Analysing the guidelines and assessment tools for Real-World Evidence studies

Michiel K.B. Hartog
3989097
Utrecht University
Graduate School of Life Sciences
Science and Business Management
Division of Pharmacoepidemiology and Clinical Pharmacology
Utrecht Institute for Pharmaceutical sciences
Junfeng Wang (examiner 1)
Aukje K Mantel-Teeuwisse (examiner 2)
Li Jiu (daily supervisor)
17-01-2023

Index

Abstract

Background: Non-randomised interventional studies, as a source of real-world data, provide valuable information for decision-makers around the world. As non-randomised studies are more susceptible to biases, their quality, in terms of risk of bias and reporting, should be evaluated, with the help of quality assessment tools. For this reason, health technology assessment (HTA) agencies rely on critical appraisal tools to evaluate non-randomised studies. However, information is lacking on which tool(s) should be used.

Objective: This study aimed to identify tools used to assess the risk of bias of non-randomised interventional studies, and to provide recommendations on which tools to use. As HTA agencies rely on the critical appraisal tools it is expected they provide recommendations on which tools to use. This study also explored the possibility of designing a new tool.

Method: We identified existing tools in three different approaches, we updated a search strategy from the study by D'Andrea et al. (2021), used the concept of snowballing to find similar publication, and we conducted a grey literature search to identify recommendations provided by the European HTA agencies. The included tools were assessed using a prespecified criteria list of 8 domains and 27 critical quality items.

Results: Of the 49 included tools, none sufficiently covered all the prespecified 27 critical quality items. A selection of tools covered almost all items (n=5), but they were not able to describe all items sufficiently. Among all the items, ethical approval is the item least covered by most tools (n=4). Based on the 27-item review, we designed the DRAGON tool, by combining items of all the included tools. The DRAGON tool covered and sufficiently described the 27 items on both methodology and reporting, and it can be used as a guideline for both methodological quality appraisal and reporting, when conduct or assessing a non-randomised intervention study.

Conclusions: None of e existing tools for assessing quality of NRSI studies were able to address all the important quality items. The DRAGON tool is a newly designed tool to assess the risk of bias and reporting of non-randomised intervention studies. To further test validity of the tool, pilot test and feedback from a panel of experts in the future are needed.

Introduction

In recent years, an interest in non-randomised studies of interventions (NRSI) has grown, as they provide useful insight into the real-world performance within biomedical and public health research. NRSI are especially useful for clinical and policy decision makers when generating hypothesis before trials are available, studying rare events, and answering research questions which would be unethical for trials. [1,2] Also, NRSI could provide valuable insight in effectiveness and cost-effectiveness of health intervention analyses for Health Technology Assessment purposes. [3]

What is real-world data

Real-world data is defined as all the data collected outside the scope of highly controlled randomised trials (RCTs). It can be gathered from a variety of sources such as electronic health records, patient registries, pharmacy and health insurance databases, social media, and patient-powered research networks. [4] Most of the real-world data is routinely collected as part of the provided healthcare. However, a portion of the real-world data is also gathered through observational (non-interventional) study design. This means an intervention is only allocated by physicians with consent of the patient. This in accordance with the terms of the marketing authorisation.

With randomization biases are minimized, as randomization ensures comparability between intervention and the control groups, by evenly distributing potential confounders. Also, information bias is minimized by the use of blinding. The major drawback of observational studies, compared to highly controlled clinical trials, is that they are prone to being biased due to the lack of randomisation. According to study design, observational studies can be categorized as cohort, case-control, cross sectional, and ecological studies.

In experimental study designs researchers introduce an intervention and study its effects. An experimental study design includes both RCTs and non-randomised controlled trials, which also belong to real-world data. In short, non-randomised controlled design studies include non-randomised controlled trial, controlled before-and-after study, interrupted time series study, cohort studies, case-control, cross-sectional study, and case series (uncontrolled longitudinal study). Due to the non-randomised design of these studies, they are more prone to bias.

In short non-randomised controlled design studies include non-randomised controlled trial, controlled before-and-after study, interrupted time series study, cohort studies, case-control, cross-sectional study, and case series (uncontrolled longitudinal study). Due to the non-randomised design of these studies, they are more prone to being biased.

Risk of Bias

A bias is a systemic error that is a threat to the validity of a study. Bias may occur due to mistakes in the design of studies, how they are conducted or how the results are analysed and interpreted. Biases can lead to either an overestimation or an under estimation of the true effect. Because the results of a study may in fact be unbiased despite a methodological flaw, it is more appropriate to consider the risk of bias. [5]

The Cochrane library defines four different types of bias for non-randomised studies for intervention. The four main types of bias are information bias, selection bias, reporting bias and confounding. [6]

Information bias

Information bias, also known as measurement bias, is any systematic difference from the truth that arises in the collection, recall, recording and handling of information in a study. This also includes how missing data is dealt with. [7] Information bias is an umbrella term used to cover a wide range of different types of biases such as, misclassification bias, observer bias, recall bias, interviewer bias, response bias, reporting bias, ascertainment bias and confirmation bias.

Misclassification bias is the result when a study participant is incorrectly categorised for exposure or outcome of interest. This can alter the outcome of interest or observed association. When an observer has certain expectations or is prejudiced to an outcome it can cause observer bias. The observer can influence the reporting by what they perceive or record. Recall bias is a systematic error that occurs when a study participant does not remember specific events or experiences accurately. This could lead to differences between study groups. Recall bias is especially a problem for case-control studies and retrospective cohort studies as the questions are asked in the end of the study or years after an event. Interviewers need to be cautious when questioning the study participants as the interviewer's expectation and/or opinions may influence the objectivity of the study participant. This is called interviewers bias and can occur during any interview of a study participant. Response bias is a term used for a wide range of tendencies for participants to respond false or incorrect to questions. Reporting bias arises when the authors selectively reveal or suppress specific items that may alter the outcome of the study. Another method to prevent information bias is to introduce blinding of the patients, administrators, and data analysis.

Selection bias

Selection bias occurs when eligible participants in a study are systematically different from the population of interest, leading to a systematic error of association or outcome. Selection biases occur in non-randomised studies either due to selection of participants or follow-up time into the study or those who left the study, missing data. [8] The best method to prevent selection bias is by randomisation. However, in non-randomised studies this is not an option. Other steps to address selection bias might be by making the intervention- and control-group as comparable as possible, including the number of participants at each stage, comparability of intervention/exposure and control group at baseline, openness of allocation process, and how missing data is handled. Ascertainment bias is one of the forms of selection bias, it occurs when specific members of a target population are more likely to be included in a sample. Confirmation bias is the tendency to be looking for specific information that is consistent with one's personal existing beliefs.

Reporting bias

Reporting bias is defined as selective revealing or repressing findings, and it is common in both randomised and non-randomised studies. [8] Reporting bias can be divided into subtypes, including publication bias, time lag bias, multiple (duplicate) publication bias, location bias, citation bias, language bias and outcome reporting bias. [9] It is possible to address reporting bias with the use of reporting guidelines that ensure the writer to stay objective and includes all aspects of the study.

Confounding

A confounder is a distortion in the estimated measure of association that influences both the dependent and independent variable meaning. The confounder could predict the outcome and it is associated with the intervention or exposure, and itis extraneous to the occurrence relation. It is possible to adjust for confounding when potential confounders are known. However not all confounders are known for each study.

Quality assessment tools and reporting guidelines

To ensure the quality of NRSI, the agencies of health technology assessment rely on quality assessment tools and guidelines. These tools rigorously address risk of bias of primary studies and offer support in transparent writing. In recent years, the substantial number of tools guidelines have become a burden for their users, as they may feel overwhelmed and confused about which one to use. Previous studies found there is no consensus between the HTA agencies on the preferred appraisal tool. The sheer volume of tools makes it hard to decide on a golden standard. Furthermore, previous studies found not all tools address the same aspects and are not able to sufficiently address all critical elements needed for a full critical appraisal.

Our objective was to provide guidance in selecting the most appropriate tool(s) for NRSI and designing a quality assessment tool by summarizing items from existing tools. This tool could be used for all types of NRSI design.

Method

Protocol

The protocol is registered at the OSF registries identifier code: osf-registrations-kcsgx-v1. The protocol can be found on the website of the OSF registry https://osf.io/kcsgx.

We conducted a systematic review with three different approaches to identify different quality assessment tools for non-randomised studies for intervention (NRSI). The first step was to update the search strategy of the study by E. D'Andrea et al. (2021). [10] This approach was chosen to identify the newest publications on tools published after the initial search strategy. The second approach was to identify similar reviews via the process known as snowballing. With this method we are able to identify similar reviews based on their references. The third approach was a grey literature search on the websites of the European HTA agencies to review the recommendations of the HTA agencies. The HTA agencies were of importance as they heavily rely on critical appraisal tool, but no consensus exists between the agencies. Furthermore, the tools specifically developed by the HTA agencies are not readily available. All publications were assessed on the following in- and exclusion criteria.

Inclusion and exclusion of articles

Articles were included if they met the following criteria:

- Published in a peer-reviewed journal
- Were published in English
- Reviews on quality assessment tools for non-randomised studies of intervention

We excluded articles that were on:

- Interrater reliability

Search and screening

As described before a three-part approach has been used to identify publications on quality assessment tools. Two reviewers (MH and LJ) independently reviewed in two different rounds. Firstly, all titles and abstract were scanned for eligibility. Secondly, positive identified were fully scanned for eligibility. Tools identified in the included were also assessed on eligibility by the two reviewers (MH and LJ). Discrepancies were resolved by consensus or by consulting a third reviewer (JW).

Approach one: Updated search

First an updated search had been performed from the publication of D'Andrea et al. 2021 [10], the search strategy can be found in the appendix. In the study of D'andrea (2021) they conducted a search on PubMed and Embase from the inception of the databases up until November 2019. We extended this search strategy up until April 2022 to identify possible new or updated versions of tools. The identified publications were assessed using Rayyan [11], a free web-tool for systematic reviews.

Approach two: Snowballing

The second stage was to use the concept of Snowballing as described by Wohlin et al. 2014. [11] Snowballing is a search approach for systematic literature studies, it refers to using the reference list of a paper or the citations to identify additional. The snowballing approach starts

with defining a starting set of articles. We chose three different publications as our starting set: D'Andrea et al. 2021 [10], Quigley et al. 2018, [13] and Faria et al. 2015 [14]. These articles were chosen as they were reviews on NRSI quality assessment tools, and they matched our eligibility criteria. This study group agreed on the starting set of three to keep the amount of publications manageable.

The process of snowballing exists of backward and forward snowballing. Backward snowballing refers to first looking at the reference list of the starting set articles. First the title of the reference is assessed. When the reference might be of interest the abstract is read, if the abstract matched the topic the full article is read to see if the inclusion criteria are met. Forward snowballing refers to identifying articles that refer to the starting set articles, and the process is in essence the same as backward snowballing by first assessing the title and abstract and if of interest the full article is assessed. [12]

The process of snowballing can be simplified by using the powerful internet-based tool 'Connected Papers". Connected papers is a unique, visual tool to help researchers and applied scientists find and explore relevant to their field of work [15]. It is connected to the Semantic Scholar Paper Corpus that has an extensive database of articles published across many scientific fields. When a primary paper is uploaded, a graph is provided to return a list of 40 articles with similar topics and show how well they are linked. A table provides further insight as it shows the number of references, how many times a paper is cited and how similar it is to the original paper used to identify other. The website of ConnectedPapers can be accesses trough link https://www.connectedpapers.com/ .

Process of snowballing
We used a starting of set three articles: D'Andrea et al. 2021 [10], Quigley et al. 2018, [13] and Faria et al. 2015. [14] Each run with the online tool Connected Papers [15] revealed 40 hits per paper, duplicates were removed, the connected papers were than judged on eligibility criteria and if met the paper was included. The included from the first run were used for a second run of the snowballing process to identify more related. Each paper revealed another 40 that were judged on the eligibility criteria.

Approach three: HTA recommendations

We reviewed recommendations for assessment tools by HTA agencies. A grey literature search has been conducted on the websites of all 32 European HTA agencies identified on the International Network of Agencies for Health Technology Assessment (INAHTA). [16] All websites were searched using four search concepts: Critical appraisal tool, Quality assessment tool, Risk of Bias, and Methodology tool. Only the first ten hits of each search were used to identify recommendations by each HTA agency.

Eligibility of appraisal tools

The identified tools were assessed on the following eligibility criteria:

- We developed and/or updated after 2002
- Domain-based or checklist or scales
- Were on methodology and/or reporting
- Risk of bias tools

We excluded:

- Tools for assessing Randomised Controlled trail only
- Tools for exposure only
- Prevalence studies
- Previous versions of tools

Two reviewers (MH and LJ) reviewed the eligibility of the mentioned tools based on the eligibility criteria. One researcher (MH) downloaded the eligible tools and extracted descriptive items; name of the tool, year of publication, originality, type of tool (checklist, rating scale, summary judgement, scales, guideline, questionnaire), scope of tool, study design(s), intended for intervention, number of items and version).

Data collection

We compiled a data extraction sheet of eight different domains based on the twelve quality domains with 45 quality items identified by Deeks et al. 2003 [17]. We chose to alter/combine the identified quality domains and items as the suggested order by Deeks as the structure was not based on the structure of a scientific article. We restructured the eight domains to the following structure: (1) Background, (2) Population, (3) Intervention, (4) outcome, (5) data collection, (6) data analysis, (7) Results, and (8) Conflict of interest. Each of our eight domain consists of multiple items with a total of 27 items. An overview of each domain and its items with accompanying description is shown in figure 1.

Two reviewers (MH and LJ) independently extracted and coded data using the computer software NVivo12. The extracted items were classified as either methodology (M) or reporting (R). Methodological quality refers to how well a study was designed and executed for the prevention of systematic errors or bias. Methodology items are used for the assessment of the risk of bias. Any item related to the method is considered a methodological item. Items were considered on reporting if it refers in any way on how to report of what to include in the final paper. An item was classified as reporting if the words state, explain, and include. Reporting is included as it impacts to what extent the reader can evaluate the publication. We aim to differentiate methodology and reporting items as both aspects are important to assess the risk of bias. Some items of the included tools were classified as both methodology and reporting (R&M) as they covered both aspects.

The items extracted and classified were then graded on either a level 1 or a level 2, depending on how well the item was covered. If an item was just a brief statement without any explanation or guidance it was graded as a level 1, if an item was covered in detail with example aspects to take into consideration it was graded as a level 2. After data extraction, any discrepancies were resolved by consensus or consulting a third reviewer (JW). Each item was later summarised using both the level 1 and level 2 information as any item could contribute to the development of a new tool.

Background	**Definition**
1. Study objective	1. Study objective focusses on the research question and hypothesis
2. Protocol	2. A priori designed protocol, submitted before data collection
3. Study design	3. What is the study design, are other designs considered
4. Ethical approval	4. Has ethical approval been asked and received
Population	
5. Sample size/power calculation	5. Calculation that determines the required sample size
6. Eligibility criteria	6. Requirements that must be met to become a participant
7. Patients	7. General information on patients
8. Participation rate	8. The number of individuals in the selected sample who eventually participated
9. Baseline characteristics	9. Demographic, medical, and other information relevant to the variables
Intervention	10. How is the intervention selected
10. Selection	11. Definition of the intervention
11. Definition	12. Measurement of intervention/exposure
12. Measurement	13. Were applicators blinded from intervention
13. Blinding	14. Duration of follow-up period is it sufficient?
14. Length of follow-up	
Outcome	15. What effect(s)/outcome(s) are selected
15. Selection	16. Definition of the outcome
16. Definition	17. How is the outcome measured
17. Measurement	18. Are the assessors blinded?
18. Blinding	
Data collection	19. Where does the data come from
19. Data source	20. Data values that are not stored and are lost
20. Missing data	21. Participants lost during the study (e.g., they moved or died)
21. Loss to follow-up	
Data analysis	22. The process of systematically cleaning, transforming, and modelling data
22. Description	23. What is the level of uncertainty in the delivered output
23. Sensitivity	24. Factors that influence both the dependent and independent variable, distortion
24. Confounding	
Results	
25. Are all the results included	25. Are results missing or hidden
26. SPIN	26. Misleading, reporting, interpretation, or extrapolation of study results
Conflict of interest	
27. Conflict of interest	27. Financial or social factors that can compromise the judgment or decisions

Table 1. List of quality assessment domains and items used for the data extraction

Results

Overview of reviews

As shown in figure 1, we identified a total of 1967 reviews on quality assessment tools, only 28 reviews met our eligibility criteria (PRISMA flow-chart, figure 2). Of the 28 articles, only 1 article was included from the updated search, 20 papers were from snowballing, and 7 papers were from grey literature search. Of the 28 reviewed articles, 230 tools were identified, after removing duplicates. After assessing their eligibility, 49 tools were included. Of the included tools, one tool was identified with the updated search, 42 were identified through the process of snowballing and 5 were identified in the grey literature search visiting the websites of the European HTA agencies. The review of D'Andrea et al. (2008) included 35 tools, with 9 tools providing separate instruments to assess cohort, case-control, and case-series. Of the 34 tools identified by D'Andrea we included 19 tools in our study.

