NHS National Services Scotland

Critical Appraisal Checklists

D00.1 (Draft)

Month 2024

NHSScotland Assure

Quality in the healthcare environment

Document Control Sheet:

Key Information

Title	Critical Appraisal Checklist
Date Published / Issued	
Date Effective From	
Version No.	V00.1 (Draft)
Document Type	SOP literature review supporting document
Document Status	
Author	HCS Technical Team
Owner	NHSScotland Assure HCS
Approver	
Approval Date	
Contact	nss.assuretechhcs@nhs.scot
File Location	

REVISION HISTORY

VERSION:	DATE:	SUMMARY OF CHANGES:	NAME:

DISTRIBUTION

NAME	TITLE	VERSION	ISSUE DATE

APPROVALS

NAME	TITLE	VERSION	ISSUE DATE

ASSOCIATED DOCUMENTATION

Any relevant applicable legislation, policies, or guidance.

DOCUMENT TITLE	DOCUMENT FILE PATH / URL
Legislation	
Regulations	
Standards	
Guidelines	
Other	

Contents

1. Introduction

This document contains all the checklists that need to be used for assessing the methodological quality/risk of bias of the study designs and reports as per the HCS Technical Team SOP for literature reviews.

The checklists in this document are well-established and sourced from recognised organisations. They were selected following a rigorous assessment by the HCS Technical Team with some checklists amended to enhance robustness, clarity, and usability.

Descriptions

For the purposes of the HCS Technical literature reviews, the concept of "population and sample" extends beyond humans to include objects, events, and other relevant entities.

Screening questions

Two initial screening questions have been included in all the checklists, adapted from the MMAT critical appraisal checklist.

The purpose of these screening questions is to determine whether the study's research questions, aims, and/or objectives are clearly defined and whether the collected data appropriately address those questions, prior to completing the critical appraisal. For expert opinion and grey literature reports, the screening question may not be applicable. In such cases, they should be marked as N/A, and the checklist questions should then be completed.

If the answer to one or both questions is "No" or "Unclear", further appraisal will not be conducted, and the study will be excluded from the review.

Screening questions (further appraisal may not be feasible or appropriate when the answer is “No” or “Unclear” to one or both screening questions):

Screening Question	Yes	No	Unclear	N/A
S1. Are there clear research questions, aims and/ or objectives?	☐	☐	☐	☐
S2. Do the collected data answer the research questions, aims and/ or objectives?	☐	☐	☐	☐

2. Case Series Studies

JBI critical appraisal checklist for case series studies

Amendments to the original checklist: Screening questions were added. Question 1 from the original JBI checklist was removed as it duplicated the screening questions incorporated into all checklists. Question 9 of the original JBI checklist was also removed, as it was deemed too subjective and overlapped with screening question 2.

Screening Question	Yes	No	Unclear	N/A
S1. Are there clear research questions, aims and/ or objectives?	☐	☐	☐	☐
S2. Do the collected data answer the research questions, aims and/ or objectives?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
1. Were there clear criteria for inclusion in the case series?	☐	☐	☐	☐
2. Was the condition measured in a standard, reliable way for all participants included in the case series?	☐	☐	☐	☐
3. Were valid methods used for identification of the condition for all participants included in the case series?				
4. Did the case series have consecutive inclusion of participants?	☐	☐	☐	☐
5. Did the case series have complete inclusion of participants?				
6. Was there clear reporting of the demographics of the participants in the study?	☐	☐	☐	☐
7. Was there clear reporting of clinical information of the participants?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
8. Were the outcomes or follow-up results of cases clearly reported?	☐	☐	☐	☐
9. Was the statistical analysis appropriate?	☐	☐	☐	☐

Guidance

Q1. Were there clear criteria for inclusion in the case series?

The authors should provide clear inclusion (and exclusion criteria where appropriate) for the study participants. The inclusion/exclusion criteria should be specified (e.g., risk, stage of disease progression) with sufficient detail and all the necessary information critical to the study.

Q2. Was the condition measured in a standard, reliable way for all participants included in the case series?

The study should clearly describe the method of measurement of the condition. This should be done in a standard (i.e. same way for all patients) and reliable (i.e. repeatable and reproducible results) way.

Q3. Were valid methods used for identification of the condition for all participants included in the case series?

