)? (Refer to the original tool for more details)	M2
13	RTI Item Bank	Is the selection of the comparison group appropriate, after taking into account feasibility and ethical considerations. Provide instruction to the abstractor based on the type of study. Interventions with community components are likely to have contamination if all groups are drawn from the same community. Interventions without community components should select groups from the same source (e.g., community or hospital) to reduce baseline differences across groups. For Case-control studies, controls should represent the population from which cases arose; that is, controls should have met the case definition if they had the outcome.	M2
14	SIGN	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	M1
17	TREND	Baseline demographic and clinical characteristics of participants in each study condition. Baseline characteristics for each study condition. Example (baseline characteristics specific to HIV prevention research): HIV serostatus disease prevention research and HIV testing behavior. Baseline comparisons of those lost to follow-up and those retained, overall and by study condition. Comparison between study population at baseline and target population of interest.	R2
19	MINORS	Baseline equivalence of groups : the groups should be similar regarding the criteria other than the studied endpoints. Absence of confounding factors that could bias the interpretation of the results.	M1
23	Atluri et al. 2008	Comparability of groups at baseline with regard to disease status and prognostic factors. Study groups comparable to non-participants with regard to confounding factors.	M1
26	Genaidy et al. 2007	Are the characteristics of study participants described? (Refer to the original tool for more details)	R2
27	Harm	Were there clearly defined groups of patients, similar in all important ways other than exposure to the treatment or other causes?	M1
31	ANDQ	Were health, demographics, and other characteristics of subjects described? Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline?	R2
32	NICE	What are the main characteristics of the study population? The groups were comparable at baseline, including all major confounding and prognostic factors	R1 & M2
33	IHE	Participants entering the study at a similar point in their disease progression? Are the characteristics of the participants included in the study described?	R1 & M1
35	AHRQ	Characteristics such as disease severity or comorbidity are unlikely to influence the intervention and outcome) or appropriate analysis methods are used to adjust for important baseline confounding?	M2

38	CARE	Demographic information (e.g. age, gender, ethnicity, occupation) ,main symptoms of the patient (his or her chief complaints), medical, family, and psychosocial history, including diet, lifestyle, and genetic information , comorbidities including past interventions and their outcomes.	R2
39	Faillie et al. 2017	Are baseline characteristics and prognostic factors comparable between different groups?	M1
40	IPM-QRBNR	Similarity of Groups at Baseline for Important Prognostic Indicators.	M1
41	Handu et al. 2016	Were distribution of disease status, prognostic factors, and other factors (eg, demographic characteristics) similar across study groups at baseline?	M2
47	OSTEBA	Note the number and characteristics of the participants down.	R1
47	RECORD	Give characteristics of study participants (e.g., demographic, clinical, and social) and information on exposures and potential confounders. (b) Indicate the number of participants with missing data for each variable of interest. (c) Cohort study: summarize follow-up time (e.g., average and total amount).	R2

Item 16 – Intervention measurement (Data quality)

	Tool	**Content**	**Level**
2	GATE- GATE	Were exposures & outcomes well Measured?' were they measured Objectively?	M1
3	MMAT	Are measurements appropriate regarding both the outcome and intervention (or exposure)? Are the measurements appropriate?	M1
4	CASP	Was the exposure accurately measured to minimize bias? Did they use subjective or objective measurements? Do the measurements truly reflect what you want them to (have they been validated)?	M2
6	JBI	Was exposure measured in a standard, valid and reliable way? Was exposure measured in the same way for cases and controls? The study should clearly describe the method of measurement of exposure. Was the condition measured in a standard, reliable way for all participants included in the case series? Were the exposures measured similarly to assign people to both exposed and unexposed groups? (Refer to the original tool for more details)	R1&M2
7	ROBINS-I	Were there deviations from the intended intervention beyond what would be expected in usual practice? If Y/PY to 4.1: Were these deviations from intended intervention unbalanced between groups and likely to have affected the outcome? If Y/PY to 4.1: Were these deviations from intended intervention unbalanced between groups and likely to have affected the outcome?	M2
8	CER-CI	Was exposure defined and measured in a valid way?	M1
10	NIH	For the analyses in this paper, were the exposure(s) of interest measured prior to the outcome(s) being measured? Was the exposure(s) assessed more than once over time?	M2

		Were the exposure measures (independent variables) clearly defined, valid, reliable, and implemented consistently across all study participants?	
13	RTI Item Bank	Are interventions/exposures assessed using valid and reliable measures, implemented consistently across all study participants? Important measures may be listed separately. When subjective or objective measures could be collected, subjective measures based on self report may be considered as being less reliable and valid than objective measures such as clinical reports and lab findings. Replicate question when needed.	M2
14	SIGN	Exposure status is measured in a standard, valid and reliable way.	M1
23	Atluri et al. 2008	Measurement method standard, valid and reliable.	M1
26	Genaidy et al. 2007	Are the exposure variables reliable? (Refer to the original tool for more details)	M2
27	Harm	Were treatments/exposures and clinical outcomes measured in the same way for both groups?	M1
31	ANDQ	Was the amount of exposure and, if relevant, subject/patient compliance measured?	M1
32	NICE	Exposure status is measured in a standard, valid and reliable way?	M1
36	Pluye et al. 2009	Justification of measurements (validity and standards).	M1
37	Heller et al. 2008	Observations/risk factors: how are the exposures measured?	M1
39	Faillie et al. 2017	Cohort, Case-control studies: Was the method for ascertaining drug use and drug use duration adequately constructed, and equal for all participants?	M2
41	Handu et al. 2016	Was the amount of exposure and, if relevant, subject/patient compliance measured?	M1
45	ENCePP	Does the protocol address the validity of the exposure measurement (e.g. precision, accuracy, use of validation sub-study) ?	M1