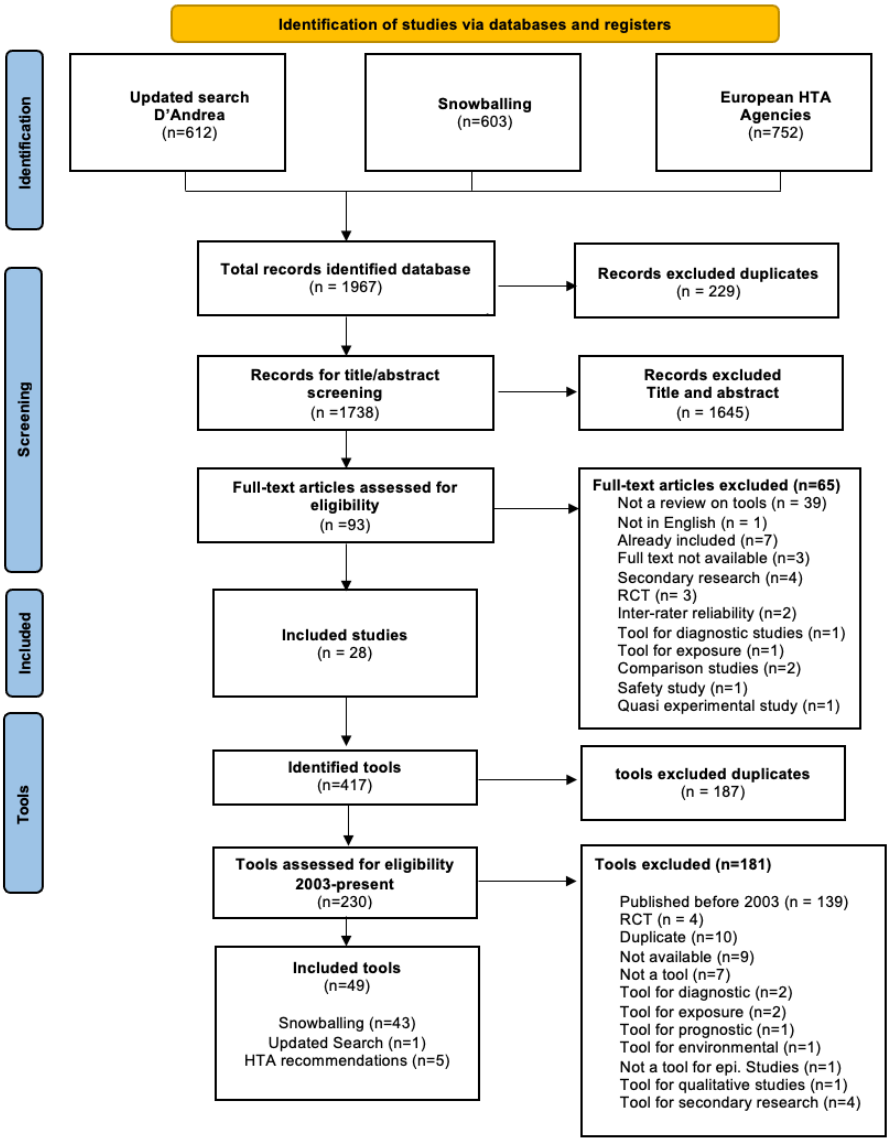

Figure 2. The PRISMA flow diagram details of our search and selection process of reviews and quality assessment tools during the review.

HTA recommendations

During the grey literature search, the recommendations from HTA agencies were reviewed, but only two HTA agencies gave a clear recommendation. The National Institute for Health, and Care Excellence (NICE), referred to the reviews performed by Sterne et al. 2016 and D'Andrea et al. 2021. The review by Sterne et al. recommended using the ROBINS-I[23] tool, while D'Andrea et al. 2021 found the ROBINS-I[23] tool and the GRACE checklist were the most comprehensive tools, and they should therefore be the tools of choice. The European Network for Health Technology Assessment (EUnetHTA), in a guideline published in 2014, considered the ACROBAT-NRSI[34] and the ROBANS, tools as the most suitable tools for the risk of bias assessment of NRSI. During the grey literature search multiple quality assessment tools were identified. However, only the NICE and EUnetHTA gave clear recommendations on what tools to use instead of publishing articles of tools.

	Tool	Year	Type of tool	Scope of tool	Study design	No. of items
1	RELEVANT (Real Life Evidence AssessmeNt Tool) [17]	2019	Checklist	Critical appraisal and reporting	Non-randomised study	21
2	RAMboMan - GATE-EPIQ [18]	2019	Ratingscale + summary judgement	Critical appraisal	Coh (+RCTs), CC	Coh 21, CC 18
3	Mixed Methods Appraisal Tool (MMAT) [19]	2018	Checklist	Critical appraisal	Non-randomised study	5
4	CASP [20]	2018	Checklist	Critical appraisal	Coh, CC	Coh 12, CC11
5	SURE [21]	2018	Questionnaire	Critical appraisal	RCT and other experimental	14
6	JBI (Joanna Briggs Insitute) [22]	2020	Checklist + summary judgement	Critical appraisal	Coh, CC	Coh 11, CC 10
7	ROBINS-I [23]	2017	Checklist + summary judgement	Critical appraisal	Non-randomised study	34 (+ 8 optional)
8	ISPOR-AMCP-NPC [24]	2016	Checklist + summary judgement	Critical appraisal	Coh, CC	32
9	GRACE - V5 [25]	2014	Checklist + summary judgement	Critical appraisal	Coh, CC	11
10	NIH-NHLBI [26]	2013	Checklist + summary judgement	Critical appraisal	Coh (+CSS), CC	Coh (+CSS) 14, CC 12
11	HEB Wales [27]	2004	Checklist + summary judgement	Critical appraisal	Coh	18
12	RoBANS [28]	2011	Ratingscale	Critical appraisal	Non-randomised study	6
13	RTI-Item Bank [29]	2011	Checklist	Critical appraisal	Non-randomised study	13
14	SIGN [30]	2014	Checklist + summary judgement	Critical appraisal	Coh, CC	Coh 14, CC 11
15	Montreal [31]	2011	Checklist	Critical appraisal	Coh, CC (+RCTs)	10
16	STROBE [32]	2019	Checklist	Reporting	Coh, CC	Coh 22, CC 22
17	TREND [33]	2004	Checklist	Reporting	Non-randomised study	22
18	ACROBAT-NRSI [34]	2014	Ratingscale	Critical appraisal	Non-randomised study	39
19	MINORS [35]	2003	Ratingscale	Methodology	Non-randomised study	12
20	GRADE [36]	2011	Checklist	Critical appraisal and reporting	RCT and non-randomised	
21	Rangel [37]	2003	Ratingscale	Critical appraisal	Coh	15
22	Thomas [38]	2004	Checklist	Critical appraisal	All study types	21
23	Atluri [39]	2008	Checklist	Critical appraisal	Coh, CC, CS	26
24	Bishop [40]	2009	Checklist	Critical appraisal	CS	17
25	Blagojevic [41]	2010	Checklist	Critical appraisal	Coh, CC	15
26	Genaidy [42]	2007	Checklist	Critical appraisal	Coh, CC, CS	22
27	Glasgow University [43]	2009	Checklist	Critical appraisal	Coh, CC	10
28	Tseng [44]	2008	Checklist	Critical appraisal	Coh	45
29	Weightman [45]	2004	Checklist	Critical appraisal	Coh, CC, CS	25
30	Wells [46]	2009	Checklist	Critical appraisal	Coh, CC	8
31	Quality Criteria Checklist: Primary research [47]	2010	Checklist	Primary and review	All study types	49
32	NICE checklist [48]	2013	Checklist	Methodology	Coh, CC	Coh 14, CC 20
33	IHE quality appraisal tool [49]	2012	Checklist	Critical appraisal	Non-randomised study	18
34	AXIS tool [50]	2016	Checklist	critical appraisal	CS	20
35	AHRQ methodology checklist [51]	2004	Checklist	Critical appraisal	CS	11
36	Pluye [52]	2009	Checklist	Critical appraisal	Mixed studies review (MRS	15
37	Heller [53]	2008	Checklist	Critical appraisal	All study types	39
38	Gagnier [54]	2013	Guideline	Reporting	Case-report	13
39	faillie [55]	2017	Checklist	Critical appraisal	Observation, RCT and revie	RCT 32, Coh 32, CC 24, NCC 25, systematic 10
40	Manchikanti [56]	2014	Guideline	Critical appraisal	Meta-analysis	16
41	Handu [57]	2016	Checklist	Critical appraisal	All study types	49
42	Viswanathan [58]	2018	Questionnaire	Critical appraisal	Observational studies	29
43	Young [59]	2009	Checklist	Critical appraisal	All study types	10
44	ISPE [60]	2016	Guideline	Protocol development	Non-randomised study and	26
45	ENCeppC [61]	2018	Checklist	Protocol development	All study types	68
46	RECORD [62]	2015	Guideline	Reporting	All study types	22
47	ISPOR-ISPE [63]	2017	Guideline	Reporting	All study types	39
48	Critical Appraisal tool -OSTEBA [64]	2019	Checklist + summary judgement	Critical appraisal	All study types	23
49	Kennedy [65]	2019	Checklist	Risk of Bias	Non-randomised study and	8

Table 2. The tools are ordered on the moment of identification. Basic characteristics of the tools included in this review. Tool name or name of the first author is used to identify the tool. The scope of the tool is as the developers classified the tool.
Coh, cohort; CC, case-control; CS. Case-series; RELEVANT, Real Life Evidence AssessmeNt tool; RAMboMan-GATE-EPIQ, Recruitment Allocation Maintenance blinded objective Measurements ANalyses-Graphic Approach To Epidemiology; CASP, The Critical Appraisals Skills Programme; SURE, Specialist Unit for Review Evidence; JBI, Joanna Briggs Institute; ROBINS-I, The Risk Of Bias In Non-randomized Studies-of Interventions; ISPOR-AMCP-NPC, The Professional Society for Health Economics and Outcome Research-the Academy of Managed Care Pharmacy-the National Pharmaceutical Council; GRACE; the Good Research for Comparative Effectiveness; NIH-NHBLI , National Institute for Health-National Heart-Lung and Blood Institute; ROBANS, Risk of Bias Assessment tool for Non-randomized Studies; RTI-Item bank; Research Triangle Institute-item bank; SIGN, Scottish Intercollegiate Guidelines Network; STROBE, the Strengthening the Reporting of Observational Studies in Epidemiology; TREND, Transparent Reporting of Evaluations with Nonrandomized Designs; ACROBAT-NRSI, A Cochrane Risk Of Bias Assessment Tool for Non-Randomized Studies of Interventions; MINORS, Methodological Index for NOn-Randomized Studies; GRADE, Grading of Recommendations, Assessment, Development and Evaluations; NICE, the National Insitute for Health and care Excellence; IHE, Insitute Health Economics; AXIS, Appraisal tool for Cross-Sectional Studies; AHRQ, Agency for Healthcare Research and Quality; ISPE, the International Society for Pharmaceutical Engineering; ENCeppC, European Network of Centres for Pharmacoepidemiology and Pharmacovigilance; RECORD, REporting of studies Conducted using Observational Routinely-collected health Data; ISPOR-ISPE, The Professional Society for Health Economics and Outcomes Research-International Society for Pharmacoepidemiology; OSTEBA, Basque office for Health Technology Assessment,

Overview of the included tools

As shown in table 2, 18 (37%) tools were published between 2002 and 2010, while 31 (63%) tools were published thereafter. The majority of the tools were checklists (n=30, 61%). For other tools, nine were a checklist combined with summary judgement, 5 were guidelines, 5 were rating scales, and 1 was a questionnaire. Nine of the tools included a summary judgement section to summarize the final judgement of the quality of the appraised tool.
Thirty-one tools were designed as critical appraisal tools, and these tools focused on different aspects in the appraisal of papers (e.g., quality of paper for systematic review, methodology review). Two tools stated they include both critical appraisal items as reporting items. Five tools state only to describe reporting items. Two of the tools state to cover only methodology items, one tool focussed only on the risk of bias assessment.

All the tools were designed for the assessment of NRSI studies or could be used for both randomised as non-randomised study design. Since the tools provided by the Critical Appraisal Skills Programme (CASP)[20], the Joanna Briggs Institute (JBI) [22], and the Scottish Intercollegiate Guidelines Network (SIGN) [30] provided multiple NRSI checklists for different study designs (e.g., cohort and case-control), we defined them as separate tools.

Quality domains and items

Study Objective (item 1)

The item study objective focusses on the research question and hypothesis, and it is not directly linked to a specific risk of bias. However, the study objective forms the foundation of any study. If not clearly described or stated, the study could be more susceptible for bias. This item was used in thirty-one of the included tools, but only one tool was classified as methodology and addressed it on a level 1 (3%) and in five tools it was classified as methodology and reporting both at level 1 (16%). In eighteen tools it was classified as reporting at a level 1 (58%) and in seven as a level 2 reporting (23%).

Protocol (item 2)

The item protocol focusses on all items involved in the protocol, deviations from the protocol could lead to performance bias which should be avoided. The protocol item is addressed by thirteen of the included tools (27%). One tool was classified as methodology level 1 (8%) and three as methodology level 2 (23%). Eight tools were classified as reporting four of level 1 (31%) and four on level 2 (31%). Only one tool (i.e., ROBINS-I) [23] sufficiently addressed both methodology and reporting.

Study design (item 3)

The study design item focusses on whether the authors choose the appropriate study design and the risk of bias per design. Cohort studies are more prone to selection bias, performance bias and detection bias, compared to case-control studies. [29] Twenty-four tools (49%) include the item study design, twelve tools as methodology level 1 (50%), three as a level 2 methodology (13%). Five tools are classified as reporting level 1 (21%) and three on reporting level 2 (12%) of which one included a methodology item on level 1, the tool on reporting level 2 and methodology level 1 is the tool by Thomas [38]. One tool is classified as methodology and reporting on both level 1 (4%).

Ethical approval (item 4)

The item ethical approval is on whether the authors sought ethical approval. Ethical approval is not linked directly to any risk of bias. The item is the least covered item in the tools, it is only represented in four of the 49 tools (8%), three are on a level 1 reporting and one on level 2 reporting.

Sample size calculation/power calculation (item 5)

The item sample size calculation and or power calculation is included how the sample size was calculated and if it was included in the paper. The item is covered by twenty-four of the tools, eight tools cover the item only on methodology, four on level 1 (8%) and four on level 2 (8%). Nine tools cover this item only on reporting level 1 (38%), and three on a reporting level 2 (13%). Two of the tools, the TREND [33] and Heller [53], cover both reporting and methodology on a level 1. The SURE checklist covers the methodology on a level 1 and reporting on a level 2.

Eligibility criteria (item 6)

The eligibility criteria item is covered by thirty-three tools, thirteen tools (38%) focussed on only the methodology, nine on a level 1 and three on a level 2. Of the thirty-three tools sixteen were on reporting, six (18%) on a level 1 and ten (30%) on a level 2. Two tools cover both the methodology and reporting on a level 1, one tool covers the methodology on level 1 and reporting on a level 2, the RELEVANT [17] tool. Only one tool that covers both methodology and reporting on a level 2, the tool developed by Handu [57].

Patients (item 7)

Item seven is on patients; this item covers the selection and definition of patients/ controls. It is covered by twenty-eight of the 49 tools, sixteen tools are on only methodology, eleven (38%) on a level 1 and five on level 2. Nine tools are on reporting with three (10%) tools on level 1 and six (21%) level 2. Three tools (10%) cover the item patients on methodology and reporting both on a level 1.

Participation rate (item 8)

The participation rate is included as a low participation rate could lead to biased results. The item is covered by nineteen tools in total, eight (42%) on methodology level 1, one (5%) on methodology level 2 the NIH-NHLBI tool [26], three (16%) on reporting level 1, four (21%) on reporting level 2, and three (16%) on reporting and methodology level 2.

Baseline characteristics (item 9)

As baseline characteristics summarizes important attributes of the participants to ensure comparability of the groups. The item baseline characteristics is covered by twenty-one tools, eight on methodology level 1 (38%), three on methodology level 2 (19%), two on reporting level1 (10%) and five on reporting level 2(24%). One tool, IHE checklist, [49] is on reporting level 1 and methodology on level 1. One tool covers the baseline characteristics on methodology level 2 and reporting level 1 the NICE checklist. [48]

Selection of intervention (item 10)

Selection of the intervention is an item that focusses on additional interventions, and relevant information on the control arm. The item is covered by only 10 tools in total (20%). Three tools are on methodology level 1 (30%), four on level 2 (50%) and one is on a level 1 reporting (10%). Only one tool covered reporting and methodology on a level 1 and level 1 respectively (10%), this is the SURE guideline.

Definition of intervention (item 11)

Defining the intervention was an item included by twenty-one tools, two of the tools were classified as a methodology level 1 (10%), ten on a level 1 reporting (48%) and seven on a level 2 reporting (33%). Two tools were classified as methodology level 2 and reporting level 1 (10%), the SURE [21] and ROBINS-I [23] tools.

Measurement of intervention (item 12)

The item measurement of intervention is on how the exposure/intervention is measured and if it is completed as intended to minimize bias. The item is covered by twenty-one of the tools and is only classified as a methodology item. Twelve tools are classified as a level 1 (58%) and seven as a level 2 (38%) methodology. Only one tool includes a reporting section besides a methodology is the JBI Case-Control [22], it covers reporting and methodology on a level 1 and level 2 respectively.

Blinding of intervention (item 13)

Blinding of the intervention is included as is can prevent bias. The item is used in nine tools. Seven were on methodology of which three on level 1 (33%), and four on level 2(44%). Two tools were on both on reporting and methodology, the TREND [33] checklist covers the reporting on a level 1 and methodology on a level 2, the MINORS [35] cover both categories on a level 2.

Length of follow-up (item 14)

The length of follow-up should be sufficient to measure an outcome of interest. The item is covered by fifteen tools and is only classified as methodology. Ten of the tools are on a level 1 (63%) and five on a level 2(38%).