Many health problems are not easily diagnosed or defined and some measures may not be capable of including or excluding appropriate levels or stages of the health problem. If the outcomes were assessed based on existing definitions or diagnostic criteria, then the answer to this question is likely to be yes. If the outcomes were assessed using observer reported, or self-reported scales, the risk of over- or under-reporting is increased, and objectivity is compromised. Importantly, determine if the measurement tools used were validated instruments as this has a significant impact on outcome assessment validity.

Q4. Did the case series have consecutive inclusion of participants?

Studies that indicate a consecutive inclusion are more reliable than those that do not. For example, a case series that states ‘we included all patients (24) with osteosarcoma who presented to our clinic between March 2005 and June 2006’ is more reliable than a study that simply states ‘we report a case series of 24 people with osteosarcoma.’

Q5. Did the case series have complete inclusion of participants?

The completeness of a case series contributes to its reliability. Studies that indicate a complete inclusion are more reliable than those that do not. A stated above, a case series that states ‘we included all patients (24) with osteosarcoma who presented to our clinic between March 2005 and June 2006’ is more reliable than a study that simply states ‘we report a case series of 24 people with osteosarcoma.’

Q6. Was there clear reporting of the demographics of the participants in the study?

The case series should clearly describe relevant participant’s demographics such as the following information where relevant: participant’s age, sex, education, geographic region, ethnicity, time period, education.

Q7. Was there clear reporting of clinical information of the participants?

There should be clear reporting of clinical information of the participants such as the following information where relevant: disease status, comorbidities, stage of disease, previous interventions/treatment, results of diagnostic tests, etc.

Q8. Were the outcomes or follow-up results of cases clearly reported?

The results of any intervention or treatment should be clearly reported in the case series. A good case study should clearly describe the clinical condition post-intervention in terms of the presence or lack of symptoms. The outcomes of management/treatment when presented as images or figures can help in conveying the information to the reader/clinician. It is important that adverse events are clearly documented and described, particularly a new or unique condition is being treated or when a new drug or treatment is used. In addition, unanticipated events, if any that may yield new or useful information should be identified and clearly described.

Q9. Was the statistical analysis appropriate?

As with any consideration of statistical analysis, consideration should be given to whether there was a more appropriate alternate statistical method that could have been used. The methods section of studies should be detailed enough for reviewers to identify which analytical techniques were used and whether these were suitable.

3. Case-Control Studies

JBI critical appraisal checklist for case-control studies

Amendments to the checklist: Screening questions added.

Screening Question	Yes	No	Unclear	N/A
S1. Are there clear research questions, aims and/ or objectives?	☐	☐	☐	☐
S2. Do the collected data answer the research questions, aims and/ or objectives?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
1. Were the groups comparable other than the presence of disease in cases or the absence of disease in controls?	☐	☐	☐	☐
2. Were cases and controls matched appropriately?	☐	☐	☐	☐
3. Were the same criteria used for identification of cases and controls?	☐	☐	☐	☐
4. Was exposure measured in a standard, valid and reliable way?	☐	☐	☐	☐
5. Was exposure measured in the same way for cases and controls?	☐	☐	☐	☐
6. Were confounding factors identified?	☐	☐	☐	☐
7. Were strategies to deal with confounding factors stated?	☐	☐	☐	☐
8. Were outcomes assessed in a standard, valid and reliable way for cases and controls?	☐	☐	☐	☐
9. Was the exposure period of interest long enough to be meaningful?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
10. Was appropriate statistical analysis used?	☐	☐	☐	☐

Guidance

Q1. Were the groups comparable other than presence of disease in cases or absence of disease in controls?

The control group should be representative of the source population that produced the cases. This is usually done by individual matching; wherein controls are selected for each case on the basis of similarity with respect to certain characteristics other than the exposure of interest. Frequency or group matching is an alternative method. Selection bias may result if the groups are not comparable.

Q2. Were cases and controls matched appropriately?

As in item 1, the study should include clear definitions of the source population. Sources from which cases and controls were recruited should be carefully looked at. For example, cancer registries may be used to recruit participants in a study examining risk factors for lung cancer, which typify population-based case control studies. Study participants may be selected from the target population, the source population, or from a pool of eligible participants (such as in hospital-based case control studies).