Item 17 – Outcome measurement (Data quality)

	Tool	**Content**	**Level**
2	GATE-GATE	were exposures & outcomes well measured?	M1
3	MMAT	Are measurements appropriate regarding both the outcome and intervention (or exposure)? Are the measurements appropriate?	M1
4	CASP	Did they use subjective or objective measurements? Do the measurements truly reflect what you want them to (have they been validated) Has a reliable system been established for detecting all the cases (for measuring disease occurrence)? Were the measurement methods similar in the different groups? Was the outcome accurately measured to minimize bias?	M2
5	SURE	Was the condition measured in a standard, reliable way for all participants included in the case series?	M1
6	JBI	Were outcomes assessed in a standard, valid and reliable way for cases and controls? Were the outcomes measured in a valid and reliable way?	M1

		Was the condition measured in a standard, reliable way for all participants included in the case series?	
7	ROBINS-I	Could the outcome measure have been influenced by knowledge of the intervention received? Were the methods of outcome assessment comparable across intervention groups? Were any systematic errors in measurement of the outcome related to intervention received?	M2
8	CER-CI	Were the primary outcomes defined and measured in a valid way?	M1
9	GRACE	Was the primary clinical outcome measured objectively rather than subject to clinical judgment (e.g., opinion about whether the patient's condition has improved)? Were primary outcomes validated, adjudicated, or otherwise known to be valid in a similar population?	M2
10	NIH	For the analyses in this paper, were the exposure(s) of interest measured prior to the outcome(s) being measured? Were the exposure measures (independent variables) clearly defined, valid, reliable, and implemented consistently across all study participants? Were the outcome measures (dependent variables) clearly defined, valid, reliable, and implemented? consistently across all study participants?	M2
13	RTI Item bank	Are outcomes assessed using valid and reliable measures, implemented consistently across all study participants? Primary outcomes should be identified for abstractors and if there is more than one, they may be listed separately. Also, identify any relevant secondary outcomes and harms. Subjective measures based on self-report tend to have lower reliability and validity than objective measures such as clinical reports and lab findings. Note for Case-control studies: consider whether the ascertainment of cases was independent of exposure.	M2
15	Montreal	What are the outcome factors and how are they measured? a) Are all relevant outcomes assessed? b) Is there measurement error?	R2
16	STROBE	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group.	R2
17	TREND	Clearly defined primary and secondary outcome measures. Methods used to collect data and any methods used to enhance the quality of measurements.	R1 & M1
18	ACROBAT-NRSI	Was the outcome measure objective? Were any systematic errors in measurement of the outcome unrelated to intervention received?	M2
21	Rangel et al. 2003	Is the diagnostic method clearly described for assessing outcome(s) of interest?	R1
25	Blagojevic et al. 2010	Appropriate and validated outcome measure.	M1
26	Genaidy et al. 2007	Are the main outcome measures reliable? Are the methods of assessing the outcome variables standard across all groups?	M1
27	Harm	Were treatments/exposures and clinical outcomes measured in the same way for both groups?	M1
28	Tseng et al. 2009	Methods for assessing outcomes described?	R1

31	ANDQ	Were outcomes clearly defined and the measurements valid and reliable? Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? Was the measurement of effect at an appropriate level of precision? 7.6 Were other factors accounted for (measured) that could affect outcomes? Were the measurements conducted consistently across groups?	M2
32	NICE	A valid and reliable method was used to determine the outcome. What outcome measure(s) is/are used?	R1&M1
33	IHE	Were relevant outcomes appropriately measured with objective and/or subjective methods? Were outcomes measured before and after intervention?	M1
34	AXIS	Were the risk factor and outcome variables measured appropriate to the aims of the study? Were the risk factor and outcome variables measured correctly using instruments/ measurements that had been trialled, piloted or published previously?	M2
35	AHRQ	Outcomes are measured using valid and consistent procedures and instruments across all study participants. Errors in measurement of the outcome are unrelated to the intervention received (i.e., no differential misclassification of outcomes) ?	M2
36	Pluye et al. 2009	Justification of measurements (validity and standards).	M1
37	Heller et al. 2008	Is there bias in the measurement? Are these outcome measures appropriate?	M1
39	Faillie et al. 2017	RCT, cohort studies: Is the time frequency of drug safety outcome assessment during the follow-up period appropriate?	M1
40	IPM-QRBNR	Outcomes Assessment Criteria for Significant Improvement: No descriptions of outcomes OR < 20% change in pain rating or functional status (0); Pain rating with a decrease of 2 or more points or more than 20% reduction OR functional status improvement of more than 20% (1); Pain rating with decrease of ≥ 2 points AND ≥ 20% change or functional status improvement of ≥ 20% (2); Pain rating with a decrease of 3 or more points or more than 50% reduction OR functional status improvement with a 50% or 40% reduction in disability score (3); Significant improvement with pain and function ≥ 50% or 3 points and 40% reduction in disability scores (4).	M2
41	Handu et al. 2016	Were outcomes clearly defined and the measurements valid and reliable?	M1
43	Young et al. 2009	Were study measures objective or subjective and is recall bias likely if they were subjective?	M2
45	ENCePP	Does the protocol describe how the outcomes are defined and measured?	R1&M1