Outcome selection (item 15)

The item outcome selection is on how outcomes are being measured and if they are appropriate. The outcome selection is covered by nine tools, four tools cover it as methodology with two tools on level 1 and three on level 2. One tool is on a reporting level 1 and two tools on a level 2, the quality criteria checklist and de ENCePP. [61]

Definition outcome (item 16)

Defining the outcome is a critical step when assessing the risk of bias and is therefore included as an item. Five of the seventeen tools are considered methodology with three on a level 1 (18%) and three on a level 2 (18%). Twelve tools are categorised as reporting, six are on a level 1 (35%), and five on a level 2 (29%).

Measurement Outcome (item 17)

Measuring the outcome is a factor that can cause bias in the form of measurement bias, it refers to systematic or non-random errors that can occur during data collection. The item outcome measurement is an item covered by thirty-four tools The majority of the tools cover this item as a methodology item, fifteen tools on a level 1 (44%) and twelve on a level 2 (35%), only four tools focussed on reporting, two on level 1 (6%) and two on level 2.

Blinding (outcome 18)

Blinding the outcome refers to the blinding of the data collection to minimize observer bias. Blinding of the outcome is covered by twenty-three of the tools. Nineteen of the twenty-three tool items were categorised as methodology, fourteen as a level 1 (61%) and five at a level 2 (22%). Only two tools are classified as a reporting item, the Newcastle-Ottawa Scale [46] on a level 1 (4%) and ISPE [60] on a level 2.

Data source (item 19)

The item data source is included to see where the data is collected and provided. The item is covered by eleven tools, four of the tools are categorized as methodology, three on a level 1 (27%) and one on a level 2 (9%). Six tools are considered on reporting, one tool is on reporting level 1 and five on level 2 (49%). One tool covers both reporting and methodology on a level 1, the GRACE guideline. [25]

Missing data (item 20)

The item missing data is included as it could lead to biased results or low power study results. Missing data is covered by twenty of the included tools, nine tools are on methodology, with four on a level 1 (20%) and five on a level 2 (25%). Seven tools are on a level 1 (35%) reporting and one on a level 2 reporting (5%). Two tools cover reporting and methodology both on a level 1 the SURE guideline [21] and the tool by Blagojevich [41]. The tool developed by Genaidy [42] is categorized as a level 2 reporting and a level 1 on methodology.

Loss to follow-up (item 21)

Loss to follow-up is an item that is present in twenty-one of the tools. Fourteen tools are considered on methodology only, six on level 1 (29%) and eight on level 2(43%). Three of the tools are on reporting level 1 and three on reporting level 2.

Description data analysis (item 22)

Data analysis description focusses on whether the data analysis is valid, relevant, and well described for the purpose of the study. The item is described in twenty-seven tools, twenty tools are considered on methodology of which fifteen are on a level 1 (56%) and six on a level 2 (22%). Seven of the tools are classified as reporting one on a level 1 (4%) and five on a level 2 (19%).

Sensitivity (item 23)

The sensitivity of the data analysis is an important aspect of the data analysis as it plays a central role in the assessment of unmeasured confounders [67]. The item is covered by eleven tools, four on level 1 (36%) methodology and four on level 2 methodology, and three on a level 1 reporting (27%).

Confounding (item 24)

The item confounding is included to address if all the relevant confounders have been assessed. Confounding could lead to results that do not reflect on the actual relationship between variables and the outcome. The item confounding was used in twenty-seven of the tools, fourteen on a level 1 (52%) methodology and eight on level 2 (30%). Three tools covered confounding at a level 1 reporting (11%). Two of the tools cover both methodology and reporting on level 1 (7%), these are JBI Cohort [22] and Weightman [45].

Are all the results included (item 25)

This item is included to assure all results are presented in a clear and proper way. This item is covered by twenty-two tools, of which six tools cover it as methodology three on level 1 (14%) and three on level 2 (14%). Two tools are on a level 1 (9%) reporting, ten on a level 2 (14%) reporting. Heller et al. covers the item as methodology level 1 and the reporting on level 2.

SPIN (item 26)

The item SPIN refers to misleading reporting, interpretation, or extrapolation of study results [AG]. This item is included to assess whether the authors have been honest and clear about their study findings. The item SPIN item is covered by seventeen of the included tools. Thirteen tools are considered on methodology ten on a level 1 (59%) and three on level 2 (24%), three tools are on a level 2 reporting.

Conflict of interest (item 27)

The conflict of interest is included as it may introduce bias when a person or company has an interest in favourable results. This interest may negatively affect the persons or companies by biasing individual judgement. This item is included by fifteen tools, six on methodology level 2 (40%), five on a level 1 (33%) reporting and three on a level 2 (20%) reporting. The critical reading sheet covers both methodology and reporting on a level 1 (7%).

Representation of each domain and item

Figure 2 shows the coverage of all tools on all the critical quality items identified. Among the eight domains, the Background domain was least represented by most quality assessment tools. None of the four items covering the Background domain were sufficiently covered by more than four tools. The least represented item was ethical approval, it is only represented by four tools in total and was only classified as a reporting question.

The items blinding of the intervention and outcome selection were less represented by the included tools as only nine tools covered these items. As figure 2 shows, some of the items are represented more as a methodology item or as a reporting item. However, both aspect matter for the assessment of the risk of bias. The best represented items were outcome measurement, eligibility criteria, and study objective, with thirty-four, thirty-three and thirty tools addressing the items, respectively.

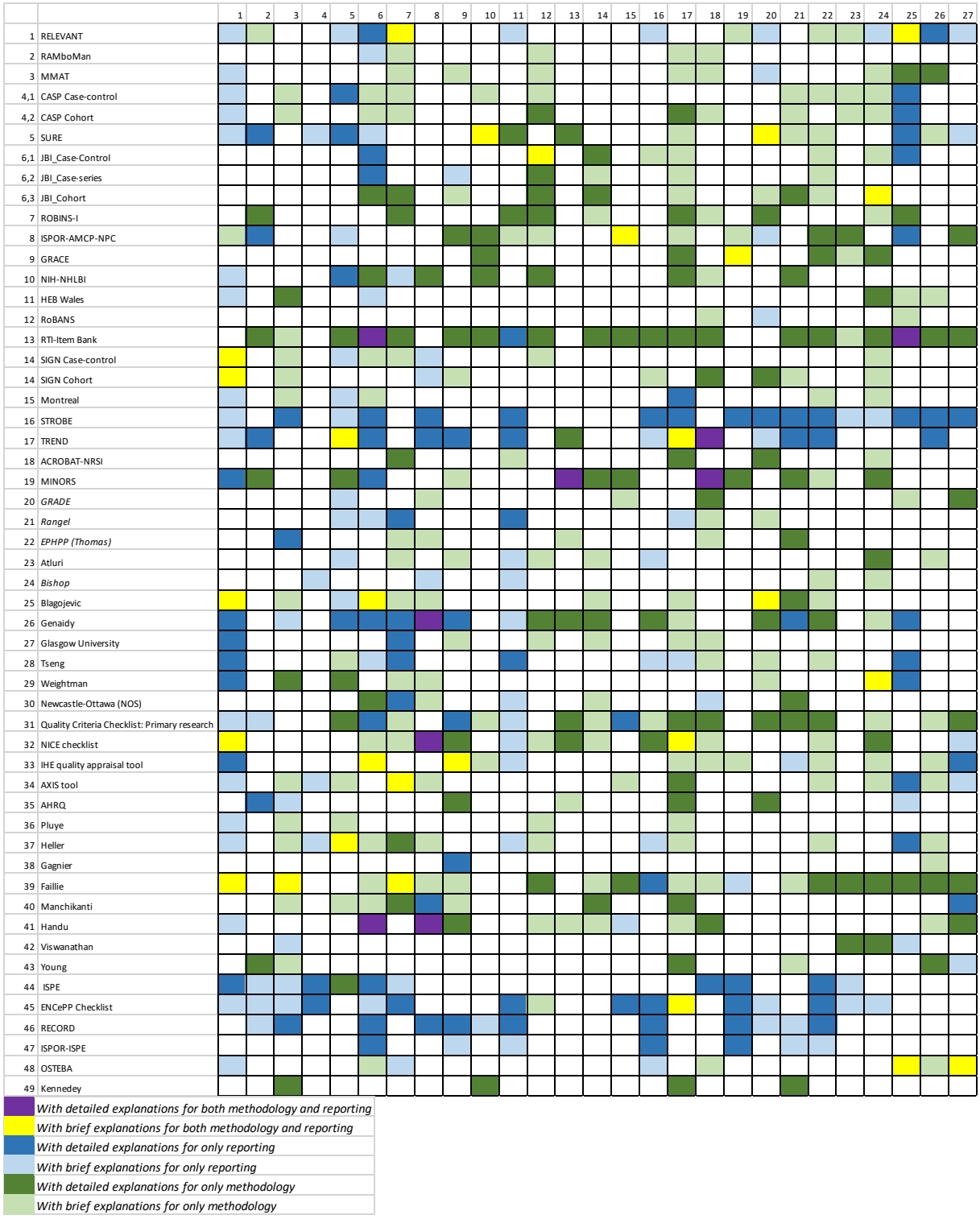

Figure 2. Overview of item coverage of the tools, only the highest classification is represented in the table. Items: **Background**: (1) Study objective, (2) Protocol, (3) Study design, (4) Ethical approval, **Population:** (5) Sample size/power calculation, (6) Eligibility criteria, (7) Patients, (8) Participation rate, (9) Baseline characteristics, **Intervention:** (10) Selection , (11) Definition, (12) Measurement, (13) Blinding, (14) Length of follow-up, **Outcome:** (15) Selection, (16) Definition, (17) Measurement, (18) Blinding, **Data collection:** (19)Data source, (20) Missing data, (21) Loss to follow-up, **Data analysis:** (22) Description, (23) Sensitivity, (24) Confounding, **Results:** (25) Are all the results included, (26) SPIN, **Conflict of interest** (27) Conflict of interest

Review of the tools covering most

Table 3.1 shows the five tools covering the highest number of items on a detailed level (dark green, dark blue, and purple). As we aim to identify the most complete tool, we do not differentiate between methodology and reporting. The RTI-item bank scores best with a coverage of nineteen detailed explanation items. The RTI-item bank mainly focusses on methodology in the assessment. The tool by Genaidy [42] ranks second, covering fourteen items in detail focussing on both methodology as reporting. The third ranking tool is the STROBE [32] checklist covering thirteen items on reporting in detail. On the fourth place comes the MINORS [35] checklist covering eleven items in detail. The fifth tool is the Quality criteria checklist [47], covering eleven items in detail with the majority of the detailed items focusses on methodology. On the basis of detailed coverage, the five tools show less (2 tools or less) or no coverage at all, for the items ethical approval, participation rate, selection of intervention, measurement of intervention, data source, and sensitivity analysis.

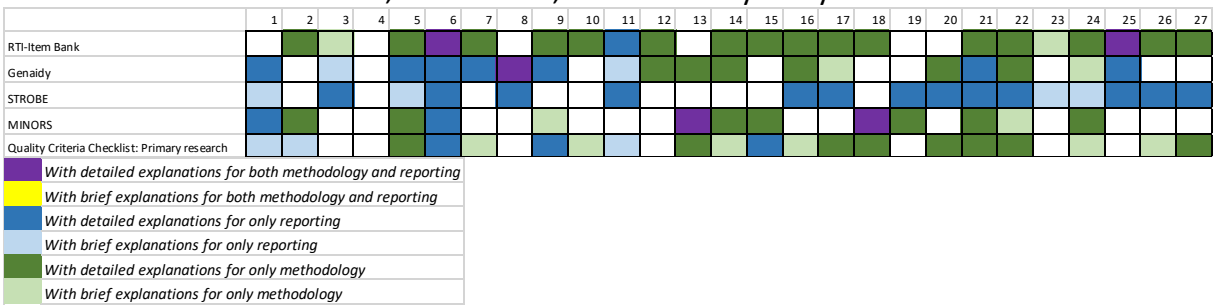

Figure 3.1 The 5 tools covering the most items on detailed level. **Background**: (1) Study objective, (2) Protocol, (3) Study design, (4) Ethical approval, **Population:** (5) Sample size/power calculation, (6) Eligibility criteria, (7) Patients, (8) Participation rate, (9) Baseline characteristics, **Intervention:** (10) Selection , (11) Definition, (12) Measurement, (13) Blinding, (14) Length of follow-up, **Outcome:** (15) Selection, (16) Definition, (17) Measurement, (18) Blinding, **Data collection:** (19)Data source, (20) Missing data, (21) Loss to follow-up, **Data analysis:** (22) Description, (23) Sensitivity, (24) Confounding, **Results:** (25) Are all the results included, (26) SPIN, **Conflict of interest** (27) Conflict of interest

When reviewing the tools on basis of overall coverage and not on the level of coverage, the RTI-item bank [29] covers most items overall, covering nineteen detailed explanation, two brief explanation and only six items are not covered. The second tool covering the most items is the Quality criteria checklist, covering eleven items in detail, nine items are addressed briefly and seven are not addressed. The third tool, is the tool by Faillie [55], covering nine items in detail, eleven are briefly addressed and seven are not addressed. The tool by Genaidy [42] is fourth, covering fourteen items in detail focussing on both methodology as reporting, four brief and nine are not addressed. The fifth tool is the STROBE [32] checklist covering thirteen items on reporting in detail, four items are covered briefly, and ten items are not addressed. Table 3.2 displays tools covering the most items overall. Both table 3.1 and 3.2 show that not every item is covered when ranking the tools, item 4 ethical approval is not addressed by any of the top-ranking tools. The participation rate, blinding of intervention, and data source are just covered by two of the tools in the ranking of tools based on the coverage of items.

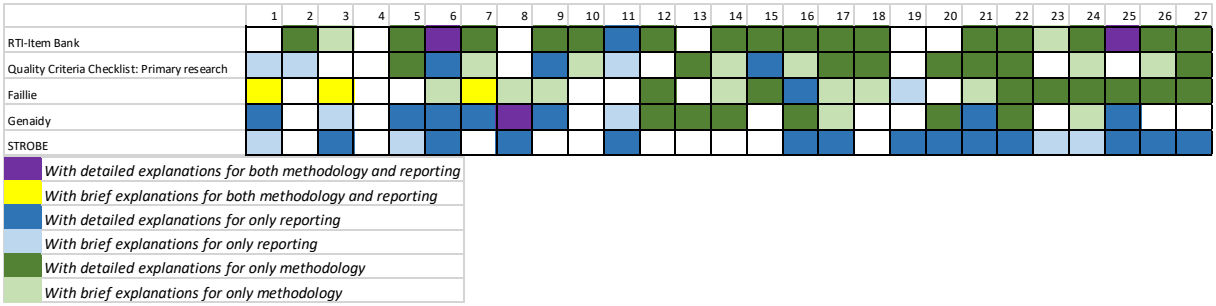

Figure 3.2. The five highest scoring tools on basis of overall coverage for all the items, focussing on all items covered basic (level 1) or detailed (level 2). **Background**: (1) Study objective, (2) Protocol, (3) Study design, (4) Ethical approval, **Population:** (5) Sample size/power calculation, (6) Eligibility criteria, (7) Patients, (8) Participation rate, (9) Baseline characteristics, **Intervention:** (10) Selection , (11) Definition, (12) Measurement, (13) Blinding, (14) Length of follow-up, **Outcome:** (15) Selection, (16) Definition, (17) Measurement, (18) Blinding, **Data collection:** (19)Data source, (20) Missing data, (21) Loss to follow-up, **Data analysis:** (22) Description, (23) Sensitivity, (24) Confounding, **Results:** (25) Are all the results included, (26) SPIN, **Conflict of interest** (27) Conflict of interest

Development "The DRAGON checklist"

As there is no consensus on the preferred tool to use for quality appraisal of NRSI, we designed a new tool that covers the eight domains and twenty-seven items on the highest level with the necessary guidance. We suggest making a new tool based on the included tools in this study, each item is assessed by content and the best aspects were combined to develop the most complete tool on both methodology and reporting.

The proposed tool will be called the Demonstrative Reporting and Appraisal Guideline for Observational and Non-randomised studies (DRAGON). The name "DRAGON" is a reference to the legendary Chinese animal, dragons were described visually as a composite of parts of different animals. Our proposed tool contains different aspects of all the included tools and therefor resembles the mystical animal. We used all the information at hand from the data extraction and constructed the dragon tool by combining and reformulating the items. Each item is shortly addressed, and key elements are indicated below.

Study objective (item 1)

The ISPORE tool [63] covers the item study objective as "Were study hypotheses or goals prespecified a priori?", the SIGN tools [30] adds if the hypothesis and goals are appropriate or not. The IHE quality appraisal tool [49] is used to add a reporting section "Is the hypothesis/aim/objective clearly stated in the abstract, introduction or methods section".

Protocol (item 2)

The RTI-item bank [29] covers the protocol extensively as *"Did execution of the study vary from the intervention protocol proposed by the investigators and therefore compromise the conclusions of the study? [PI: Consider intensity, duration, frequency, route, setting, and timing of intervention/exposures. Drop if not relevant for body of literature.]".* However, the RTI-item bank does not include a section on the publication of the protocol a priori, this could be added by included as other tools propose. These tools provide no guidance in writing the protocol or important aspects, the AHRQ [51] and ROBINS-I [23] include examples as "Specify the review question, participant, experimental intervention, comparator, outcomes, list of potential confounders, list of co-interventions that could differ between groups."

Study design (item 3)

The HEB Wales tool [27] clearly states two points of interest for the study design "*Has an acceptable method been chosen (e.g., interventional without randomisation, before-and after study)? Is the choice of study method appropriate?*". This tool does not provide guidance on what is appropriate, the RTI-item bank [29] provides more guidance "Is the study design prospective, retrospective, or mixed? [Abstractor: 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 analyses data from past records. The question is not applicable to cross-sectional studies.]". The RECORD [62] guideline provides guidance on how to present the study design and relevant information.