Q3. Were the same criteria used for identification of cases and controls?

It is useful to determine if patients were included in the study based on either a specified diagnosis or definition. This is more likely to decrease the risk of bias. Characteristics are another useful approach to matching groups, and studies that did not use specified diagnostic methods or definitions should provide evidence on matching by key characteristics. A case should be defined clearly. It is also important those relating to diagnosis of the disease that controls must fulfil all the eligibility criteria defined for the cases except for.

Q4. Was exposure measured in a standard, valid and reliable way?

The study should clearly describe the method of measurement of exposure. Assessing validity requires that a 'gold standard' is available to which the measure can be compared. The validity of exposure measurement usually relates to whether a current measure is appropriate or whether a measure of past exposure is needed.

Case control studies may investigate many different ‘exposures’ that may or may not be associated with the condition. In these cases, reviewers should use the main exposure of

interest for their review to answer this question when using this tool at the study level. Reliability refers to the processes included in an epidemiological study to check repeatability of measurements of the exposures. These usually include intra-observer reliability and inter-observer reliability.

Q5. Was exposure measured in the same way for cases and controls?

As in item 4, the study should clearly describe the method of measurement of exposure. The exposure measures should be clearly defined and described in detail. Assessment of exposure or risk factors should have been carried out according to same procedures or protocols for both cases and controls.

Q6. Were confounding factors identified?

Confounding has occurred where the estimated intervention exposure effect is biased by the presence of some difference between the comparison groups (apart from the exposure investigated/of interest). Typical confounders include baseline characteristics, prognostic factors, or concomitant exposures (e.g. smoking). A confounder is a difference between the comparison groups and it influences the direction of the study results. A high quality study at the level of case control design will identify the potential confounders and measure them (where possible). This is difficult for studies where behavioral, attitudinal or lifestyle factors may impact on the results.

Q7. Were strategies to deal with confounding factors stated?

Strategies to deal with effects of confounding factors may be dealt within the study design or in data analysis. By matching or stratifying sampling of participants, effects of confounding factors can be adjusted for. When dealing with adjustment in data analysis, assess the statistics used in the study. Most will be some form of multivariate regression analysis to account for the confounding factors measured. Look out for a description of statistical methods as regression methods such as logistic regression are usually employed to deal with confounding factors/ variables of interest.

Q8. Were outcomes assessed in a standard, valid and reliable way for cases and controls?

Read the methods section of the paper. If for e.g. lung cancer is assessed based on existing definitions or diagnostic criteria, then the answer to this question is likely to be yes. If lung cancer is assessed using observer reported, or self-reported scales, the risk of over- or under-reporting is increased, and objectivity is compromised. Importantly, determine if the measurement tools used were validated instruments as this has a significant impact on outcome assessment validity.

Having established the objectivity of the outcome measurement (e.g. lung cancer) instrument, it's important to establish how the measurement was conducted. Were those

involved in collecting data trained or educated in the use of the instrument/s? (e.g. radiographers). If there was more than one data collector, were they similar in terms of level of education, clinical or research experience, or level of responsibility in the piece of research being appraised?

Q9. Was the exposure period of interest long enough to be meaningful?

It is particularly important in a case control study that the exposure time was sufficient enough to show an association between the exposure and the outcome. It may be that the exposure period may be too short or too long to influence the outcome.

Q10. Was appropriate statistical analysis used?

As with any consideration of statistical analysis, consideration should be given to whether there was a more appropriate alternate statistical method that could have been used. The methods section should be detailed enough for reviewers to identify which analytical techniques were used (in particular, regression or stratification) and how specific confounders were measured.

For studies utilising regression analysis, it is useful to identify if the study identified which variables were included and how they related to the outcome. If stratification was the analytical approach used, were the strata of analysis defined by the specified variables? Additionally, it is also important to assess the appropriateness of the analytical strategy in terms of the assumptions associated with the approach as differing methods of analysis are based on differing assumptions about the data and how it will respond.

4. Cohort Studies

JBI critical appraisal checklist for cohort studies

Amendments to the original checklist: Screening questions added. Question 1 from the original JBI checklist was divided into two separate questions to address both aspects individually. Guidance was also amended to address this change.