		Does the protocol address the validity of outcome measurement (e.g. precision, accuracy, sensitivity, specificity, positive predictive value, use of validation sub-study) ?	
48	Kennedy et al. 2019	Comparison group outcome matching is assessed in multi-arm studies to establish whether there were statistically significant baseline differences in study outcome measures. As above, study arms include intervention, control, or comparison groups. Outcome measures are those which the intervention is trying to change; they generally include things like knowledge, attitudes, behaviors, or biological outcomes. There may be one or more outcome measures in any given study. If the study arms are equivalent on outcome measures at baseline, this criterion is met. If there are statistically significant differences between one or more of the study arms on outcome measures at baseline, this criterion is not met.	M2

Item 18 - Blinding of outcome (Data quality)

	Tool	**Content**	**Level**
2	GATE-GATE	Were outcomes measured blind to whether participant was in EG or CG (or vice versa)?	M1
3	MMAT	Are outcome assessors blinded to the intervention provided?	M1
4	CASP	Were the subjects and/or the outcome assessor blinded to exposure (does this matter)	M1
7	ROBINS-I	Were outcome assessors aware of the intervention received by study participants?	M1
10	NIH	Were the outcome assessors blinded to the exposure status of participants?	M1
12	ROBANS	Blinding of outcome assessments.	M1
13	RTI Item bank	Were the outcome assessors blinded to the intervention or exposure status of participants? There may be circumstances where clinical evaluators cannot be blinded to exposure status. Drop if not relevant to the body of literature.	M2
14	SIGN	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable. Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	M2
17	TREND	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to study condition assignment; if so, statement regarding how the blinding was accomplished and how it was assessed.	R2 & M2
19	MINORS	Unbiased assessment of the study endpoint: blind evaluation of objective endpoints and double-blind evaluation of subjective endpoints. Otherwise the reasons for not blinding should be stated.	R2 & M2
20	GRADE	Was there blinding of outcome assessment (i.e. no potential for detection bias)?	M2
21	Rangel et al. 2003	If comparison groups were used, was any attempt made to blind evaluators during the analysis of data?	M1
22	Thomas et al. 2004	Was (were) the outcome assessor(s) aware of the intervention or exposure status of participants?	M1
27	Harm	Was the assessment of outcomes either objective or blinded to exposure?	M1
28	Tseng et al. 2009	Was any attempt made to blind evaluators during the analysis of data?	M1

30	NOS	Assessment of outcome: independent blind assessment; record linkage.	R1
31	ANDQ	Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded?	M2
32	NICE	Investigators were kept 'blind' to other important confounding and prognostic factors.	M1
33	IHE	Blind assessment of outcomes.	M1
39	Faillie et al. 2017	Was the blinding method of drug safety outcome assessment appropriate considering the nature of the adverse event?	M1
41	Handu et al. 2016	Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded?	M2
44	ISPE	For any endpoint or covariate status ascertainment (in a cohort study or trial) or exposure ascertainment (in a case–control study) that requires adjudication, all measures taken to assure blinding of the adjudicators to the exposure (cohort) or outcome (case–control) status of the subject should be outlined in the protocol.	R2
47	OSTEBA	Was the assessment of the results of both tests blind?	M1

Item 19 - Missing data (Data quality)

	Tool	**Content**	**Level**
1	RELEVANT	The extent of missing data is reported.	R1
3	MMAT	Are there complete outcome data? Almost all the participants contributed to almost all measures.	R1
5	SURE	Data analysis Are the statistical methods well described? Consider: How missing data was handled; were potential sources of bias (confounding factors) controlled for; How loss to follow-up was addressed.	R1&M1
6	JBI	Was follow up complete, and if not, were the reasons to loss to follow up described and explored? Were strategies to address incomplete follow up utilized?	M1
7	ROBINS-I	Were outcome data available for all, or nearly all, participants? Were participants excluded due to missing data on intervention status? Were participants excluded due to missing data on other variables needed for the analysis?	M2
8	CER-CI	Was the extent of missing data reported?	R1
12	ROBANS	Incomplete outcome data. Attrition bias caused by the inadequate handling of incomplete outcome data.	R1
14	SIGN	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed. Comparison is made between full participants and those lost to follow up, by exposure status.	M2
16	STROBE	Explain how missing data were addressed.	R2