Ethical approval (item 4)

In all the tools ethical approval was classified as a reporting item, the ENCePP [61] states it as "Have requirements of Ethics Committee/Institutional Review Board been described? Has any outcome of an ethical review procedure been addressed?". This explanation covers the item well and could not be improved by any additions.

Sample size/power calculation (item 5)

The RTI-item bank [29] provides clear guidance on methodology for the sample size item, it states "*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? [PI: Specify a different percent, if clinically relevant for each outcome of interest. Question relates to precision; 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. PIs 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.]".* The RTI-item bank [63] does not specify how the sample was selected what could lead to bias, the tool by Weightman does include this aspect. Furthermore, it does not address any aspect of reporting, Genaidy [42], provides clear guidance on the reporting.

Eligibility criteria (item 6)

The RTI-item bank [29] provides guidance on both methodology and reporting in a structured and clear manner. The tool focusses on measuring the inclusion/exclusion criteria using a valid and reliable method, and if the inclusion/exclusion criteria are uniformly applied to all comparison groups/arms of the study. The tool further specifies how the inclusion/exclusion criteria should be stated. The RTI-item bank provides clear guidance but lacks certain aspects other tools focus on such as: are inclusion/exclusion criteria appropriate (Blagojevic [41] and Heller [53]), is there selection bias (Montreal [31] and Handu [57]), does this selection bias threaten the external validity of the study (Montreal) [31]. The STROBE [32] and ISPE checklist [60] specify to provide the rationale for the inclusion and exclusion criteria and their impact on the number of subjects. By adding these aspects more guidance is provided to assess the methodology and help with the reporting.

Patients (item 7)

Heller [53] provides questions for the assessment of the methodology, they focus on, appropriateness of the sampling, representativeness of the population, external validity and is the sample relevant. The RTI-item bank [29] asked if the strategy for recruitment was similar across the study. The tool by Heller [53] does not include questions on the reporting, the ENCePP guideline [61] provides guidance in basic information of the patients such as study time period, age and sex, country of origin, disease/indication, duration of follow-up.

Participation rate (item 8)

Genaidy [42] provides some guidance on methodology and reporting, it asks if the participation rate is adequate and if the record of ascertainment is available. This part addresses the methodology but lacks clear guidance on what is considered adequate, the NIH-HBLBI guideline states that at least 50% should be included, the EPHPP [38] guideline gives three options to consider a good participation rate 1)80-100%, 2)60-79%, 3) less than 60%, 4) not applicable and 5) can't tell. The Grade tool [36] states a limit of 80% participants are enrolled. There is no consensus on the exact level of participation rate, the percentage provided by the GRADE [36] tool of 80% should be considered a lower limit. The STROBE checklist [32] provides extra guidance on the number of individuals at each stage of the study for full disclosure on the participation rate.

Baseline characteristics (item 9)

The tool of the Glasgow University [43] states, "Were there clearly defined groups of patients, similar in all important ways other than exposure to the treatment or other causes?", the NICE guideline [48] states "The groups were comparable at baseline, including all major confounding and prognostic factors". None of the tools provide specific guidance on what aspects might be relevant, the reporting question could provide further guidance on the methodology as they give example. Gagnier [54] provides demographic information such as: age, gender, ethnicity, occupation, disease status and family history. These characteristics should be stated.

Selection of intervention (item10)

This item is included to assess the intervention and possible cointerventions across the study groups. The ISPORE-AMCP-NPC tool [24] states "are any relevant interventions missing? This question addressed whether the interventions analysed in the study include ones of interest to the decision maker and whether all relevant comparators have been considered". The quality criteria checklist adds "Are the intensity and duration of the interventions and exposure factor sufficient to produce a meaningful effect?". When selecting an intervention all aspects should be considered and reported in the protocol and final paper.

Definition of intervention (item 11)

Defining the intervention is an item mainly on reporting, the RTI-item bank [29] and TREND statement [33] checklist provides similar and detailed guidance on defining the intervention. The RTI-item bank [29] states, "what is the level of detail in describing the intervention or exposure?" and the TREND [33] statement checklist states, "Details of the intervention intended for each study condition and how and when they were actually administered, specifically including what was given". None of the other tools provide specific examples to include except for the RTI-item bank [29] and the TREND [33] statement checklist. The ROBINS-I

tools [23] included a section "could classification of intervention status have been affected by knowledge of the outcome or risk of outcome?", this focusses on the methodology.

Measurement of intervention (item 12)

The RTI-item bank [29] states "Are interventions/exposures assessed using valid and reliable measures, implemented consistently across all study participants?". This definition is found across most tools, only the ROBINS-I [23] tool adds a section if there were any deviations from the intended intervention beyond what could be expected in usual practice. The JBI case-control tool indicates that the study should clearly describe the method or measurement of exposure.

Blinding of the intervention (item 13)

The NICE guideline [48] states "Participants receiving care were kept 'blind' to treatment allocation. The knowledge of assignment to a particular treatment group may affect outcomes such as a study. Individuals administering care were kept 'blind to treatment allocation". The quality criteria checklist includes the reason for blinding as it could introduce bias. The TREND [33] statement checklist included a small section on how the blinding was accomplished.

Length of follow-up (item 14)

The RTI-item bank [29] provides the profound guidance on the length of follow-up "Is the length of follow-up the same for all groups? Is the length of time following the intervention/exposure sufficient to support the evaluation or primary outcomes and harms?" The MINORS [35] add "is the follow-up period appropriate to the aim of the study, this aspect differs between studies and should be included". The length of follow-up should be clearly stated in the protocol and final paper.

Outcome selection (item 15)

The MINORS [35] states "Endpoints appropriate to the aim of the study", the RTI-item bank [29] specifies is as; does not take other relevant secondary outcomes and harms into consideration. The Quality criteria checklist includes multiple reporting aspects such as, were primary and secondary endpoints described and relevant, were other factors accounted for (measured) that could affect outcomes. The ENCePP [61] includes a section specific for HTA purposes, does the protocol describe specific outcomes relevant for Health Technology Assessment? (e.g., HRQoL, QALYs, DALYS, health care services utilization, burden of disease or treatment, compliance, disease management)

Outcome definition (item 16)

The NICE checklist states "The study used a precise definition of outcome(s)", the RTI-item bank [29] specifies the important outcomes should be pre-specified by the researchers. They often include a section on measurement of the outcome, this will be covered by the item measurement outcome. The STROBE [32] guideline includes a reporting question, "clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers". The RECORD [62] checklist suggests including a complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers.

Outcome measurement (item 17)

The outcome measurement is covered extensively by different tools, the AHRQ [51] states "outcomes are measured using valid and consistent procedures and instruments across all study participants" and "errors in measurement of the outcome are unrelated to the intervention received". The ROBINS-I [23] tool adds "could the outcome measure have been influenced by knowledge of the intervention received? The GRACE guideline [25] states, "was the primary clinical outcome measured objectively rather than subject to clinical judgment (e.g., opinion about whether the patient's condition has improved)?". The STROBE [32] and TREND [33] checklists provide clear guidance, "state the methods used to collect data and any methods used to enhance the quality of measurements. 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.".

Blinding outcome (item 18)

The Quality Criteria Checklist [47] defines the item as "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)". Multiple tools (MMAT [19], CASP [29], ROBINS-I [23], NIH-NHLBI [26], RTI-item bank [29]) state "were outcome assessors aware of the intervention received by study participants". The TREND [33] guideline states, "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", this covers the reporting part completely.

Data source (item 19)

The RELEVANT [17] guideline states, "The data source (or database), as described, contains adequate exposures (if relevant and outcome variables to answer the research question".
The item data source focusses more on reporting as it should be clearly stated it origin of datasets, the ISPE [60] covers potential data sources extensively for reporting purposes.

Missing data (item 20)

The ROBINS-I [23] tool covers the item missing data very well, it states "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 variable needed for the analysis?". The SIGN guideline [30] included percentages of individuals or clusters recruited into each arm. Most tools include to what extent there is missing data and how missing data was handled.

Loss to follow-up (item 21)

The item loss to follow-up is frequent and similar across all items, most tools specify the follow-up rate >80%, meaning loss to follow-up should not exceed 20%. The most frequent formulation of the loss to follow-up item is "Was the follow-up rate over all study groups ≥ 80%?". The MINORS [35] states "loss to follow-up should be less than 5%, all patients should be included in the follow-up. Otherwise, the proportion lost to follow-up should not exceed the proportion expiring the major endpoint". On reporting STROBE [32] provides clear guidance "Report number 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)", The RECORD [62] guideline states, "report the method of how loss to follow-up was handled".

Description of data analysis (item 22)

The RTI-item bank [29] extensively covers the description of data analysis item on both methodology and reporting. It covers the methodology in two questions, "Does the analysis control for baseline differences between groups, are the statistical methods used to assess the primary benefit outcomes appropriate to the data". The JBI case control [22] and Relevant [17] checklists focus on potential confounding factors. The Quality Criteria Checklist [47] focusses on multiple aspects similar as the aforementioned tools, it differs only by including an item "were correct statistical tests used and assumptions of the test not violated?". The STROBE [32] and TREND [33] checklist provide clear guidance on the reporting of the description of data analysis.

Sensitivity data analysis (item 23)

The sensitivity analysis is just covered by a few tools, the level of coverage is for most of the tools just basic. The GRACE [25] tool describes the sensitivity analysis as "Were any meaningful analyses conducted to test key assumptions on which primary results are bases?", this cover the methodology part very well. However, Viswanathan [58] adds "Were processes used to reduce uncertainty in individual judgements such as dual independent assessment of risk of bias with an unbiased reconciliation method", they add "avoid the presentation of risk-of-bias assessment solely as a numerical score; at minimum consider sensitivity analyses of these scores.". The tools covering the sensitivity as a reporting item all state "Describe any sensitivity analyses", the ISPE [60] includes a section to be for the development of the protocol "Any sensitivity analyses should be described. Details of the statistical analysis may be specified later, but before analysis begins".

Confounding (item 24)

The majority of tools covering the item confounding state, "Are confounding factors considered/identified?", the RTI-item bank [29] covers the item more extensive, "Are confounding and/or effect modifying variables assessed using valid and reliable measures across all study participants?" and "Were the important confounding and effect modifying variables taken into account in the design and/or analysis?". The NICE [48] and Quality Criteria Checklist [47] includes a section "Were groups comparable at baseline?", this refers to potential confounding factors. STROBE [32] states for the reporting question "Clearly define all outcome, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable", Weightman [45] adds, "include an explanation of how potential confounding factors have been controlled for".

Are all the results included (item 25)

The HEB Wales guideline [27] states, "Were all important outcomes/results considered?". The GRADE tool [36] adds "Were data reported consistently for the outcome of interest (i.e., no potential selective reporting)?", the addition of the GRADE tool [36] points out the option of selective reporting of the author. On reporting the tool of Genaidy [42] focusses basic characteristics of study participants, adverse effects, and main finding in general. Viswanathan [58] states "Present findings and conclusions transparently, balancing the competing considerations of simplicity of presentation with burden on the reader".

SPIN (item 26)

The RTI-item bank [29] provides some questions regarding SPIN it states, “Are results believable taking study limitations into consideration?”, this item is the essence of the SPIN item but does not provide clear guidance in the assessment. The MMAT [19] states, “Are the findings adequately supported by the results? Is there coherence between qualitative data sources, collection, analysis, and interpretation?”. Young [59] adds “Does the data justify the conclusions?”. For reporting STROBE [32] provides clear guidance “Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies and other relevant evidence.” The RELEVANT [17] guideline adds to include the clinical relevance of the results”.

Conflict of interest (item 27)

The item Conflict of interest is mainly focussed on reporting however, Handu [57] provides a question on methodology for the assessment of bias “Is bias due to study’s funding or sponsorship unlikely?”. The RTI-item bank [29] provides a similar question “Is there bias due to study’s funding or sponsorship?”. As a reporting question conflict of interest is well defined by the STROBE checklist [32] “Is the source of funding and role of the funders for the present study and, if applicable, for the original study on which the present article is based”. None of the other tools covers any different aspect on reporting the conflict of interest.

Table 3. the first version of the DRAGON tool

Domain/Item number	**Signalling question**
Domain 1 Background	
Item 1 Study objective	
M1.1	Was the study objective prespecified before the study was conducted?
M1.2	Was the study objective specific?
M1.3	Was the study objective relevant to the available literature?
R1.1	State the study objective, in terms of population, intervention(s), comparator(s), outcomes, settings (e.g., location and timing), and hypotheses to be evaluated.
R1.2	State the study rationale, i.e., how the research would fill in a research gap.
R1.3	Provide literature that supported relevance of the study objective.
Item 2 Protocol	
M2.1	Was the study protocol published in a journal or registry before the study was conducted?
M2.2	Did implementation of the study vary from the study protocol? If yes, was the variation likely to affect study validity?
R2.1	The study protocol should at least describe the study objective, population, intervention(s), comparators(s), co-intervention(s) that might differ between intervention(s) and comparator(s), and settings.
R2.2	Specify relevant confounders and reasons on why they are relevant.
R2.3	Specify methods to address and present confounders and justification on method selection.
R2.4	Provide information on how the study protocol was updated.
R2.5	Describe deviations from the protocol with reasons.
Item 3 Study design	
M3.1	Were concepts on study design used correctly?

M3.2	Was the study design appropriate to the study objective?
M3.3	Were the methods on data collection and analysis appropriate to the study design?
R3.1	State what study design was used in the title and abstract of the study.
R3.2	State all key elements that were specific to a study design.
R3.3	State how methods on data collection and analysis were relevant to study design
Item 4 Ethical approval	
M4.1	Was ethical approval relevant to the study? If yes, was it received?
R4.1	State relevance of ethical approval.
R4.2	Describe requirements and procedure of the Ethics Committee or Institutional Review Board.
Domain 2 Population	
Item 5 Sample size/power calculation	
M5.1	Was sample size justified, e.g., through sample size or power calculation?
M5.2	Was the sample size adequate to detect a clinically significant difference?
R5.1	Provide sample size or power calculation, with details on effect size, type I or II errors, and number of confounders.
R5.2	Judge magnitude of sample size with reasons.
Item 6 Eligibility criteria	
M6.1	Were eligible criteria prespecified?
M6.2	Was there risk of selection bias? If yes, did the bias threaten representativeness of subjects to the target population?
M6.3	Could all eligible criteria be measured validly and reliably?
M6.4	Were eligible criteria implemented uniformly across intervention groups?
R6.1	Discuss representativeness of included subjects.
R6.2	Discuss risks of omitting criteria that were critical to the study.
R6.3	Provide rationales for eligible criteria.
Item 7 Patients	
M7.1	Was the target population clearly defined, in terms of study time period, age and sex, location, disease, indication, and duration of follow-up?
M7.2	Were all subgroup population included?
M7.3	Were all subjects recruited from the same source population?
R7.1	Define target population, in terms of study time period, age and sex, location, disease, indication, and duration of follow-up.
R7.2	Describe source population, e.g., participating institutions and how subjects were recruited.
Item 8 Participation rate	
M8.1	Was the participation rate adequate (80%) to avoid the non-response bias in each study group? If not, was it sufficiently explained?
R8.1	Report participation rate in each of the groups being studied, and record reasons for non-participation.
Item 9 Baseline characteristics	
M9.1	Were the demographic characteristics, disease status, confounding, and prognostic factors comparable across study groups?
M9.2	Were the study groups selected appropriately (e.g., by restriction)?
R9.1	State the demographic characteristics (e.g., age, gender, ethnicity, occupation), symptoms, medical and psychosocial history, comorbidities, lifestyle, and genetic information.
R9.2	Report the methods used to select study groups.