Screening Question	Yes	No	Unclear	N/A
S1. Are there clear research questions, aims and/ or objectives?	☐	☐	☐	☐
S2. Do the collected data answer the research questions, aims and/ or objectives?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
1. Were the two groups similar?	☐	☐	☐	☐
2. Were the two groups recruited from the same population?	☐	☐	☐	☐
3. Were the exposures measured similarly to assign people to both exposed and unexposed groups?	☐	☐		
4. Was the exposure measured in a valid and reliable way?	☐	☐	☐	☐
5. Were confounding factors identified?				
6. Were strategies to deal with confounding factors stated?	☐	☐	☐	☐
7. Were the groups/participants free of the outcome at the start of the study (or at the moment of exposure)?				
8. Were the outcomes measured in a valid and reliable way?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
9. Was the follow-up time reported and sufficient to be long enough for outcomes to occur?	☐	☐	☐	☐
10. Was the follow-up complete, and if not, were the reasons to loss to follow-up described and explored?	☐	☐	☐	☐
	☐	☐	☐	
12. Was appropriate statistical analysis used?	☐	☐	☐	☐

Guidance

Q1. Were the two groups similar?

Check the paper carefully for descriptions of participants to determine if patients within and across groups have similar characteristics in relation to exposure (e.g. risk factor under investigation). The two groups selected for comparison should be as similar as possible in all characteristics except for their exposure status, relevant to the study in question. The authors should provide clear inclusion and exclusion criteria that they developed prior to recruitment of the study participants.

Q2. Were the two groups recruited from the same population?

Check whether the participants in both groups (exposed and unexposed) were drawn from the same population. This ensures that both groups are comparable and that differences in outcomes are not due to differences in population characteristics. Ideally, the study should describe the source and selection process for participants in both groups in detail. If the two groups were recruited from different populations, this could introduce selection bias, impacting the validity of the study's findings. Look for information on how participants were selected, any inclusion or exclusion criteria, and whether recruitment was conducted in similar settings and time periods for both groups.

Q3. Were the exposures measured similarly to assign people to both exposed and unexposed groups?

A high-quality study at the level of cohort design should mention or describe how the exposures were measured. The exposure measures should be clearly defined and described in detail. This will enable reviewers to assess whether or not the participants received the exposure of interest.

Q4. Was the exposure measured in a valid and reliable way?

The study should clearly describe the method of measurement of exposure. Assessing validity requires that a 'gold standard' is available to which the measure can be compared. The validity of exposure measurement usually relates to whether a current measure is appropriate or whether a measure of past exposure is needed.

Reliability refers to the processes included in an epidemiological study to check repeatability of measurements of the exposures. These usually include intra-observer reliability and inter-observer reliability.

Q5. Were confounding factors identified?

Confounding has occurred where the estimated intervention exposure effect is biased by the presence of some difference between the comparison groups (apart from the exposure investigated/of interest). Typical confounders include baseline characteristics, prognostic factors, or concomitant exposures (e.g. smoking). A confounder is a difference between the comparison groups and it influences the direction of the study results. A high-quality study at the level of cohort design will identify the potential confounders and measure them (where possible). This is difficult for studies where behavioural, attitudinal or lifestyle factors may impact on the results.

Q6. Were strategies to deal with confounding factors stated?

Strategies to deal with effects of confounding factors may be dealt with in the study design or in data analysis. By matching or stratifying sampling of participants, effects of confounding factors can be adjusted for. When dealing with adjustment in data analysis, assess the statistics used in the study. Most will be some form of multivariate regression analysis to account for the confounding factors measured. Look out for a description of statistical methods as regression methods such as logistic regression are usually employed to deal with confounding factors/variables of interest.

Q7. Were the groups/participants free of the outcome at the start of the study (or at the moment of exposure)?

The participants should be free of the outcomes of interest at the start of the study. Refer to the ‘methods’ section in the paper for this information, which is usually found in descriptions of participant/sample recruitment, definitions of variables, and/or inclusion/exclusion criteria.

Q8. Were the outcomes measured in a valid and reliable way?

Read the methods section of the paper. If for e.g. lung cancer is assessed based on existing definitions or diagnostic criteria, then the answer to this question is likely to be yes. If lung cancer is assessed using observer reported, or self-reported scales, the risk of over- or under-reporting is increased, and objectivity is compromised. Importantly, determine if

the measurement tools used were validated instruments as this has a significant impact on outcome assessment validity.