		Indicate number of participants with missing data for each variable of interest.	
17	TREND	Methods for imputing missing data, if used.	R1
18	ACROBAT-NRSI	Are outcome data reasonably complete? Was intervention status reasonably complete for those in whom it was sought? Are data reasonably complete for other variables in the analysis?	M2
21	Rangel et al. 2003	Do the authors address whether there is any missing data?	M1
25	Blagojevic et al. 2010	Loss and dropout at follow-up <25%. Adequate description and discussion of dropouts.	R1 & M1
26	Genaidy et al. 2007	Have the characteristics of subjects lost after entry into the study or subjects not participating from among the eligible population been described? Have the details of unavailable records been described?	R1 & M2
28	Tseng et al. 2009	Do the authors address whether there is any missing data? If not explicitly addressed, answer 'No' unless it is obvious there is no missing data.	M1
29	NHS Wales	Is there an explanation of how missing data have been handled?	M1
31	ANDQ	Was method of handling withdrawals? Were follow up methods described and the same for all groups? Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.)? Were all enrolled subjects/patients (in the original sample) accounted for? Were reasons for withdrawals similar across groups?	M2
35	AHRQ	Outcome data are reasonably complete and proportion of participants and reasons for missing data are similar across groups. Confounding variables that are controlled for in the analysis are reasonably complete across participants. Appropriate statistical methods are used to account for missing data (i.e., intention-to-treat analyses using appropriate imputation techniques). Intervention status is reasonably complete and does not differ systematically between groups.	M2
45	ENCePP	Does the plan describe methods for handling missing data?	R1
49	RECORD	Explain how missing data were addressed.	R1

Item 20 - Length of follow-up (Data quality)

	Tool	**Content**	**Level**
6	JBI	Was the exposure period of interest long enough to be meaningful? Was the follow-up period of sufficient duration to detect differences addressed? (Refer to the original tool for more details)	M2
10	NIH	Was the timeframe sufficient so that one could reasonably expect to see an association between exposure and outcome if it existed?	M2

11	Weightman et al. 2004	Was follow up for long enough?	M1
13	RTI Item Bank	Is the length of follow-up the same for all groups? For Case-control studies, are cases and controls matched on length of follow-up? When follow-up was the same for all study participants, the answer is yes. If different lengths of follow-up were adjusted by statistical techniques, (e.g., survival analysis), the answer is yes. Studies in which differences in follow-up were ignored should be answered no. Is the length of time following the intervention/exposure sufficient to support the evaluation of primary outcomes and harms? Primary outcomes (including harms) should be identified for abstractors. Important measures may be listed separately. Abstractors should be provided with specific criteria for sufficient length of follow-up based on prior research or theory. Drop if entire body of evidence is cross-sectional or if minimal length of follow-up period is specified through inclusion criteria.	M2
19	MINORS	Follow-up period appropriate to the aim of the study: the follow-up should be sufficiently long to allow the assessment of the main endpoint and possible adverse events.	M2
23	Atluri et al. 2008	Length of follow-up adequate for question.	M1
25	Blagojevic et al. 2010	Length of follow-up is about 36 months.	M1
26	Genaidy et al. 2007	Is the minimum follow-up time since initial exposure sufficient enough to detect a relationship between exposure/intervention and outcome?	M2
27	Harm	Was the follow-up of study patients sufficiently long for the outcome to occur?	M1
30	NOS	Was follow-up long enough for outcomes to occur? a) yes (select an adequate follow up period for outcome of interest) ¯ b) no.	M1
31	ANDQ	Was the period of follow-up long enough for important outcome(s) to occur?	M1
32	NICE	The study had an appropriate length of follow-up.	M1
39	Faillie et al. 2017	Was the duration of follow-up adequate to assess the drug safety outcome?	M1
40	IPM-QRBNR	Duration of Follow-up with appropriate interventions: less than 3 months or less for epidural or facet joint procedures, etc., and 6 months for intradiscal procedures and implantables (1) 3-6 months for epidural or facet joint procedures, etc., or one year for intradiscal procedures or implantables (2) 6-12 months for epidurals or facet joint procedures, etc., and 2 years or longer for discal procedures and implantables (3) 18 months or longer for epidurals and facet joint procedures, etc., or 5 years or longer for discal procedures and implantables (4).	M2
41	Handu et al. 2016	Was the period of follow-up long enough for important outcome(s) to occur?	M1

Item 21 - Loss to follow-up (Data quality)

	Tool	Content	Level
4.1	CASP	Does the study adequately address biased loss to follow-up? Was the follow up of subjects complete enough?	M1
5	SURE	Was follow-up ≥ 80%?	M1
6	JBI	Was follow up complete, and if not, were the reasons to loss to follow up described and explored? Were strategies to address incomplete follow up utilized? (Refer to the original tool for more details)	M2
10	NIH	Was loss to follow-up after baseline 20% or less? (Refer to the original tool for more details)	M2
13	RTI Item Bank	Did attrition from any group exceed [x] percent? Attrition is measured in relation to the time between baseline (allocation in some instances) and outcome measurement for both retrospective and prospective studies and could include data loss from crossover. Attrition rates may vary by outcome and time of measurement. Specify the criterion to meet relevant standards for the topic. Specify measurement period of interest, if repeated measures. Cochrane standard for attrition is 20 percent for shorter term (<1 year) and 30 percent for longer term (>1year). Drop of entire body of evidence is cross-sectional. Did attrition differ between groups by more than 20 percent? [PI: If appropriate, modify difference criterion to meet relevant standards for the topic. Attrition rates may vary by outcome and time of measurement. Drop if entire body of evidence is cross-sectional or case series.]	M2
14	SIGN	Comparison is made between full participants and those lost to follow up, by exposure status.	M1
16	STROBE	Report numbers of individuals at each stage of study—e.g. numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed.	R2
17	TREND	Follow-up: the number of participants who completed the follow-up or did not complete the follow-up (i.e., lost to follow-up), by study condition.	R2
19	MINORS	Loss to follow up less than 5% : all patients should be included in the follow up. Otherwise, the proportion lost to follow up should not exceed the proportion experiencing the major endpoint	M2
22	Thomas et al. 2004	Withdrawals and dropouts Follow-up rate of >80% of participants; Follow-up rate of 60–79% of participants; Follow-up rate of <60% of participants or withdrawals and dropouts not described.	M2
25	Blagojevic et al. 2010	All subjects aged 50 or over at follow-up; Loss and dropout at follow-up <25%.	M2
26	Genaidy et al. 2007	Are the participation rate(s) reported? Are ascertainments of record availability described? Are subject losses or unavailable records after entry into the study taken into account?	R2