Domain 3 Intervention	
Item 10 Intervention selection	
M10.1	Were all interventions of interest included?
M10.2	Was the intensity and duration of the intervention sufficient to produce a meaningful effect?
M10.3	If the intensity and duration of an intervention vary, did the study investigate the impact of variety on outcomes?
R10.1	Justify the inclusion of interventions of interests.
R10.2	Report the intensity and duration of interventions.
Item 11 Intervention definition	
M11.1	Were interventions clearly defined?
M11.2	Were interventions defined without knowledge of subsequent outcomes?
R11.1	Provide information on intervention administration, including content, dosage, duration, frequency.
R11.2	Provide codes for classifying interventions, of provide explanations if codes were not available.
R11.3	Provide reference information (e.g., journal articles) that supported intervention definitions.
Item 12 Intervention measurement	
M12.1	Were interventions measured in a standard and objective way?
M12.2	Were interventions measured consistently across all study groups and participants?
M12.3	Did measured interventions deviate from usual practice?
R12.1	Report methods used to measure interventions in all study groups and discuss the method validity.
R12.2	Discuss the risk of deviation of measure interventions from usual practice?
Item 13 Intervention blinding	
M13.1	Were patients blinded to allocation of treatment groups?
M13.2	Were clinicians and investigators blinded to allocation of treatment groups?
R13.1	State how the intervention blinding was accomplished and assessed, or provide reasons if blinding was not possible.
Item 14 Length of follow-up	
M14.1	Was the length of follow-up sufficiently long to investigate a relationship between interventions and outcomes?
M14.2	Was the length of follow-up same for all groups?
R14.1	Report the length of follow-up and explain why the length is sufficient.
Domain 4 Outcome	
Item 15 Outcome selection	
M15.1	Whether the included outcomes were meaningful to the patients the decision makers were concerned with?
M15.2	Were surrogate outcomes avoided?
R15.1	Explain why the selected outcomes are relevant to the study objective.
Item 16 Outcome definition	
M16.1	Whether the primary and secondary outcomes were clearly defined?
M16.2	Were all important outcomes pre-specified?
M16.3	Was the same outcome definition applied for all study groups and patients?
R16.1	Provide outcome definitions with codes. If codes were not available, provide explanations.
Item 17 Outcome measurement	

M17.1	Were outcomes measured using a valid and objective way?
M17.2	Were outcomes measured consistently across study groups?
M17.3	Were errors in measurement of the outcome are unrelated to the intervention received (i.e., no differential misclassification of outcomes)?
R17.1	Describe the methods used to enhance quality of outcome measurements.
R17.2	Describe how validity of outcome measurements (e.g., precision, accuracy, sensitivity, specificity, positive predictive value) was addressed.
R17.3	Describe comparability of measurement methods across study groups.
Item 18 Outcome blinding	
M18.1	If outcomes were not assessed objectively, were data collectors and outcome assessors blinded for outcomes of patients, to avoid detection bias?
M18.2	Were data collectors and outcome assessors blinded for other important confounding and prognostic factors?
R18.1	Specify methods used to ensure blinding of outcome assessment, or provide explanations if blinding was not available.
Domain 5 Data collection	
Item 19 Data source	
M19.1	Were interventions and outcomes available in sufficient details in data sources?
M19.2	Were data sources prespecified before the study was conducted?
R19.1	Describe the name, type (e.g., medical records, questionnaire, etc.; individual data or aggregate data) of data sources, data linkage, and settings where data were collected.
R19.2	Specify data extraction date, data sampling, data format, and data cleaning methods.
Item 20 Missing data	
M20.1	Were outcome data and confounding variables reasonably complete?
M20.2	Were proportion of patients and reasons for missing data similar across study groups?
M20.3	Were robust methods used to address missing data?
R20.1	Describe the extent of missing data, including number and proportion of patients with missing data for each variable of interest.
R20.2	Describe methods used to address missing data.
Item 21 Loss to follow-up	
M21.1	Was the follow-up rate in all study groups at least 80%, or the proportion lost to follow up should not exceed the proportion experiencing the major outcomes.
R21.1	Report number and proportion of patients completing follow-up and calculate the follow-up rate. Also, provide reasons for loss to follow-up.
R21.2	State the method used for addressing loss to follow-up.
Domain 6 Data analysis	
Item 22 Data description	
M22.1	Were statistical analyses appropriate for the study design and type of outcome indicators?
M22.2	Were robust statistical methods used to compare study groups for primary outcomes?
M22.3	Were any meaningful analyses conducted to test key assumptions on which results are based?
R22.1	Describe statistical methods used for comparing study groups for outcomes.
R22.2	Describe statistical methods used to test key assumptions on which results are based.
R22.3	Describe software used for statistical analyses, including software packages, versions, and analytic procedures.
Item 23 Sensitivity	
M23.1	Was statistical uncertainty of the findings (e.g., p values, confidence intervals) evaluated?

M23.2	Were sensitivity analyses or subgroup analyses performed to address uncertainties, e.g., on key assumptions, outcome definitions, loss to follow-up, or level of risk of bias?
R23.1	Prespecify sensitivity analyses before the study was conducted.
R23.2	Describe sensitivity or subgroup analyses and explain what uncertainty the analyses were used to address.
Item 24 Confounding	
M24.1	Were all important confounders identified?
M24.2	Were confounders or effect modifiers accounted for in study design or analysis, e.g., through restriction, stratification, interaction terms, multivariate analysis, propensity score matching, instrumental variables, or other approaches?
M24.3	Was the method for ascertaining confounders equal for all participants?
R24.1	Specify all likely sources of potential confounding and provide definitions.
R24.2	Explain how potential confounders were controlled for.
Domain 7 Results	
Item 25 Are all the results included	
M25.1	Were all important results presented, without potential of selective reporting?
M25.2	Were results presented in an understandable way, e.g., in adequately labelled tables and graphs?
M25.3	Were results consistent in primary and secondary analyses?
M25.4	Were the main findings of the study clearly described?
R25.1	Present findings and conclusions transparently, balancing the competing considerations of simplicity of presentation with burden on the reader.
R25.2	Provide both adjusted and unadjusted results, based on confounding.
R25.3	Provide implications of study results and discuss whether and to what extent study objectives were achieved.
Item 26 Are all conclusions reasonable	
M26.1	Were results believable, after taking study limitations into consideration?
M26.2	Was the conclusion adequately supported by the results and accumulated data?
R26.1	Discuss clinical relevance of results.
R26.2	Interpret results, while take into account study limitations, known information from other studies, potential bias, and multiplicative analyses.
Domain 8 Conflict of interest	
Item 27 Conflict of interest	
M27.1	Was there no apparent conflict of interest, or was there bias due to study's funding or sponsorship?
R27.1	Specify the funding sources, role of funders, and competing interest.

Discussion

We conducted a literature review of quality assessment tools for non-randomised studies of interventions and assessed the level of coverage for both methodology and reporting, based on 27 items in 8 domains. We identified 49 quality assessment tools that showed great variability in item coverage in different tools. Based on level of coverage, the best tools are the RTI-item bank, Genaidy et al. (2007), STROBE, MINORS, and the Quality Criteria checklist. However, none of the included 49 tools covered all the 27 items. The least covered items included: ethical approval, blinding intervention, and outcome selection. Furthermore, this study aimed to clarify the recommendations for the use of quality assessment tools given by different European HTA agencies. Only two HTA agencies, i.e., the NICE and EUnetHTA, provided a clear recommendation on which appraisal tool to use. A consensus should be reached on what tool to recommend, and this should be clearly stated by all HTA agencies.

To our knowledge, this review is the first thematic analysis that compared the level of coverage on pre-specified items. Previous reviews compared quality assessment tools for NRSI but from a different perspective, or they only tried to identify tools. Quigley et al. 2018, aimed to identify commonly used tools to assess bias in NRSs and to determine those recommended by HTA agencies. One of our aims was similar but we focussed on European HTA agencies. D'Andrea al. 2021 addressed whether critical elements that influence the validity of NRS findings for comparative safety and effectiveness of medication. D'Andrea al. 2021 evaluated tools using 8 prespecified critical domains. They found that most tools evaluated methods for selecting study participants, measurement of exposure and outcome, and measurement and control for confounders. These findings were consistent with our findings. Due to the differences in domains and items, we recommend different tools of choice as they represent our domains and items more.
We found that the coverage of items differed significantly among the tools, and the RTI-item bank covered the highest number of items. The success of the RTI-item bank might be due to the origin of the tool, as when developing the tool, the authors used the recommendations of quality domains and items of Deeks et al. 2003 [13]. In addition, our study implied that none of the tool's covered all the items on its own or on a sufficiently high level, and therefore, it is hard to recommend one tool of choice. We found tools covering critical quality items just on a basic level could provide useful insight and should not be discarded as inferior. All the items from the included tools contributed to the development of the DRAGON tool. The proposed DRAGON tool might provide a solution as it combines various aspects from all the tools and covers all the important items.

Using the DRAGON tool

The DRAGON tool is a tool that can be used in multiple ways. Firstly, it can be used for quality assessment purposes by a reviewer. Secondly, it can be used for designing study methods, by taking key elements and points of concern into consideration. The reporting questions could be used by authors to address reporting biases, when writing an article. When conducting the critical appraisal, each item should be included, and each question included in the DRAGON needs to be answered, otherwise there is potential risk of bias.

Implications

As stated before, HTA agencies should come to an agreement on what tools to use and give clear recommendations. The EUnetHTA is a consortium of European HTA agencies led by the Zorg Institute Netherlands (ZIN). They are in the ideal position to provide uniformly guidance on the tools, for the European Union, and should come to an agreement on the preferred tools. This review gives a selection of tools that could be used but it should be noted that not every critical quality item is addressed by the selection of tools. The proposed DRAGON tool offers a complete guidance on all items on both methodology and for the reporting of NRSIs. Before the DRAGON tool can be used it must be rigorously tested and adjustments made accordingly. If the HTA agencies prefer a selection of tools the DRAGON tool is one of the candidates for this selection as it covers all the critical quality items.

Limitations

Our study had some limitations. Firstly, we only conducted two rounds of the snowballing. If we had continued, we might have identified more reviews on quality assessment tools. Secondly, we only focused on tools published or updated as of 2002. Older tools still might provide complete quality assessment tools. Thirdly, only one reviewer collected the most recent versions of the tools mentioned. Consequently, some updated versions of tools might be missing, resulting in underperformance of some of the tools. Fourthly, we extracted data and classified items as detailed (level 2) or briefly covered (level 1), and this type of classification is not entirely objective. A more experienced reviewer would classify items on a different level compared to a non-experienced reviewer. The two reviewers (MH and JL) reached a consensus on all fronts before continuing.

The proposed Dragon tool might be a more complete tool but is still not finished. The tool should be finalized by making a user guide and editing the lay-out to a more user-friendly format. Afterwards the tool need be tested in several steps, including face validity testing, cognitive testing, content validity testing and interrater reliability testing. [29] Adjustments to the proposed DRAGON tool can be made in any step of the process. Afterwards, the ease of use may be discussed by experienced reviewers.

Conclusion

None of the existing tools for assessing quality of NRSI studies were able to address all the important quality items. Ethical approval, blinding intervention, and outcome selection were rarely addressed by existing tools. We developed a quality assessment tool (DRAGON) by combining the criteria or signalling questions of all the included tools. The proposed DRAGON tool could cover all important quality items, but its validity needs to be tested in the future research. Also, consensus is needed among HTA agencies on selection of these tools.

References

1. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W. What is real-world data? A review of definitions based on literature and stakeholder interviews. Value in Health. 2017 Jul 1;20(7):858-65.
2. Johnson L.J., Principles and Practice of Clinical Research. Chapter 17 - Design of Observational Studies. U.S. Food and Drug Administration, Silver Spring, MD, United States. 3 November 2017, Version of Record 3 November 2017. doi.10.1016/B978-0-12-849905-4.00017-4
3. Hogervorst, M. A., Pontén, J., Vreman, R. A., Mantel-Teeuwisse, A. K., & Goettsch, W. G. (2022). Real World Data in Health Technology Assessment of Complex Health Technologies. Frontiers in pharmacology, 297.
4. RWE Navigator, Putting real-world healthcare data to work [Internet]. [Place unknown; publisher unknown]; Cited 12-01-2023. Available from URL:https://rwe-navigator.eu/
5. Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.
6. Higgins JPT, Altman DG, Sterne JAC (editors). Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Churchill R, Chandler J, Cumpston MS (editors), Cochrane Handbook for Systematic Reviews of Interventions version 5.2.0 (updated June 2017), Cochrane, 2017. Available from www.training.cochrane.org/handbook.Catalogue of bias collaboration. Bankhead CR, Spencer E, Nunan D. Information bias. In: Sackett Catalogue of Biases 2019. https://catalogofbias.org/biases/information-bias/
7. Sterne JAC, Hernán MA, McAleenan A, Reeves BC, Higgins JPT. Chapter 25: Assessing risk of bias in a non-randomized study. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.3 (updated February 2022). Cochrane, 2022. Available from www.training.cochrane.org/handbook.
8. Cumpston M, Lasserson T, Chandler J, Page MJ. Chapter III: Reporting the review. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.3 (updated February 2022). Cochrane, 2022. Available from www.training.cochrane.org/handbook.
9. 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.
10. Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. Rayyan — a web and mobile app for systematic reviews. Systematic Reviews (2016) 5:210, DOI: 10.1186/s13643-016-0384-4.
11. Wohlin, C. (2014, May). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In Proceedings of the 18th international conference on evaluation and assessment in software engineering (pp. 1-10).
12. 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.

13. Faria, R., Hernandez Alava, M., Manca, A., Wailoo, A.J. NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness for Technology Appraisal: Methods for comparative individual patient data. 2015. Available from http://www.nicedsu.org.uk
14. Connected papers [Internet]. [place unknown: publisher unknown]; [updated date (unknown); cited January 12, 2023]. Available from URL: https://www.connectedpapers.com/
15. INAHTA International Network of Agencies for Health Technology Assessment [Internet]. Canada: Edmonton Canada; [updated date (unknown); cited January 12,2023]. Available from URL: https://www.inahta.org/
16. Deeks JJ., Dinnes J., D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health tehcnol Assess 2003;**7**(27)
17. 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.
18. School of Population Health. EPIQ (effective practice, informatics, and quality improvement). faculty of medical and health sciences, University of Auckland, 2019. Available:https://www.fmhs.auckland.ac.nz/en/soph/about/ourdepartments/epidemiology-and-biostatistics/research/epiq.html [Accessed April 2022].
19. Hong QN, Pluye P, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, Gagnon M-P, Griffiths F, Nicolau B, O'Cathain A, Rousseau M-C, Vedel I. Mixed Methods Appraisal Tool (MMAT), version 2018. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada.
20. Critical Appraisal Skills Programme (2018). CASP (Cohort Study,) Checklist. [Online]Available at: URL. https://casp-uk.net/casp-tools-checklists/ Accessed: April 2022.
21. Specialist Unit for Review Evidence (SURE) 2018. Questions to assist with the critical appraisal of randomised controlled trials and other experimental studies available at: http://www.cardiff.ac.uk/specialist-unit-for-review-evidence/resources/critical-appraisal-checklists [Accessed April 2022]
22. The Joanna Briggs Insitute. Critical Appraisal tools: JBI'[s critical appraisal tools assist in assessing the trustworthiness, relevance ans results of published papers, 2020. Available from URL: https://jbi.global/critical-appraisal-tools [Accessed April 2022]
23. 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.
24. 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.
25. 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.

26. The National Institue of Health and The National Heart, Lung, and Blood Institute (NIH-NLHBI). Study Quality assessment tools. Available from URL: https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools [Accessed: April 2022]
27. 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.
28. Kim SY, Park JE, Lee YJ, et al. Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. J Clin Epidemiol 2013; 66:408–14.
29. 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.
30. Scottish Intercollegiate Guidelines Network. Available: https://www.sign.ac.uk/checklists- and-notes.html [Accessed April 2022]
31. University of Montreal. Critical appraisal Worksheet. University of Montreal, 2011. Available: https:// guides. bib. umontreal. ca/ uploads/ uploads/ original/ critical- appraisal- worksheet. pdf? 1296211861 [Accessed April 2022].
32. The STROBE Statement. Strengthening the reporting of observational studies (cohort, case-control, and cross-sectional), 2007. Available: https://www. strobe- statement. org/ index. php? id=strobe- aims [Accessed April 2022].
33. 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.
34. 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.
35. 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.
36. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Cjulbegovic B, Falck-Ytter Y,Norris SL, Williams Jr JW, Atkins D, Meerpohl J,Schünemann HJ. GRADE guidelines: 4. Rating the quality of evidence-stucy limitations (risk of bias). Journal of Clinical Epidemiology 64 (2011) 407-415
37. Rangel SJ, Kelsey J, Colby CE, Anderson J, Moss RL. Development of a quality assessment scale for retrospective clinical studies in pediatric surgery. J Pediatr Surg 2003;38:390–96.
38. Thomas, B.H., Ciliska, D., Dobbins, M. and Micucci, S. (2004), A Process for Systematically Reviewing the Literature: Providing the Research Evidence for Public Health Nursing Interventions. Worldviews on Evidence-Based Nursing, 1: 176-184. https://doi.org/10.1111/j.1524-475X.2004.04006.x
39. 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.
40. 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.

41. 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.
42. 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.
43. Glasgow University. Critical Appraisal checklist for an article on harm or causation. [Internet] https://www.colleaga.org/tools/harm-critical-appraisal-worksheet Unpublished checklist. [Accessed April 2022].
44. 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.
45. Weightman, A. L., Mann, M. K., Sander, L., & Turley, R. L. (2004). Questions to assist with the critical appraisal of an observational study e.g., cohort, case-control, cross-sectional. Health Evidence Bulletins – Wales, 2009
46. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. Quality Assessment Scales for Observational Studies. Ottawa Health Research Institute, 2004.
47. Academy of Nutrition and Dietetics. Quality criteria checklist: primary research. Chicago: Academy of Nutrition and Dietetics; [date unknown]. https://www.andeal.org/vault/2440/web/files/QCC_3.pdf. [Accessed April 2022].
48. NICE. The guidelines manual. Appendix B: methodology checklist: systematic reviews and meta-analyses. London: NICE; 2012. https://www.nice.org.uk/process/pmg6/resources/the-guidelines-manual-appendicesbi-2549703709/chapter/appendix-b-methodology-checklist-systematicreviews-and-meta-analyses. [Accessed April 2022].
49. Moga C, Guo B, Schopflocher D, Harstall C. Development of a Quality Appraisal Tool for Case Series Studies Using a Modified Delphi Technique. Edmonton AB: Institute of Health Economics. 2012.
50. 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.
51. 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.
52. 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.
53. 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.
54. Gagnier 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. Young JM, Solomon MJ. How to critically appraise an article. Nature Clinical Practice Gastroenterology & Hepatology. 2009 Feb;6(2):82-91.
60. Public Policy Committee, International Society of Pharmacoepidemiology. Guidelines for good pharmacoepidemiology practice (GPP). pharmacoepidemiology and drug safety. 2016 Jan;25(1):2-10.
61. The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Guide on Methodological Standards in Pharmacoepidemiology (Revision 10). EMA/95098/2010.
62. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. (2015) The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Med 12(10): e1001885. doi:10.1371/journal.pmed.1001885
63. 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.
64. López de Argumedo M, Reviriego E, Gutiérrez A, Bayón JC. *Actualización del Sistema de Trabajo Compartido para Revisiones Sistemáticas de la Evidencia Científica y Lectura Crítica* (Plataforma FLC 3.0). Ministerio de Sanidad, Servicios Sociales e Igualdad. Servicio de Evaluación de Tecnologías Sanitarias del País Vasco; 2017. Informes de Evaluación de Tecnologías Sanitarias: OSTEBA.
65. 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.
66. *(Reeves B. Systematic reviews incorporating evidence from nonrandomized study designs: reasons for cations when estimating health effects.)*
67. Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology. 2011 Jan;22(1):42-52. doi: 10.1097/EDE.0b013e3181f74493. PMID: 21052008; PMCID: PMC3073860.