Having established the objectivity of the outcome measurement (e.g. lung cancer) instrument, it is important to establish how the measurement was conducted. Were those involved in collecting data trained or educated in the use of the instrument/s? (e.g. radiographers). If there was more than one data collector, were they similar in terms of level of education, clinical or research experience, or level of responsibility in the piece of research being appraised?

Q9. Was the follow up time reported and sufficient to be long enough for outcomes to occur?

The appropriate length of time for follow up will vary with the nature and characteristics of the population of interest and/or the intervention, disease or exposure. To estimate an appropriate duration of follow up, read across multiple papers and take note of the range for duration of follow up. The opinions of experts in clinical practice or clinical research may also assist in determining an appropriate duration of follow up. For example, a longer timeframe may be needed to examine the association between occupational exposure to asbestos and the risk of lung cancer. It is important, particularly in cohort studies that follow up is long enough to enable the outcomes. However, it should be remembered that the research question and outcomes being examined would probably dictate the follow up time.

Q10. Was follow up complete, and if not, were the reasons to loss to follow up described and explored?

It is important in a cohort study that a greater percentage of people are followed up. As a general guideline, at least 80% of patients should be followed up. Generally, a dropout rate of 5% or less is considered insignificant. A rate of 20% or greater is considered to significantly impact on the validity of the study. However, in observational studies conducted over a lengthy period of time a higher dropout rate is to be expected. A decision on whether to include or exclude a study because of a high dropout rate is a matter of judgement based on the reasons why people dropped out, and whether dropout rates were comparable in the exposed and unexposed groups.

Reporting of efforts to follow up participants that dropped out may be regarded as an indicator of a well conducted study. Look for clear and justifiable description of why people were left out, excluded, dropped out etc. If there is no clear description or a statement in this regard, this will be a 'No'.

Q11. Were strategies to address incomplete follow up utilised?

Some people may withdraw due to change in employment or some may die; however, it is important that their outcomes are assessed. Selection bias may occur as a result of

incomplete follow up. Therefore, participants with unequal follow up periods must be considered in the analysis, which should be adjusted to allow for differences in length of follow up periods. This is usually done by calculating rates which use person-years at risk, i.e. considering time in the denominator.

Q12. Was appropriate statistical analysis used?

As with any consideration of statistical analysis, consideration should be given to whether there was a more appropriate alternate statistical method that could have been used. The methods section of cohort studies should be detailed enough for reviewers to identify which analytical techniques were used (in particular, regression or stratification) and how specific confounders were measured.

For studies utilising regression analysis, it is useful to identify if the study identified which variables were included and how they related to the outcome. If stratification was the analytical approach used, were the strata of analysis defined by the specified variables? Additionally, it is also important to assess the appropriateness of the analytical strategy in terms of the assumptions associated with the approach as differing methods of analysis are based on differing assumptions about the data and how it will respond.

5. Cross-Sectional Analytical Studies

JBI critical appraisal checklist for cross-sectional analytical studies

Amendments to the original checklist: Screening questions added. Question 2 from the original JBI checklist was split into two distinct parts: one addressing the population and the other focusing on the setting. This change reflects the importance of context in the HCS technical team literature reviews and allows for a more targeted assessment of each component. The guidance has also been amended to address this change.

Screening Question	Yes	No	Unclear	N/A
S1. Are there clear research questions, aims and/ or objectives?	☐	☐	☐	☐
S2. Do the collected data answer the research questions, aims and/ or objectives?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
1. Were the criteria for inclusion in the sample clearly defined?	☐	☐	☐	☐
2. Were the study subjects described in detail?	☐	☐	☐	☐
3. Was the setting described in detail?	☐			
4. Was the exposure measured in a valid and reliable way?	☐	☐	☐	☐
5. Were objective, standard criteria used for measurement of the condition?				
6. Were confounding factors identified?	☐	☐	☐	☐
7. Were strategies to deal with confounding factors stated?				
8. Were the outcomes measured in a valid and reliable way?	☐	☐	☐	☐
9. Was appropriate statistical analysis used?	☐	☐	☐	☐

Guidance

Q1. Were the criteria for inclusion in the sample clearly defined?