30	NOS	Adequacy of follow up of cohorts a) complete follow up - all subjects accounted for ¯ b) subjects lost to follow up unlikely to introduce bias - small number lost - > ____ % (select an adequate %) follow up, or description provided of those lost) ¯ c) follow up rate < ____% (select an adequate %) and no description of those lost d) no statement.	M2
31	ANDQ	Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.)	M2
33	IHE	Was the loss to follow-up reported?	R1
39	Faillie et al. 2017	RCT, cohort studies: Does the study adequately address biased loss to follow-up?	M1
43	Young et al. 2009	Were there important losses to follow-up?	M1
46	ISPE-ISPOR	Reporting on follow-up time should include: Censoring criteria The criteria that censor follow up.	R1
48	Kennedy et al. 2019	Attrition of participants is measured at the final study follow-up. This is related to incomplete reporting, or loss-to-follow-up, that may introduce bias if participants who are retained are different than those who are not retained. One rule of thumb suggests that < 5% loss leads to little bias, while > 20% poses serious threats to validity [34]. This criterion is measured across the entire study population (all study arms). If the entire study group had a follow-up rate of 80% or more, this criterion is met. If the follow-up rate was less than 80% at the final assessment, this criterion is not met. For studies that are post-intervention only or serial cross-sectional in nature, this criterion should be listed as not applicable.	M2
49	RECORD	If applicable, explain how loss to follow-up was addressed.	R1

Item 22 - Description (Data analysis)

	Tool	**Content**	**Level**
1	RELEVANT	Potential confounders are addressed; Study groups are compared at baseline.	M1
4	CASP	Is the analysis appropriate to the design.	M1
5	SURE	Are the statistical methods well described?	M1
6	JBI	Were strategies to deal with confounding factors stated? Was appropriate statistical analysis used?	M1
8	CER-CI	Were analyses of subgroups or interaction effects reported for comparison groups?	M2
9	GRACE	Were any meaningful analyses conducted to test key assumptions on which primary results are based? (E.g., were some analyses reported to evaluate the potential for a biased assessment of exposure or outcome, such as analyses where the impact of varying exposure and/or outcome definitions was tested to examine the impact on results?)	M2

13	RTI Item Bank	Does the analysis control for baseline differences between groups? [PI: Drop if entire body of evidence is case series or case control. Define adequate control. List critical baseline differences that need to be controlled.] In cases of high loss to follow-up (or differential loss to follow-up), is the impact assessed (e.g., through sensitivity analysis or other adjustment method)? Are the statistical methods used to assess the primary benefit outcomes appropriate to the data? [Abstractor: Question relates to precision and may not be relevant for systematic reviews that are able to pool data. The statistical techniques used must be appropriate to the data and take into account issues such as controlling for dose-response, small sample size, clustering, rare outcomes, and multiple comparisons. In normally distributed data the standard error, standard deviation, or confidence intervals should be reported. In non-normally distributed data, interquartile range should be reported. For cohort studies, if the outcome has a greater than 10 percent prevalence, consider if the risk ratio and relative risk need to be calculated] Are the statistical methods used to assess the main harm or adverse event outcomes appropriate to the data? [Abstractor: Question relates to precision and may not be relevant for systematic reviews that are able to pool data. The statistical techniques used must be appropriate to the data and take into account issues such as controlling for dose-response, small sample size, clustering, rare outcomes, and multiple comparisons. In normally distributed data, the standard error, standard deviation, or confidence intervals should be reported. In non-normally distributed data, inter-quartile range should be reported.]	M2
15	Montreal	Are statistical tests considered?	M1
16	STROBE	(a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) Cohort study—If applicable, explain how loss to follow-up was addressed Case-control study—If applicable, explain how matching of cases and controls was addressed 12 Cross-sectional study—If applicable, describe analytical methods taking account of sampling strategy.	R2
17	TREND	Description of the smallest unit that is being analyzed to assess intervention effects (e.g., individual, group, or community): • Statistical methods used to compare study groups for primary outcome(s), including complex methods for correlated data; • Statistical methods used for additional analyses, such as subgroup analyses and adjusted analysis • Methods for imputing missing data, if used; • Statistical software or programs used.	R2
19	MINORS	Adequate statistical analyses : whether the statistics were in accordance with the type of study with calculation of confidence intervals or relative risk	M1