Supplemental material

Table 1 Included papers snowballing process

	Authors	Title	Publication date
1	D'Andrea et al.	How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools.	March 2021
2	Quigley et al.	Critical appraisal of nonrandomized studies—a review of recommended and commonly used tools. Journal of Evaluation in Clinical Practice.	February 2019
3	Faria et al.	NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness for Technology Appraisal: Methods for comparative individual patient data.	(unknown) 2015
4	Sanderson et al.	Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography	April 2007
5	Deeks et al.	Evaluating non-randomised intervention studies.	(Unknown) 2003
6	Jarde et al.	Methodological quality assessment tools of non-experimental studies: A systematic review	May 2012
7	Farrah et al.	Risk of bias tools in systematic reviews of health interventions: an analysis of PROSPERO-registered protocols	November 2019
8	Ma et al.	Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is beter.	February 2020
9	Lohr et al.	Rating the strength of scientific evidence: relevance for quality improvement programs	February 2004
10	Crowe et al.	A review of critical appraisal tools show they lack rigor: Alternative tool structure is proposed	January 2011
11	Patole et al.	Systematic Reviews and Meta-Analyses of Non-randomised Studies	June 2021
12	Losilla et al.	Three risk of bias tools lead to opposite conclusions in observational Research synthesis	September 2018
13	Page et al.	Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review	January 2018
14	Waddington et al.	Quasi-experimental study designs seriesdpaper 6: risk of bias assessment	September 2017
15	Saunders et al.	Assessing the Methodological Quality of Nonrandomized Intervention Studies	March 2003
16	Brand et al.	Research Pearls: Checklists and Flowcharts to Improve Research Quality	July 2020
17	Lundh et al.	Systematic review finds that appraisal tools for medical research studies address conflicts of interest superficially.	April 2020
18	Yao et al.	Clinical research methods for treatment, diagnosis, prognosis, etiology, screening, and prevention: A narrative review	May 2020
19	Liebherz et al.	How to assess the quality of psychotherapy outcome studies: A systematic review of quality assessment criteria	September 2016
20	Tate et al.	Use of Reporting Guidelines in Scientific Writing: PRISMA, CONSORT, STROBE, STARD and Other Resources	May 2011

Table 2 Updated search strategy

Items	N.	Terms
Critical appraisal tool	#1	"critical" [All Fields] AND "appraisal" [All Fields] AND "tools" [All Fields]
	#2	"critical" [All Fields] AND "appraisal" [All Fields]
	#3	("critical" [All Fields] AND "review" [All Fields]) OR "critical review" [All Fields]) AND form [All Fields]
	#4	("systematic review" [Publication Type] OR "systematic reviews as topic" [MeSH Terms] OR "systematic review"[All Fields]) AND form [All Fields]
	#5	appraisal [All Fields] AND ("research design" [MeSH Terms] OR ("research" [All Fields] AND "design" [All Fields]) OR "research design" [All Fields] OR ("research" [All Fields] AND "methodology" [All Fields]) OR "research methodology"[All Fields])
	#6	("research design" [MeSH Terms] OR ("research" [All Fields] AND "design" [All Fields]) OR "research design" [All Fields]) AND ("review" [Publication Type] OR "review literature as topic" [MeSH Terms] OR "review"[All Fields])
Study reporting tool	#7	"study" [All Fields] AND "reporting" [All Fields] AND "tool" [All Fields]
	#8	"study" [All Fields] AND "reporting" [All Fields]
	#9	"reporting" [All Fields] AND "form" [All Fields] AND ("Studies"[Journal] OR "studies"[All Fields])
	#10	"reporting" [All Fields] AND ("Studies"[Journal] OR "studies"[All Fields])
Tool	#11	"checklist" [MeSH Major Topic] OR "scale*" [Title/Abstract]
	#12	"surveys and questionnaires"[MeSH Major Topic] OR "questionnaire*" [Title/Abstract]
	#13	("tool*" [All Fields] OR "instrument*" [All Fields] OR "checklist*" [All Fields] OR "questionnaire*" [All Fields]) AND ("quality" [All Fields] OR "method*" [All Fields] OR "bias" [All Fields])
Study design	#14	"cohort studies"[MeSH Terms] OR cohort studies [Text Word] OR cohort stud* [All Fields]
	#15	"case-control studies" [MeSH Terms] OR case-control studies [Text Word] OR case control stud* [All Fields]
	#16	Non [All Fields] AND ("random allocation"[MeSH Terms] OR randomized [Text Word]) AND stud* [All Fields]
Systematic review	#17	"systematic review" [Publication Type] OR "systematic reviews as topic"[MeSH Terms] OR "systematic review"[All Fields]
Filters	#18	"humans"[MeSH Terms]
	#19	"Review" [ptyp] OR "systematic" [sb]
	#20	("2019/11"[Date - Publication] : "2022/04"[Date - Publication])
Strings		
1st search – tools*	#21	(#1 OR #2 OR #3 OR #4 OR #5 OR #6) AND (#14 OR #15 OR #16) AND #18 AND #20
	#22	(#7 OR #8 OR #9 OR #10) AND (#14 OR #15 OR #16) AND #18 AND #20
	#23	(#11 OR #12 OR #13) AND (#14 OR #15 OR #16) AND #18 AND #20
2 nd search - systematic reviews of tools*	#24	(#1 OR #2 OR #3 OR #4 OR #5 OR #6) AND (#14 OR #15 OR #16) AND #18 AND (#17 OR #19) AND #20
	#25	(#7 OR #8 OR #9 OR #10) AND (#14 OR #15 OR #16) AND #18 AND (#17 OR #19) AND #20
	#26	(#11 OR #12 OR #13) AND (#14 OR #15 OR #16) AND #18 AND (#17 OR #19) AND #20

Updated search string of the study by D'Andrea et al (2021), we added filter #20 to review papers publicized between November 2019 and April 2022. The search has been reproduced in Medline.

Table 3 Data extraction

Study objective (Background) Item 1

	Tool		Michiel	Li
1	RELEVANT	Clearly stated research question	R1	R1
3	MMAT	Are there clear research questions?	R1	R1
4	CASP Case-Control	Did the study address a clearly focused issue?	R1	R1
4	CASP cohort	Did the study address a clearly focused issue?	R1	R1
5	SURE	Does the study address a clearly focused question/hypothesis	R1	R1
8	ISPORE-AMCP-NPC	Were the study hypotheses or goals prespecified a priori?	M1	M1
10	NIH-NHLBI	Was the research question or objective in this paper clearly stated?	R1	R1
11	HEB Wales	Does the paper address a clearly focused issue?	R1	R1
14	SIGN Case control	The study addresses an appropriate and clearly focused question	R1&M1	R1 & M1
14	SIGN cohort	The study addresses an appropriate and clearly focused question.	R1&M1	R1 & M1
15	Montreal	What is the research question?	R1	R1
16	STROBE	- State specific objectives, including any prespecified hypotheses	R1	R1
17	TREND	Specific objectives and hypotheses	R1	R1
19	MINORS	A clearly stated aim: the question addressed should be precise and relevant in the light of available literature	R2	R2
25	Blagojevic	Clearly defined and appropriate study objective	R1&M1	R1 & M1
26	Genaidy	Is the hypothesis/aim/objective of the study clearly described?	R2	R2
27	Glasgow University	Is there a clearly focused question? Consider • Patients • Exposure • Outcome	R2	R2
28	Tseng	Specific objectives or hypotheses stated (i.e. broadly outlined method for comparison indicated)?	R2	R2
29	Weightman	Does the paper address a clearly focused issue? In terms of: aims of the investigation? setting (location and dates)? the population studied? the variables measured?	R2	R2
31	Quality criteria checklist	Was de research question clearly stated?	R1	R1
32	NICE	The study addresses an appropriate and clearly focused question.	R1&M1	R1 & M1
33	IHE quality appraisal	Is the hypothesis/aim/objective of the study clearly stated in the abstract, introduction or methods section?	R2	R2
34	AXIS	Were the aims/objectives of the study clear?	R1	R1
36	Pluye	Qualitative objective or question	R1	R1
37	Heller	Is the research question and/or hypothesis stated clearly?	R1	R1
39	Faillie	Are study objectives clearly specified and appropriate?	R1&M1	R1 & M1
41	Handu	Was the research question clearly stated?	R1	R1
44	ISPE	a statement of research objectives, specific aims, and rationale; 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	R2	R2

		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	R1
49	Critical reading sheet	Describe the objectives of the study Is the study based on a clearly defined research question?	R1	R1

Protocol (Background) Item 2

	Tool		Michiel	Li
1	Relevant	Evidence of a priori design, e.g. protocol registration in a dedicated website	M1	M1
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	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	R2
8	ISPORE-AMCP-NPC	Was there evidence that a formal study protocol including an analysis plan was specified before executing the study? *(for more details on the item we refer to the original tool)*	R2	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? [PI: Consider intensity, duration, frequency, route, setting, and timing of intervention/exposures. Drop if not relevant for body of literature.]	M2	M2
17	Trend	Description of protocol deviations from study as planned, along with reasons	R2	R2
19	MINORS	Prospective collection of data: data were collected according to a protocol established before the beginning of the study	M2	M2
31	Quality Criteria Checklist	In RCT or other intervention trial, were protocols described for all regimens studied?	R1	R1
35	AHRQ methodological checklist	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) • 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	R2	R2
43	Young	Deviations from the planned protocol can affect the validity or relevance of a study. *(for more details on the item we refer to the original tool)*	M2	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. *(for more details on the item we refer to the original tool)*	R1	R1
45	ENCEPP	Ethical considerations, as described in Chapter 14. The study protocol should also explain how the results will be interpreted, avoiding misuse of p-values and statistical significance (see Chapter 4.1).	R1	R1
47	RECORD	Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code	R1	R1

Study design (Background) Item 3

	Tool		Michiel	Li
4	CASP	Did the authors use an appropriate method to answer their question?	M1	M1
11	HEB Wales	- Has an acceptable method been chosen (eg interventional without randomization, before-and after study)? - Is the choice of study method appropriate?	M2	M2
13	RTI	Is the study design prospective, retrospective, or mixed? [Abstractor: 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	M1
14	SIGN Case-Control	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	M1
14	SIGN Cohort	Is the paper really a cohort study? If in doubt, check the study design algorithm available from SIGN and make sure you have the correct checklist.	M1	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	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	R2
22	Thomas	STUDY DESIGN 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)) 6 Interrupted time series 7 Other specify _____ 8 Can't tell Was the study described as randomized? If NO, go to Component C. No Yes If Yes, was the method of randomization described? (See dictionary) No Yes If Yes, was the method appropriate? (See dictionary) No Yes	R2&M1	R2 & M1
25	Blagojevic	Prospective study design	M1	M1
26	Genaidy	Is the study design clearly described?	R1	R1
29	Weightman	- 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	M2
34	AXIS	Was the study design appropriate for the stated aim(s)?	M1	M1
35	AHRQ	Determine study design of each (individual) study	R1	R1
36	Pluye	Appropriate qualitative approach or design or method	M1	M1
37	Heller	What is the study type? Is the study type appropriate for the research question? Is there a comparison group?	M1	M1
39	Faillie	Is study design clearly specified and appropriate?	R1&M1	R1 & M1
40	Manchikanti	Ranking different study designs on their strengths (points) Case report (0)	M1	M1

		Retrospective cohort (1) Prospective cohort (2) Prospective Case control (3) Prospective controlled, nonrandomized (4)		
42	Viswanathan	Determine study design of each (individual) study	R1	R1
43	Young	Was the study design appropriate for the research question?	M1	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	R1
45	ENCEPP	Is the study design described? (e.g. cohort, case-control, cross-sectional, other design)	R1	R1
47	RECORD	Present key elements of study design early in the paper. 4.a: Include details of the specific study design (and its features) and report the use of multiple designs if used. 4.b: 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	R2
50	Kennedey	*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.* *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.*	*M2*	*M2*

Ethical approval (Background) Item 4

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

Sample size/Power calculation (population) Item 5

	Tool		Michiel	
1	Relevant	Sample size/Power pre-specified	R1	R1
4	CASP Case Control	- Was there a power calculation - was there a sufficient number of cases selected - was there a sufficient number of controls selected	R2	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	R2&M1
8	ISPORE-AMCP-NPC	Were sample size and statistical power to detect difference addressed?	R1	R1
10	NIH-NHLBI	Was a sample size justification, power description, or variance and effect estimates provided?	R2	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? [PI: Specify a different percent, if clinically relevant for each outcome of interest. Question relates to precision; 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. PIs 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	M2
14	SIGN Case Control	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	R1
15	Montreal	Was the sample size adequate to detect a clinically/socially significant result?	R1	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	R1
17	TREND	How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules	R1&M1	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	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	R1
21	Rangel	Can the number of surgeons who participated in the study be determined?	R1	R1
23	Atluri	- Sample size justification - Power calculation provided	R1	R1
25	Blagojevic	Sample size calculation given or ~20 subjects per variable included in multivariate analysis	R1	R1
26	Genaidy	Are sample size calculations performed and reported? Yes – Clearly described Y . Calculations are performed, and, all details are reported for effect size, type I or II errors and number of confounders.	R2	R2

		Partial – Somewhat described P . Calculations are performed, and, not all details are reported. No – Not described N . No mention of any calculations.		
28	Tseng	Calculation to justify sample size?	M1	M1
29	Weightman	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	M2
31	Quality Criteria Checklist	If negative findings, was a power calculation reported to address type 2 error?	M2	M2
34	AXIS	Was the sample size justified?	M1	M1
36	Pluye	Appropriate sampling and sample	M1	M1
37	Heller	Was sample size/power calculated and appropriate?	R1&M1	R1 & M1
40	Manchikanti	Sample Size Less than 100 participants without appropriate sample size determination (0) At least 100 participants in the study without appropriate sample size determination (0) 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	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. For safety studies, it may be useful to specify the sample size that can minimally detect a pre-specified risk with a pre-specified power, for example, “the study has an 80% power to detect a relative risk of 3 or greater for drug x compared with treatment with other drugs commonly used in this condition.”	M2	M2

Eligibility criteria (Population) Item 6

	Tool		Michiel	Li
1	Relevant	- Population justified - Flow chart explaining all exclusions and individuals screened or selected at each stage of defining the final sample	R2&M1	R2 & M1
2	RAMboMan	Eligible population recruitment process	R1	R1
4	CASP Case Control	Were the cases recruited in an acceptable way? Were the controls selected in an acceptable way?	M1	M1
4	CASP Cohort	Was the cohort recruited in an acceptable way?	M1	M1
5	SURE	Population/Problem? Can you identify the setting & eligibility criteria?	R1	R1
6	JBI Case series	Were there clear criteria for inclusion in the case series? Did the case series have complete inclusion of participants? *(for more details on the item we refer to the original tool)*	R2	R2
6	JBI Case-Control	Were the same criteria used for identification of cases and controls? *(for more details on the item we refer to the original tool)*	R2	R2
6	JBI Cohort	Were the two groups similar and recruited from the same population?	M2	M2
10	NIH-NHLBI	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	M2
11	HEB Wales	Are the inclusion/exclusion criteria given?	R1	R1
13	RTI-item bank	Are critical inclusion/exclusion criteria clearly stated (does not require the reader to infer)? [Principal Investigator (PI): 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] [Abstractor: Use "Partially" if only some criteria are stated or if some criteria are not clearly stated (corresponding to directions provided by the PI). 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? [PI: 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? [PI: Drop question if not relevant to entire body of evidence (e.g., all case-series, singlearm studies).]	R2&M2	R2&M2
14	SIGN Case Control	The same exclusion criteria are used for both cases and controls.	M1	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?	M1	M2
16	STROBE	(a) Cohort study—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up Case-control study—	R2	R2

		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 (b) 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		
17	TREND	Eligibility criteria for participants, including criteria at different levels in recruitment/sampling plan (e.g., cities, clinics, subjects)	R2	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	R2
21	Rangel	Are selection and/or exclusion criteria for cases clearly stated?	R1	R1
25	Blagojevic	- Inclusion and exclusion criteria are clear and appropriate - Representative sample e.g., general population sample should not exclude subgroups	R1&M1	R1 & M1
26	Genaidy	Are the eligibility criteria for subject selection clearly described? Yes – Clearly described Y . 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 P . Criteria are not clearly described. No – Not described N . Criteria are not described.	R2	R2
28	Tseng	Are selection and/or exclusion criteria for cases clearly stated?	R1	R1
30	Newcaslte-Ottawa (NOS) (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 (eg surgical records) ¯ b) structured interview ¯ c) written self report d) no description 4) Demonstration that outcome of interest was not present at start of study a) yes ¯ b) no 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.)	M2	M2
31	Quality Criteria Checklist	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? 2.3 Were health, demographics, and other characteristics of subjects described? 2.4 Were the subjects/patients a representative sample of the relevant population?	R2	R2
32	NICE	The same exclusion criteria are used for both cases and controls	M1	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	R1 & M1
37	Heller	Are exclusion criteria appropriate?	M1	M1
39	Faillie	Were inclusion and exclusion criteria implemented uniformly across study groups?	M1	M1
40	Manchikanti	Inclusiveness of Population Population A study's population is clinically relevant to assessing methodological quality and bias risk (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 (4)	M1	M1
41	Handu	Was the selection of study subjects/patients free from bias? 2.1 Were inclusion/exclusion criteria specified (eg, 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	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	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	R1
47	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. See explanatory document for guidance related to matched designs.	R2	R2
48	ISPOR-ISPE	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,……	R2	R2
49	Critical reading sheet	Was the participant selection method suitable?	M1	M1

Patients (Population) Item 7

	Tool	Michiel	Michiel	Li
1	Relevant	- Population defined - Population justified	R1&M1	R1 & M1
2	RAMboMan	Recruitment of participants 'who are the findings applicable to?'	M1	M1
3	MMAT	Are the participants representative of the target population?	M1	M1
4	CASP Case-Control	- Were the cases recruited in an acceptable way? - were the controls representative of the defined population (geographically and/or temporally)	M1	M1
4	CASP cohort	- Was the cohort recruited in an acceptable way?	M1	M1
6	JBI Cohort	Were the groups/participants free of the outcome at the start of the study (or at the moment of exposure)?	M2	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	M2
10	NIH-NHLBI	Was the study population clearly specified and defined?	R1	R1
13	RTI-item bank	Was the strategy for recruiting participants into the study the same across study groups/arms of the study? [PIs: 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	M2
14	SIGN Case-Control	The cases and controls are taken from comparable populations.	M1	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	M2
21	Rangel	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	R2
22	Thomas (EPHPP)	Are the individuals selected to participate in the study likely to be representative of the target population?	M1	M1
23	Atluri	Subjects similar to populations in which the test would be used and with a similar spectrum of disease	M1	M1
25	Blagojevic	Representative sample e.g., general population sample should not exclude subgroups	M1	M1
26	Gendaidy	Is the source of subject population (including sampling frame) clearly described? Yes – Clearly described	R2	R2

		Y . 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 P . Details are not clearly described. No – Not described N . No mention of source population. Are newly incident cases taken into account? Not Applicable NA . Cohort design . Intervention design . Cross-sectional design . Proportional design Yes No Randomization: 23. Are the study subjects randomized to groups? Not Applicable NA . Cohort design . Case-control design . Cross-sectional design . Proportional design Yes No Y N . Random allocation was made in intervention designs. . Non-randomized or method of randomization in intervention designs would not ensure random allocation. Y N Newly incident cases in case-control designs. . Prevalent cases in case-control designs.		
27	Glasgow University	Were there clearly defined groups of patients, similar in all important ways other than exposure to the treatment or other causes?	R2	R2
28	Tseng	Was the patient population from which the cases were selected from adequately described or identified (e.g. geographically?	R2	R2
29	Weightman	Is the population studied appropriate? - Was the sample representative of its target population? - How was the sample selected – random, stratified?	M1	M1
30	Newcaslte-Ottawa (NOS) (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	R2
31	Quality criteria Checklist	Was the selection of study subjects/patients free from bias? Were study groups comparable	M1	M1
32	NICE	- The cases and controls are taken from comparable populations	M1	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	R1 & M2

37	Heller	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? 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?	M2	M2
39	Faillie	Are all the subjects recruited from the same source population? Is the origin of controls clearly specified?	R1&M1	R1 & M1
40	Manchikanti	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	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	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	R2
49	Critical reading sheet	Describe the location and study period. Is the target population of the study adequately defined? Please, note it down.	R1	R1

Participation rate (Population) Item 8

	Tool		Michiel	Li
10	NIH-NHLBI	Was the participation rate of eligible persons at least 50%?	M2	M2
14	SIGN Case-control	What percentage of each group (cases and controls) participated in the study?	R1	R1
14	SIGN cohort	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	R1	R1
16	STROBE	Report numbers of individuals at each stage of study—eg 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	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 Allocation and intervention exposure: the number of participants assigned to each study condition and the number of participants who received each intervention	R2	R2
20	GRADE	Were more than 80% of participants enrolled in trials included in the analysis (i.e. no potential reporting bias)?	M1	M1
22	Thomas (EPHPP)	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	M1
24	Bishop	Where response rate // (No. of participants in the study/No. of people invited to take part) /// 100	R1	R1
25	Blagojevic	Baseline response ~70%	M1	M1
26	Genaidy	- 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? Not Applicable *(for more details on the item we refer to the original tool)*	R2&M2	R2 & M2
29	Weightman	Did the study achieve a good response rate?	M1	M1
30	Newcaslte-Ottawa (NOS) (NOS)	Non-Response rate a) same rate for both groups ¯ b) non respondents described c) rate different and no designation	M1	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	R2 & M2
34	AXIS	Does the response rate raise concerns about non-response bias?	M1	M1
37	Heller	In a cross-sectional study, is the item-specific response rate adequate?	M1	M1
39	Faillie	Are the number of participants clearly reported throughout the study?	M1	R1
40	Manchikanti	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	R2
41	Handu	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	R2 & M2
47	RECORD	RECORD 13.1: 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	R2

Baseline characteristics (Population) Item 9

	Tool		Michiel	Li
3	MMAT	Are the groups comparable at baseline?	M1	M1
6	JBI Case-Series	Was there clear reporting of the demographics of the participants in the study? - Was there clear reporting of clinical information of the participants?	R1	R1
6	JBI Cohort	- Were the two groups similar and recruited from the same population?	M1	M1
8	ISPORE-AMCP-NPC	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)? *(for more details on the item we refer to the original tool)*	M2	M2
13	RTI-item bank	Is the selection of the comparison group appropriate, after taking into account feasibility and ethical considerations. [PI: 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	M2
14	SIGN Cohort	-The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	M1	M1
17	TREND	Baseline demographic and clinical characteristics of participants in each study condition • Baseline characteristics for each study condition relevant to specific 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	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	M1
23	Atluri	- 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	M1
26	Genaidy	Are the characteristics of study participants described? *(for more details on the item we refer to the original tool)*	R2	R2
27	Glasgow University	Were there clearly defined groups of patients, similar in all important ways other than exposure to the treatment or other causes?	M1	M1
31	Quality Criteria Checklist	- 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	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	R1 & M2
33	IHE quality appraisal	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	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	M2

38	Gagnier	Demographic information (eg, 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 whenever possible and details about relevant comorbidities including past interventions and their outcomes	R2	R2
39	Faillie	Are baseline characteristics and prognostic factors comparable between different groups?	M1	M1
40	Manchikanti	Similarity of Groups at Baseline for Important Prognostic Indicators No groups or groups dissimilar with significant influence on outcomes Groups dissimilar without significant influence on outcomes Groups similar	M1	M1
41	Handu	Were distribution of disease status, prognostic factors, and other factors (eg, demographic characteristics) similar across study groups at baseline?	M2	M2
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	R2
49	Critical reading sheet	Note the number and characteristics of the participants down	R1	R1

Selection (Intervention) Item 10

	Tool		Michiel	
4	CASP Case-Control	-Were the cases recruited in an acceptable way? -Were the controls selected in an acceptable way?	M1	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	R1 & M2
8	ISPORE-AMCP-NPC	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 *(for more details on the item we refer to the original tool)*	M2	M2
9	GRACE	Was the study (or analysis) population restricted to new initiators of treatment or those starting a new course of treatment?	M2	M2
10	NIH-NHLBI	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	M2
13	RTI-item bank	Is the selection of the comparison group appropriate, after taking into account feasibility and ethical considerations. [PI: 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	M2
31	Quality criteria checklist	Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect?	M1	M1
33	IHE quality appraisal	Were additional interventions (cointerventions) clearly reported in the study?	M1	M1
47	RECORD	Use of any comparator groups should be outlined and justified.	R1	R1
50	Kennedey	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 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.	M2	M2

Definition (Intervention) Item 11

	Tool		Michiel	Li
1	Relevant	(If relevant), exposure (e.g. treatment) is clearly defined	R1	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	R1&M2
7	ROBINS-I	3.1 Were intervention groups clearly defined? 3.2 Was the information used to define intervention groups recorded at the start of the intervention? 3.3 Could classification of intervention status have been affected by knowledge of the outcome or risk of the outcome?	R1&M2	R1&M2
8	ISPORE-AMCP-NPC	Was exposure defined and measured in a valid way?	M1	M1
13	RTI-item bank	What is the level of detail in describing the intervention or exposure? [PI: 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. PI needs to establish criteria for high, medium, or low response.]	R2	R2
16	STROBE	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	R2	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	R2
18	ACROBAT-NRSI	Is intervention status well defined?	M1	M1
21	Rangel	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	R2
23	Atluri	Clear definition of exposure	R1	R1
24	Bishop	A definition of CAM and/or a list of specific CAM therapies is provided to participants	R1	R1
26	Genaidy	Are all the exposure variables/intervention(s) clearly described?	R1	R1
28	Tseng	Is the surgical technique/intervention adequately described (e.g., specifically referenced article)?	R2	R2
30	Newcaslte-Ottawa (NOS) (NOS)	- Is the case definition adequate? - Same method of ascertainment for cases and controls	R1	R1
31	Quility Criteria Checklist	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	R1
32	NICE	Cases are clearly defined and differentiated from controls	R1	R1
33	IHE quality appraisal	Was the intervention clearly described in the study?	R1	R1
37	Heller	Intervention features (for an intervention study): is the intervention described adequately?	R1	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	R2

47	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	R2
48	ISPOR-ISPE	The type of exposure that is captured or measured, e.g., drug versus procedure, new use, incident, prevalent, cumulative, timevarying.	R1	R1

Measurement (Intervention) item 12

	Tool		Michiel	Li
2	Ramboman-GATE	Were exposures & outcomes well Measured?' were they measured Objectively?	M1	M1
3	MMAT	- Are measurements appropriate regarding both the outcome and intervention (or exposure)? - Are the measurements appropriate?	M1	M1
4	CASP Case-Control	Was the exposure accurately measured to minimize bias?	M1	M1
4	CASP Cohort	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	M2
6	JBI Case-Control	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	R1&M1	R1&M1
6	JBI Case-series	Was the condition measured in a standard, reliable way for all participants included in the case series? *(For more details on the item we refer to the original tool)*	M2	M2
6	JBI Cohort	Were the exposures measured similarly to assign people to both exposed and unexposed groups? Was the exposure measured in a valid and reliable way?	M2	M2
7	ROBINS-I	4.1. Were there deviations from the intended intervention beyond what would be expected in usual practice? 4.2. If Y/PY to 4.1: Were these deviations from intended intervention unbalanced between groups and likely to have affected the outcome?	M2	M2
8	ISPORE-AMCP-NPC	Was exposure defined and measured in a valid way?	M1	M1
10	NIH-NHLBI	- 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? - Were the exposure measures (independent variables) clearly defined, valid, reliable, and implemented consistently across all study participants?	M2	M2
13	RTI-item bank	Are interventions/exposures assessed using valid and reliable measures, implemented consistently across all study participants? [PI: Important measures may be listed separately. PI may need to establish a threshold for what would constitute acceptable measures based on study topic. When subjective or objective measures could be collected, subjective measures based on selfreport may be considered as being less reliable and valid than objective measures such as clinical reports and lab findings. Replicate question when needed.]	M2	M2
14	SIGN Case-Control	Exposure status is measured in a standard, valid and reliable way.	M1	M1

23	Atluri	Measurement method standard, valid and reliable	M1	M1
26	Genaidy	Are the exposure variables reliable? *(For more details on the item we refer to the original tool)*	M2	M2
27	Glasgow university	Were treatments/exposures and clinical outcomes measured in the same way for both groups?	M1	M1
31	Quality criteria checklist	Was the amount of exposure and, if relevant, subject/patient compliance measured?	R1	R1
32	NICE	Exposure status is measured in a standard, valid and reliable way	M1	M1
36	Pluye	Justification of measurements (validity and standards)	M1	M1
37	Heller	Observations/risk factors: how are the exposures measured?	M1	M1
39	Faillie	Cohort, case-control studies: Was the method for ascertaining drug use and drug use duration adequately constructed, and equal for all participants?	M2	M2
41	Handu	Was the amount of exposure and, if relevant, subject/patient compliance measured?	M1	M1
45	ENCePP	Does the protocol address the validity of the exposure measurement? (e.g., precision, accuracy, use of validation sub-study)	M1	M1

Blinding (intervention) item 13

	Tool		Michiel	Li
5	SURE	Was allocation to intervention or comparator groups concealed? Is it possible for those allocating to know which group they are allocating people to? As above, methods such as alternating participants coin toss, birth dates, record numbers, days of the week will not allow appropriate allocation concealment. Were participants/investigators blinded to group allocation? If NO, was assessment of outcomes blinded?	M2	M2
17	TREND	Whether or not participants, those administering the intervention, 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	R1&M2	R1 & 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	R2 & M2
22	Thomas	Were the study participants aware of the research question?	M1	M1
26	Genaidy	Are the observers blinded to: subject groupings when the exposure/intervention assessment was made or the disease status of subjects when conducting exposure assessment?	M2	M2
31	Quality Criteria Checklist	Was blinding used to prevent introduction of bias? -In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate? -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? -In case control study, was case definition explicit and case ascertainment not influenced by exposure status?	M2	M2
32	NICE	- Participants receiving care were kept 'blind' to treatment allocation The knowledge of assignment to a particular treatment group may affect outcomes such as a study - Individuals administering care were kept 'blind' to treatment allocation	M2	M2
35	AHRQ	- Participants are blinded to intervention group assignment - Providers are blinded to participant intervention group assignment	M1	M1
41	Handu	In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate? - In case control study, was case definition explicit and case ascertainment not influenced by exposure status?	M1	M1

Length of follow-up (intervention) item 14

	Tool		Michiel	Li
6	JBI Case Control	Was the exposure period of interest long enough to be meaningful? *(For more details on the item we refer to the original tool)*	M2	M2
8	JBI Cohort	Was the follow-up period of sufficient duration to detect differences addressed?	M1	M1
10	NIH-NHLBI	Was the timeframe sufficient so that one could reasonably expect to see an association between exposure and outcome if it existed?	M2	M2
11	HEB Wales	Was follow up for long enough?	M1	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 followup? Abstractor: 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? [PI: 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	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	M2
23	Atluri	Length of follow-up adequate for question	M1	M1
25	Blagojevic	Length of follow-up ~36 months	M1	M1
26	Genaidy	Is the minimum follow-up time since initial exposure sufficient enough to detect a relationship between exposure/intervention and outcome?	M2	M2
27	Glasgow University	Was the follow-up of study patients sufficiently long for the outcome to occur?	M1	M1
30	Newcaslte-Ottawa (NOS) (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	M1
31	Quality Criteria Checklist	Was the period of follow-up long enough for important outcome(s) to occur?	M1	M1
32	NICE	The study had an appropriate length of follow-up	M1	M1
39	Faillie	Was the duration of follow-up adequate to assess the drug safety outcome?	M1	M1
40	Manchikanti	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	M2

41	Handu	Was the period of follow-up long enough for important outcome(s) to occur?	M1	M1

Outcome selection (outcome) Item 15

	Tool		Michiel	Li
8	ISPORE-AMCP-NPC	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	R1 & M1
13	RTI-item bank	Are outcomes assessed using valid and reliable measures, implemented consistently across all study participants? [PI: 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	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	M2
20	GRADE	Was an objective outcome used? -Was the included outcome not a surrogate outcome?	M1	M1
31	Quality Criteria Checklist	- 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	R2
34	AXIS	Were the risk factor and outcome variables measured appropriate to the aims of the study?	M1	M1
39	Faillie	Was the method for ascertaining the drug safety outcome adequately constructed and equal for all participants? *(For more details on the item we refer to the original tool)*	M2	M2
41	Handu	Were primary and secondary endpoints described and relevant to the question?	R1	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	R2

Definition (Outcome) Item 16

	Tool		Michiel	Li
1	Relevant	Primary outcomes defined	R1	R1
8	ISPORE-AMCP-NPC	Were the primary outcomes defined and measured in a valid way?	M1	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. [PI: 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	M2
14	SIGN Cohort	The outcomes are clearly defined.	M1	M1
16	STROBE	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable *(For more details on the item we refer to the original tool)*	R2	R2
17	TREND	Clearly defined primary and secondary outcome measures	R1	R1
23	Atluri	Primary/secondary outcomes clearly defined	R1	R1
26	Genaidy	Are the main outcomes clearly described? *(For more details on the item we refer to the original tool)*	M2	M2
28	Tseng	Is there a clearly defined single primary outcome?	R1	R1
31	Critical Criteria Checklist	- Were primary and secondary endpoints described and relevant to the question? - Were outcomes clearly defined and the measurements valid and reliable	M1	M1
32	NICE	- The study used a precise definition of outcome - A valid and reliable method was used to determine the outcome 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).	M2	M2
37	Heller	What are the outcome factors?	R1	R1
39	Faillie	- 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). - clear / standardized definition of the drug safety outcome (e.g., diagnostic codes, clinical and laboratory data).	R2	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	R2

47	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	R2
48	ISPOR-ISPE	Reporting on outcome definition should include: The date of an event occurrence. Codes, frequency and temporality of codes, diagnosis position, care setting Validation	R2	R2
49	Critical reading sheet	Is the test used for comparison adequately defined? Are the outcomes of interest adequately defined? Please, note it down.	R1	R1