24	Bishop et al. 2010	Adjust for potential confounders in statistical analysis.	M1
25	Blagojevic et al. 2010	Appropriate analysis.	M1
26	Genaidy et al. 2007	Are the statistical methods clearly described? Is prior history of disease and/or symptoms collected and included in the analysis? Is there adequate adjustment for covariates and confounders in terms of individual variables in the analyses? Is there adequate adjustment for covariates and confounders in terms of environment variables (other than exposure) in the analyses?	M2
28	Tseng et al. 2009	Statistical methods described? Statistical software identified?	M1
31	ANDQ	Was the statistical analysis appropriate for the study design and type of outcome indicators? Were statistical analyses adequately described the results reported appropriately Were correct statistical tests used and assumptions of test not violated? Were statistics reported with levels of significance and/or confidence intervals? Was “intent to treat” analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)? Was clinical significance as well as statistical significance reported?	M2
32	NICE	All groups were followed up for an equal length of time (or analysis was adjusted to allow for differences in length of follow-up).	M1
33	IHE	Were the statistical tests used to assess the relevant outcomes appropriate?	M1
34	AXIS	Is it clear what was used to determined statistical significance and/or precision estimates? (eg, p values, CIs) Were the methods (including statistical methods) sufficiently described to enable them to be repeated?	M1
37	Heller et al. 2008	Are statistical tests appropriate and correct?	M1
39	Faillie et al. 2017	Does the analysis adequately adjust for identified confounding factors? Does the analysis address time-dependent confounders? Are the statistical methods used to analyze the drug safety outcome appropriate? Is a survival analysis performed when there are individual differences in length of follow-up?	M2
44	ISPE	Methods for data analysis; Data analysis comprises comparisons and methods for analyzing and presenting results, categorizations, and procedures to control sources of bias and their influence on results, for example, possible impact of biases due to selection bias, misclassification, confounding, and missing data. For instance, the statistical procedures to be applied to the data to obtain point estimates and confidence intervals of measures of occurrence or association should be presented.	R2

		Any sensitivity analyses should be described. Details of the statistical analysis may be specified later, but before analysis begins.	
45	ENCePP	Are the statistical methods and the reason for their choice described? Is study size and/or statistical precision estimated? Are stratified analyses included? Does the plan describe methods for analytic control of confounding? Does the plan describe methods for analytic control of outcome misclassification?	R2
46	ISPE-ISPOR	Reporting on statistical software should include: Statistical software program used, The software package, version, settings, packages or analytic procedures.	R1
47	RECORD	Describe the methods used to evaluate whether the assumptions have been met. 12.1.b: Describe and justify the use of multiple designs, design features, or analytical approaches.	R2

Item 23 - Sensitivity analysis (Data analysis)

	Tool	**Content**	**Level**
1	RELEVANT	The authors describe the statistical uncertainty of their findings (e.g. p values, confidence intervals).	M1
4	CASP	How precise was the estimate of the treatment effect? How precise are the results?	M1
8	CER-CI	Were sensitivity analyses performed to assess the effect of key assumptions or definitions on outcomes?	M2
9	GRACE	Were any meaningful analyses conducted to test key assumptions on which primary results are based?	M2
13	RTI Item Bank	In cases of high loss to follow-up (or differential loss to follow-up), is the impact assessed (e.g., through sensitivity analysis or other adjustment method)?	M1
16	STROBE	Describe any sensitivity analyses.	R1
39	Faillie et al. 2017	Cohort, Case-control studies: Do sensitivity analyses account for different exposure windows, induction/lag periods? (Refer to the original tool for more details)	M2
42	Viswanathan et al. 2018	Use processes to reduce uncertainty in individual judgments such as dual independent assessment of risk of bias with an unbiased reconciliation method. Avoid the presentation of risk-of-bias assessment solely as a numerical score; at minimum, consider sensitivity analyses of these scores. ~ When summarizing the evidence, consider conducting sensitivity analyses to evaluate whether including the studies with high or unclear risk-ofbias influence the estimate of effect or heterogeneity.	M2
44	ISPE	Any sensitivity analyses should be described. Details of the statistical analysis may be specified later, but before analysis begins, as part of a protocol amendment to the study protocol, or more typically as a separate document, usually referred to as a Statistical Analysis Plan.	R1

45	ENCePP	Are relevant sensitivity analyses described?	R1

Item 24 – Bias adjustment (Data Analysis)

	Tool	Content	Level
1	RELEVANT	Possible biases and/or confounding factors described.	R1
3	MMAT	Are the confounders accounted for in the design and analysis?	M1
4	CASP	Have the authors taken account of the potential confounding factors in the design and/or in their analysis?	M1
5	SURE	Were potential sources of bias (confounding factors) controlled for?	M1
6	JBI	Were confounding factors identified? Were strategies to deal with confounding factors stated?	R1&M1
7	ROBINS-I	List the confounding domains relevant to all or most studies. Were adjustment techniques used that are likely to correct for the presence of selection biases? Was an appropriate analysis used to estimate the effect of starting and adhering to the intervention? (Refer to the original tool for more details)	R2&M2
9	GRACE	Were important confounding and effect modifying variables taken into account in the design and/or analysis? Appropriate methods to take these variables into account may include restriction, stratification, interaction terms, multivariate analysis, propensity score matching, instrumental variables, or other approaches.	R2&M2
11	Weightman et al. 2004	Is confounding and bias considered? • (cohort study) Were the assessors blind to the different groups? • (cohort study) Could selective drop out explain the effect? • (Case-control study) How comparable are the cases and controls with respect to potential confounding factors? • (Case-control study) Were interventions and other exposures assessed in the same way for cases and controls? • (Case-control study) Is it possible that overmatching has occurred in that cases and controls were matched on factors related to exposure?	M2
13	RTI Item Bank	Are confounding and/or effect modifying variables assessed using valid and reliable measures across all study participants? [PI: Some characteristics may require that sources for establishing their validity and/or reliability be described or referenced. If so, provide instruction to abstractors.] Were the important confounding and effect modifying variables taken into account in the design and/or analysis (e.g., through matching, stratification, interaction terms, multivariate analysis, or other statistical adjustment)? [PI: Provide instruction to abstractors on adequate adjustment for confounding and testing for effect modification.]	M2
14	SIGN	How well was the study done to minimise the risk of bias or confounding?	M1