Measurement (Outcome) Item 17

	Tool		Michiel	Li
2	RAMboMan-GATE	were exposures & outcomes well Measured?'	M1	M1
3	MMAT	- Are measurements appropriate regarding both the outcome and intervention (or exposure)? - Are the measurements appropriate?	M1	M1
4	CASP Cohort	- 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	M2
5	SURE	Was the condition measured in a standard, reliable way for all participants included in the case series?	M1	M1
6	JBI Case Control	Were outcomes assessed in a standard, valid and reliable way for cases and controls?	M1	M1
6	JBI Case series	Was the condition measured in a standard, reliable way for all participants included in the case series?	M1	M1
6	JBI Cohort	Were the outcomes measured in a valid and reliable way?	M1	M1
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	M2
8	ISPORE-AMCP-NPC	Were the primary outcomes defined and measured in a valid way?	M1	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	M2
10	NIH-NHLBI	- 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	M2
13	RTI-item bank	Are outcomes assessed using valid and reliable measures, implemented consistently across all study participants? [PI: 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	M2
15	Montreal	What are the outcome factors and how are they measured? a) Are all relevant outcomes assessed?	R2	R2

		b) Is there measurement error?		
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	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	R1 & M1
18	ACROBAT-NRSI	- Was the outcome measure objective? - Were any systematic errors in measurement of the outcome unrelated to intervention received?	M2	M2
21	Rangel	Is the diagnostic method clearly described for assessing outcome(s) of interest?	R1	R1
25	Blagojevic	Appropriate and validated outcome measure	M1	M1
26	Genaidy	- Are the main outcome measures reliable? - Are the methods of assessing the outcome variables standard across all groups?	M1	M1
27	Glasgow University	- Were treatments/exposures and clinical outcomes measured in the same way for both groups?	M1	M1
28	Tseng	Methods for assessing outcomes described?	R1	R1
31	Quality Criteria Checklist	Were outcomes clearly defined and the measurements valid and reliable 7.4 Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? 7.5 Was the measurement of effect at an appropriate level of precision? 7.6 Were other factors accounted for (measured) that could affect outcomes? 7.7 Were the measurements conducted consistently across groups?	M2	M2
32	NICE	-A valid and reliable method was used to determine the outcome -What outcome measure(s) is/are used?	R1&M1	R1&M1
33	IHE	- Were relevant outcomes appropriately measured with objective and/or subjective methods? -Were outcomes measured before and after intervention?	M1	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	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	M2
36	Pluye	Justification of measurements (validity and standards)	M1	M1
37	Heller	Is there bias in the measurement? Are these outcome measures appropriate?	M1	M1
39	Faillie	- RCT, cohort studies: Is the time frequency of drug safety outcome assessment during the follow-up period appropriate?	M1	M1
40	Manchikanti	Outcomes Assessment Criteria for Significant Improvement	M2	M2

		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 (2) Significant improvement with pain and function ≥ 50% or 3 points and 40% reduction in disability scores (4)		
41	Handu	Were outcomes clearly defined and the measurements valid and reliable?	M1	M1
43	Young	Were study measure's objective or subjective and is recall bias likely if they were subjective?	M2	M2
45	ENCePP	Does the protocol describe how the outcomes are defined and measured? Does the protocol address the validity of outcome measurement? (e.g., precision, accuracy, sensitivity, specificity, positive predictive value, use of validation sub-study)	R1&M1	R1&M1
50	Kennedey	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	M2

Blinding (Outcome) Item 18

	Tool		Michiel	Li
2	RAMboMan-GATE	Were outcomes measured Blind to whether participant was in EG or CG (or vice versa)?	M1	M1
3	MMAT	Are outcome assessors blinded to the intervention provided?	M1	M1
4	CASP Cohort	Were the subjects and/or the outcome assessor blinded to exposure (does this matter)	M1	M1
7	ROBINS-I	Were outcome assessors aware of the intervention received by study participants?	M1	M1
10	NIH-NHLBI	Were the outcome assessors blinded to the exposure status of participants?	M1	M1
12	ROBANS	Blinding of outcome assessments	M1	M1
13	RTI-item bank	Were the outcome assessors blinded to the intervention or exposure status of participants? [PI: There may be circumstances where clinical evaluators cannot be blinded to exposure status. Drop if not relevant to the body of literature.]	M2	M2
14	SIGN Cohort	- 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	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	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	R2 & M2
20	GRADE	Was there blinding of outcome assessment (i.e., no potential for detection bias)?	M2	M2
21	Rangel	If comparison groups were used, was any attempt made to blind evaluators during the analysis of data?	M1	M1
22	Thomas EPHPP	Was (were) the outcome assessor(s) aware of the intervention or exposure status of participants?	M1	M1
27	Glasgow University	Was the assessment of outcomes either objective or blinded to exposure?	M1	M1
28	Tseng	Was any attempt made to blind evaluators during the analysis of data?	M1	M1
30	Newcaslte-Ottawa (NOS) (NOS)	Assessment of outcome -Independent blind assessment -Record linkage -self report -No description	R1	R1
31	Quality Criteria Checklist	- 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? - In diagnostic study, were test results blinded to patient history and other test results?	M2	M2
32	NICE	Investigators were kept 'blind' to other important confounding and prognostic factors	M1	M1
33	IHE quality appraisal	Blind assessment of outcomes	M1	M1

39	Faillie	Was the blinding method of drug safety outcome assessment appropriate considering the nature of the adverse event?	M1	M1
41	Handu	- 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	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	R2
49	Critical reading sheet	Was the assessment of the results of both tests blind?	M1	M1

Data source (Data collection) Item 19

	Tool		Michiel	Li
1	Relevant	The data source (or database), as described, contains adequate exposures (if relevant) and outcome variables to answer the research question	M1	M1
8	ISPORE-AMCP-NPC	-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	M1
9	GRACE	Were the primary outcomes adequately recorded for the study purpose (e.g., available in sufficient detail through data sources)?	R1&M1	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	R2
19	MINORS	Prospective collection of data: data were collected according to a protocol established before the beginning of the study	M2	M2
33	IHE quality appraisal	Case series collected in more than one centre (multicentre study)	M1	M1
39	Faillie	Secondary databases studies: Are the characteristics of the database clearly described?	R1	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	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	R2
47	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	R2
48	ISPOR-ISPE	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*	R2	R2

		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 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).*		

Missing data (Data collection) Item 20

	Tool		Michiel	Li
1	Relevant	The extent of missing data is reported	R1	R1
3	MMAT	- Are there complete outcome data? - Almost all the participants contributed to almost all measures	R1	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	R1&M1
6	JBI Cohort	- 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	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	M2
8	ISPORE-AMCP-NPC	Was the extent of missing data reported?	R1	R1
12	ROBANS	- Incomplete outcome data - Attrition bias caused by the inadequate handling of incomplete outcome data	R1	R1
14	SIGN Cohort	- 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	M2
16	STROBE	- Explain how missing data were addressed - Indicate number of participants with missing data for each variable of interest	R2	R2
17	TREND	Methods for imputing missing data, if used	R1	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	M2
21	Rangel	Do the authors address whether there is any missing data?	M1	M1
25	Blagojevic	- Loss and dropout at follow-up <25% - Adequate description and discussion of dropouts	R1&M1	R1 & M1
26	Genaidy	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	R1 & M2
28	Tseng	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	M1
29	Weightman	Is there an explanation of how missing data have been handled?	M1	M1
31	Quality Criteria Checklist	- 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	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	M2
45	ENCePP	Does the plan describe methods for handling missing data?	R1	R1
47	RECORD	Explain how missing data were addressed.	R1	R1

Loss to follow-up (Data collection) Item 21

	Tool		Michiel	Li
4	CASP Case-Control	RCT, cohort studies: Does the study adequately address biased loss to follow-up?	M1	M1
4	CASP Cohort	Was the follow up of subjects complete enough?	M1	M1
5	SURE	Was follow-up ≥ 80%?	M1	M1
6	JBI Cohort	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? *(For more details on the item we refer to the original tool)*	M2	M2
10	NIH-NHLBI	Was loss to follow-up after baseline 20% or less? *(For more details on the item we refer to the original tool)*	M2	M2
13	RTI-item bank	Did attrition from any group exceed [x] percent? [PI: 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	M2
14	SIGN Cohort	Comparison is made between full participants and those lost to follow up, by exposure status.	M1	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	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	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	M2
22	Thomas	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	M2
25	Blagojevic	- All subjects aged 50 or over at follow-up - Loss and dropout at follow-up <25%	M2	M2
26	Genaidy	- 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	R2
30	Newcaslte-Ottawa	- 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	M2	M2

	(NOS) (NOS)	description provided of those lost) ¯ c) follow up rate < ____% (select an adequate %) and no description of those lost d) no statement		
31	Quality Criteria Checklist	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	M2
33	IHE	Was the loss to follow-up reported?	R1	R1
39	Faillie	RCT, cohort studies: Does the study adequately address biased loss to follow-up?	M1	M1
43	Young	Were there important losses to follow-up?	M1	M1
47	RECORD	If applicable, explain how loss to follow-up was addressed.	R1	R1
48	ISPOR-ISPE	Reporting on follow-up time should include: Censoring criteria the criteria that censor follow up.	R1	R1
50	Kennedey	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	M2

Description (Data analysis) Item 22

	Tool		Michiel	Li
1	Relevant	- Potential confounders are addressed - Study groups are compared at baseline	M1	M1
4	CASP Case-Control	is the analysis appropriate to the design	M1	M1
5	SURE	Data analysis Are the statistical methods well described?	M1	M1
6	JBI Case-Control	Were strategies to deal with confounding factors stated? Was appropriate statistical analysis used?	M1	M1
6	JBI Case-Series	Was statistical analysis appropriate?	M1	M1
6	JBI Cohort	Was appropriate statistical analysis used?	M1	M1
8	ISPORE-AMCP-NPC	Were analyses of subgroups or interaction effects reported for comparison groups?	M2	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	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	M2

15	Montreal	Are statistical tests considered?	M1	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	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 analyses • Methods for imputing missing data, if used • Statistical software or programs used	R2	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	M1
24	Bishop	Adjust for potential confounders in statistical analysis	M1	M1
25	Blagojevic	Appropriate analysis	M1	M1
26	Genaidy	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	M2
28	Tseng	Statistical methods described? Statistical software identified?	M1	M1
31	Quality Criteria Checklist	Was the statistical analysis appropriate for the study design and type of outcome indicators? 8.1 Were statistical analyses adequately described the results reported appropriately 8.2 Were correct statistical tests used and assumptions of test not violated? 8.3 Were statistics reported with levels of significance and/or confidence intervals? 8.4 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)? 8.5 Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)? 8.6 Was clinical significance as well as statistical significance reported?	M2	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	M1
33	IHE Quality appraisal	Were the statistical tests used to assess the relevant outcomes appropriate	M1	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	M1
37	Heller	Are statistical tests appropriate, and correct?	M1	M1
39	Faillie	- 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	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. Any sensitivity analyses should be described. Details of the statistical analysis may be specified later, but before analysis begins, as part of a p	R2	R2
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	R2
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	R2
48	ISPOR-ISPE	Reporting on statistical software should include: Statistical software program used, The software package, version, settings, packages or analytic procedures.	R1	R1

Sensitivity (Data analysis) Item 23

	Tool		Michiel	Li
1	Relevant	The authors describe the statistical uncertainty of their findings (e.g., p values, confidence intervals)	M1	M1
4	CASP Case-Control	How precise was the estimate of the treatment effect?	M1	M1
4	CASP Cohort	How precise are the results?	M1	M1
8	ISPORE-AMCP-NPC	Were sensitivity analyses performed to assess the effect of key assumptions or definitions on outcomes?	M2	M2
9	GRACE	Were any meaningful analyses conducted to test key assumptions on which primary results are based?	M2	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	M1
16	STROBE	Describe any sensitivity analyses	R1	R1
39	Faillie	Cohort, case-control studies: Do sensitivity analyses account for different exposure windows, induction/lag periods? *(For more details on the item we refer to the original tool)*	M2	M2
42	Viswanathan	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-of bias influence the estimate of effect or heterogeneity.	M2	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	R1
45	ENCePP	Are relevant sensitivity analyses described?	R1	R1

Confounding (Data Analysis) Item 24

	Tool		Michiel	Li
1	Relevant	Possible biases and/or confounding factors described	R1	R1
3	MMAT	Are the confounders accounted for in the design and analysis?	M1	M1
4	CASP Case-control	Have the authors taken account of the potential confounding factors in the design and/or in their analysis?	M1	M1
4	CASP Cohort	Have the authors identified all important confounding factors?	M1	M1
6	JBI Case Control	Were confounding factors identified?	M1	M1
6	JBI Cohort	Were confounding factors identified? Were strategies to deal with confounding factors stated?	R1&M1	R1&M1
7	ROBINS-I	List the confounding domains relevant to all or most studies	M1	M1
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.	M2	M2
11	HEB Wales	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	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	M2
14	SIGN Case-Control	How well was the study done to minimise the risk of bias or confounding?	M1	M1
14	SIGN Cohort	- How well was the study done to minimise the risk of bias or confounding? - The main potential confounders are identified and taken into account in the design and analysis.	M1	M1
15	Montreal	What important potential confounders are considered?	M1	M1
16	STROBE	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	R1	R1

18	ACROBAT-NRSI	Is confounding of the effect of intervention unlikely in this study?	M1	M1
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	M2
23	Atluri	Assessment of confounding Study groups comparable to non-participants with regard to confounding factors	M2	M2
24	Bishop	Adjust for potential confounders in statistical analysis	M1	M1
26	Genaidy	Are the important covariates and confounders described in terms of individual variables?	M1	M1
29	Weightman	- Have confounding and bias been considered? - Is there an explanation of how potential confounding factors have been controlled for?	R1&M1	R1&M1
31	Quality Criteria Checklist	If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were pre-existing 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 pre-existing differences accounted for by using appropriate adjustments in statistical analysis? Were other factors that could affect outcomes (e.g., confounders) measured or accounted for?	M1	M1
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	M2
33	IHE quality appraisal	Study groups comparable to nonparticipants with regard to confounding factors Discussion of possible confounders	M1	M1
35	AHRQ	For nonrandomized studies, specify likely sources of potential confounding	R1	R1
37	Heller	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	M1
39	Faillie	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?	M2	M2
42	Viswanathan	~ 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	M2	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	R1

Are all the results included (results) Item 25

	Tool		Michiel	Li
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	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., 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	M2
4	CASP Case-Control	What are the bottom-line results? How precise was the estimate of the treatment effect? HINT: 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	R2
4	Casp Cohort	What are the results of this study? • what are the bottom-line results • have they reported the rate or the proportion between the exposed/unexposed, the ratio/rate difference • how strong is the association between exposure and outcome (RR) • what is the absolute risk reduction (ARR)	R2	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	R2
6	JBI Case-Series	- 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	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? different *subgroups*?	M2	M2
8	ISPORE-AMCP-NPC	- 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	R2
11	HEB Wales	Were all important outcomes/results considered?	M1	M1
12	ROBANS	- Selective outcome reporting - Reporting bias caused by the selective reporting of outcomes	M1	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	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	R2
20	Grade	Were data reported consistently for the outcome of interest (i.e., no potential selective reporting)?	M1	M1
26	Genaidy	- 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?	R2	R2

		- Are outcome data reported by levels of exposure?		
28	Tseng	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? 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.	R2	R2
29	Weightman	-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	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	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	R1
37	Heller	What are the main results and are they presented in an understandable way? Have measures of absolute risk as well as relative risk	R2&M1	R2 & M1

		been included? [For any intervention study] Have the resource and cost implications of implementing the intervention and cost-effectiveness of the intervention been described?		
39	Faillie	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	M2
42	Viswanathan	Present findings and conclusions transparently, balancing the competing considerations of simplicity of presentation with burden on the reader	R1	R1
49	Critical reading sheet	Are the outcomes properly summarized and described?	R1&M1	R1&M1

SPIN (Results) Item 26

	Tool		Michiel	Li
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	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	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	M1
11	HEB Wales	Are the authors' conclusions adequately supported by the information cited?	M1	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	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	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	R2
23	Atluri	Conclusions supported by results with possible biases and limitations taken into consideration	M1	M1
31	Quality Criteria Checklist	Are conclusions supported by results	M1	M1
33	IHE quality appraisal	Conclusions of the study supported by results	M1	M1
34	AXIS	Were the authors' discussions and conclusions justified by the results	M1	M1
37	Heller	Have the results been interpreted appropriately?	M1	M1
38	Gagnier	Rationale for conclusions (including assessments of cause and effect)	M1	M1
39	Faillie	Is publication bias assessed? "refer to tool"	M2	M2
41	Handu	Are conclusions supported by results with biases and limitations taken into consideration?	M1	M1
43	Young	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	M2	M2

		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.		
49	Critical reading sheet	Are the conclusions justified?	M1	M1

Conflict of interest Item 27

	Tool		Michiel	Li
1	Relevant	Potential conflicts of interest, including study funding, are stated	R1	R1
5	SURE	Is any sponsorship/conflict of interest reported?	R1	R1
8	ISPORE-AMCP-NPC	-Were there any potential conflicts of interest? -If there were potential conflicts of interest, were steps taken to address these?	M2	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	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	R2
20	GRADE	There was no industry influence on studies included in the review?	M2	M2
31	Quality Criteria Checklist	10. Is bias due to study's funding or sponsorship 10.1 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? unlikely?	M2	M2
32	NICE	How was the study funded?	R1	R1
33	IHE	Are both competing interest and source of support for the study reported?	R2	R2
34	AXIS	Were there any funding sources or conflicts of interest that may affect the authors' interpretation of the results?	R1	R1
39	Faillie	-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	M2
40	Manchikanti	Funding and Sponsorship Trial included industry employees with or without proper disclosure	R2	R2
41	Handu	10. Is bias due to study's funding or sponsorship unlikely? 10.1 Were sources of funding and investigators' affiliations described? 10.2 Was there no apparent conflict of interest?	M2	M2
43	Young	Are there any conflicts of interest?	R1	R1
49	Critical reading sheet	Is the existence or absence of conflicts of interest properly described? When possible, specify the financial source.	R1&M1	R1&M1