		The main potential confounders are identified and taken into account in the design and analysis.	
15	Montreal	What important potential confounders are considered? Does this selection bias threaten the external validity of the study?	M2
16	STROBE	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	R2
18	ACROBAT-NRSI	Is confounding of the effect of intervention unlikely in this study? Were the controls sampled from the population that gave rise to the cases, or using another method that avoids selection bias?	M2
19	MINORS	Baseline equivalence of groups: the groups should be similar regarding the criteria other than the studied endpoints. Absence of confounding factors that could bias the interpretation of the results.	M2
23	Atluri et al. 2008	Assessment of confounding. Study groups comparable to non-participants with regard to confounding factors.	M2
24	Bishop et al. 2010	Adjust for potential confounders in statistical analysis.	M1
26	Genaidy et al. 2007	Are the important covariates and confounders described in terms of individual variables?	R1&M1
29	NHS Wales	Have confounding and bias been considered? Is there an explanation of how potential confounding factors have been controlled for?	R1&M1
31	ANDQ	If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)? If a cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? Were other factors that could affect outcomes (e.g., confounders) measured or accounted for? Are biases and study limitations identified and discussed?	M2
32	NICE	The groups were comparable at baseline, including all major confounding and prognostic factors Investigators were kept 'blind' to other important confounding and prognostic factors.	M2
33	IHE	Study groups comparable to nonparticipants with regard to confounding factors. Discussion of possible confounders.	M1
35	AHRQ	For nonrandomized studies, specify likely sources of potential confounding.	R1
37	Heller et al. 2008	Has confounding been dealt with adequately? What important confounders are considered, and how are they addressed? Has confounding been dealt with adequately? Are there other confounders that should have been addressed?	M1
39	Faillie et al. 2017	Was the method for ascertaining confounders adequately constructed, and equal for all participants? Does the analysis adequately adjust for identified confounding factors? Does the analysis address time-dependent confounders? Is publication bias assessed?	M2

42	Viswanathan et al. 2018	For nonrandomized studies, specify likely sources of potential confounding. Make judgments about each risk-of-bias category (or item in a tool), using the preselected appropriate criteria for that study design and for each predetermined outcome.	R2&M2
45	ENCePP	Does the protocol address ways to measure confounding? (e.g. confounding by indication) Does the plan describe methods for analytic control of confounding?	R1

Item 25 - Are all the results presented (Results presentation)

	Tool	**Content**	**Level**
1	RELEVANT	Results are clearly presented for all primary and secondary endpoints as well as confounders. Are the results of this study directly applicable to the patient group targeted by this guideline? Confidence intervals are provided.	R1 & M1
3	MMAT	Quantitative and qualitative component in a mixed methods study" (Plano Clark and Ivankova, 2015, p. 40). Look for information on how qualitative and quantitative phases, results, and data were integrated (Pluye et al. 2009 et al., 2018). For instance, how data gathered by both research methods was brought together to form a complete picture (e.g., joint displays) and when integration occurred (e.g., during the data collection-analysis or/and during the interpretation of qualitative and quantitative results). 5.3. Are the outputs of the integration of qualitative and quantitative components adequately interpreted? Explanations This criterion is related to meta-inference, which is defined as the overall interpretations derived from integrating	M2
4	CASP	What are the bottom-line results? How precise was the estimate of the treatment effect? Consider • size of the p-value • size of the confidence intervals • have the authors considered all the important variables • how was the effect of subjects refusing to participate evaluated.	R2
5	SURE	Were all important outcomes assessed? Were outcome measures reliable (eg objective or subjective measures)? Are effect sizes, confidence intervals/standard deviations provided? Were all outcome measurements complete? Are the authors' conclusions adequately supported by the results?.	R2
6	JBI	Was there clear reporting of clinical information of the participants? Were the outcomes or follow up results of cases clearly reported? Was there clear reporting of the presenting site(s)/clinic(s) demographic information?	R2
7	ROBINS-I	Is the reported effect estimate likely to be selected, on the basis of the results, from … - multiple outcome *measurements* within the outcome domain? - multiple *analyses* of the intervention-outcome relationship?	M2

		- different *subgroups*?	
8	CER-CI	Was the number of individuals screened or selected at each stage of defining the final sample reported? Did the authors describe and report the key components of their statistical approaches? Were confounder-adjusted estimates of treatment effects reported? Did the authors describe the statistical uncertainty of their findings?	R2
11	Weightman et al. 2004	Were all important outcomes/results considered?	M1
12	ROBANS	Selective outcome reporting. Reporting bias caused by the selective reporting of outcomes.	M1
13	RTI Item bank	Are any important primary outcomes missing from the results? [PI: Identify all primary outcomes, including timing of measurement, that one would expect to be reported in the study.] Are any important harms or adverse events that may be a consequence of the intervention/exposure missing from the results? [PI: Identify all important harms, including timing of measurement, that one would expect be reported in the study. Drop if not relevant to body of literature.]	R2&M2
16	STROBE	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included; (b) Report category boundaries when continuous variables were categorized; (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period; Summarise key results with reference to study objectives.	R2
20	GRADE	Were data reported consistently for the outcome of interest (i.e., no potential selective reporting)?	M1
26	Genaidy et al. 2007	Are the characteristics of study participants described? Have all important adverse effects been reported that may be consequences of the intervention(s)? Are the main findings of the study clearly described? Are outcome data reported by levels of exposure?	R2
28	Tseng et al. 2009	Interpretation of results provided? Explicitly address study hypotheses/objectives? Was the patient population from which the cases were selected from adequately described or identified (e.g. geographically)? Are study capture rates provided? If stated that 'all' patients were captured within a given period, then answer 'Yes.' Are relevant baseline demographic and clinical data given for each group?	R2

		Are actual numbers, alone or in addition to percentages, furnished for all demographic variables? Are actual numbers, alone or in addition to percentages, furnished for all results? Is the number and nature of complications addressed? For longitudinal studies, is attrition of subjects and reason for attrition recorded? Are exact P-values for significant results provided (<0.01 acceptable)? Check text for data not reported in tables/figures. Are exact P-values for insignificant results provided? Check text for data not reported in tables/figures.	
29	NHS Wales	Are tables/graphs adequately labelled and understandable? Are you confident with the authors' choice and use of statistical methods, if employed? If sub-group/interactions analyses have been undertaken is there an explanation of how/why sub-groups have been formed? Is there an explanation of how potential confounding factors have been controlled for? Is there an explanation of how missing data have been handled? Are both unadjusted and adjusted (ie for confounding) results given if appropriate? Is the precision of estimates (95% CI) given? Do you believe the results?	R2
34	AXIS	Were the basic data adequately described? Does the response rate raise concerns about non-response bias? If appropriate, was information about non-responders described? Were the results internally consistent? Were the results for the analyses described in the methods, presented	R2
35	AHRQ	Outcomes are prespecified and all prespecified outcomes are reported No evidence that the intended measures, analyses, or subgroup analyses are selectively concealed	R1
37	Heller et al. 2008	What are the main results and are they presented in an understandable way? Have measures of absolute risk as well as relative risk been included? [For any intervention study] Have the resource and cost implications of implementing the intervention and cost-effectiveness of the intervention been described?	R2 & M1
39	Faillie et al. 2017	Are the results consistent in primary and secondary analyses? Are confounding effects consistent with known associations? Is there a clear flow chart of the studies?	M2
42	Viswanathan et al. 2018	Present findings and conclusions transparently, balancing the competing considerations of simplicity of presentation with burden on the reader.	R1
47	OSTEBA	Are the outcomes properly summarized and described?	R1&M1

Item 26 - Reasonable conclusions from results (Results presentation)

	Tool	**Content**	**Level**

1	RELEVANT	The clinical relevance of the results is discussed. Results are clearly presented for all primary and secondary endpoints as well as confounders. Results consistent with known information or if not, an explanation is provided.	R2
3	MMAT	Are the findings adequately derived from the data? Is the interpretation of results sufficiently substantiated by data? Is there coherence between qualitative data sources, collection, analysis and interpretation.	M2
5	SURE	Are the authors' conclusions adequately supported by the results? Are the conclusions the same in the abstract and the full text?	M1
11	Weightman et al. 2004	Are the authors' conclusions adequately supported by the information cited?	M1
13	RTI Item bank	Are results believable taking study limitations into consideration? [Abstractor:This question is intended to capture the overall quality of the study. Consider issues that may limit your ability to interpret the results of the study. Review responses to earlier questions for specific criteria.]	M2
16	STROBE	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence.	R2
17	TREND	Interpretation of the results, taking into account study hypotheses, sources of potential bias, imprecision of measures, multiplicative analyses, and other limitations or weaknesses of the study Discussion of results taking into account the mechanism by which the intervention was intended to work (causal pathways) or alternative mechanisms or explanations.	R2
23	Atluri et al. 2008	Conclusions supported by results with possible biases and limitations taken into consideration.	M1
31	ANDQ	Are conclusions supported by results?	M1
33	IHE	Conclusions of the study supported by results.	M1
34	AXIS	Were the authors' discussions and conclusions justified by the results?	M1
37	Heller et al. 2008	Have the results been interpreted appropriately?	M1
38	CARE	Rationale for conclusions (including assessments of cause and effect).	M1
39	Faillie et al. 2017	Is publication bias assessed? (Refer to the original tool for more details)	M2
41	Handu et al. 2016	Are conclusions supported by results with biases and limitations taken into consideration?	M1
43	Young et al. 2009	Do the data justify the conclusions? The next consideration is whether the conclusions that the authors present are reasonable on the basis of the accumulated data. Sometimes an overemphasis is placed on statistically significant findings that invoke differences that are too small to be of clinical value; alternatively, some researchers might dismiss large and potentially important differences between groups that are not statistically significant, often because sample sizes were small. Other issues to be wary of are whether the authors generalized their findings to broader groups of patients or contexts than was reasonable given their study sample, and whether statistically significant associations have been misinterpreted to imply a cause and effect.	M2
47	OSTEBA	Are the conclusions justified?	M1