The authors should provide clear inclusion and exclusion criteria that they developed prior to recruitment of the study participants. The inclusion/exclusion criteria should be specified (e.g., risk, stage of disease progression) with sufficient detail and all the necessary information critical to the study.

Q2. Were the study subjects described in detail?

The study sample should be described in sufficient detail so that other researchers can determine if it is comparable to the population of interest to them. The authors should provide a clear description of the population from which the study participants were selected or recruited, including demographics, location, and time period.

Q3. Was the setting described in detail?

This question focuses on the thoroughness of the description of the physical, social, and cultural setting of the study. Reviewers should assess whether the environmental context is detailed enough to influence the study's outcomes. This includes specifics about, for example, the location (e.g., urban vs. rural), type of facility (e.g., hospital, community centre), and any relevant environmental factors (e.g., facility design, ambient conditions).

Note that the setting may not be relevant for some literature reviews. Reviewers should assess whether the setting could affect the prevalence or manifestation of the condition being studied. If the setting is relevant, check if the description is comprehensive. If the setting is not relevant, and the details of the setting are missing, mark this as 'Not Applicable'.

Q4. Was the exposure measured in a valid and reliable way?

The study should clearly describe the method of measurement of exposure. Assessing validity requires that a 'gold standard' is available to which the measure can be compared. The validity of exposure measurement usually relates to whether a current measure is appropriate or whether a measure of past exposure is needed.

Reliability refers to the processes included in an epidemiological study to check repeatability of measurements of the exposures. These usually include intra-observer reliability and inter-observer reliability.

Q5. Were objective, standard criteria used for measurement of the condition?

It is useful to determine if patients were included in the study based on either a specified diagnosis or definition. This is more likely to decrease the risk of bias. Characteristics are another useful approach to matching groups, and studies that did not use specified

diagnostic methods or definitions should provide evidence on matching by key characteristics

Q6. Were confounding factors identified?

Confounding has occurred where the estimated intervention exposure effect is biased by the presence of some difference between the comparison groups (apart from the exposure investigated/of interest). Typical confounders include baseline characteristics, prognostic factors, or concomitant exposures (e.g. smoking). A confounder is a difference between the comparison groups and it influences the direction of the study results. A high-quality study at the level of cohort design will identify the potential confounders and measure them (where possible). This is difficult for studies where behavioural, attitudinal or lifestyle factors may impact on the results.

Q7. Were strategies to deal with confounding factors stated?

Strategies to deal with effects of confounding factors may be dealt with in the study design or in data analysis. By matching or stratifying sampling of participants, effects of confounding factors can be adjusted for. When dealing with adjustment in data analysis, assess the statistics used in the study. Most will be some form of multivariate regression analysis to account for the confounding factors measured.

Q8. Were the outcomes measured in a valid and reliable way?

Read the methods section of the paper. If for e.g. lung cancer is assessed based on existing definitions or diagnostic criteria, then the answer to this question is likely to be yes. If lung cancer is assessed using observer reported, or self-reported scales, the risk of over- or under-reporting is increased, and objectivity is compromised. Importantly, determine if the measurement tools used were validated instruments as this has a significant impact on outcome assessment validity.

Having established the objectivity of the outcome measurement (e.g. lung cancer) instrument, it's important to establish how the measurement was conducted. Were those involved in collecting data trained or educated in the use of the instrument/s? (e.g. radiographers). If there was more than one data collector, were they similar in terms of level of education, clinical or research experience, or level of responsibility in the piece of research being appraised?

Q9. Was appropriate statistical analysis used?

As with any consideration of statistical analysis, consideration should be given to whether there was a more appropriate alternate statistical method that could have been used. The methods section should be detailed enough for reviewers to identify which analytical techniques were used (in particular, regression or stratification) and how specific confounders were measured.

For studies utilising regression analysis, it is useful to identify if the study identified which variables were included and how they related to the outcome. If stratification was the analytical approach used, were the strata of analysis defined by the specified variables? Additionally, it is also important to assess the appropriateness of the analytical strategy in terms of the assumptions associated with the approach as differing methods of analysis are based on differing assumptions about the data and how it will respond.

6. Cross-Sectional Descriptive Studies

CEBMa critical appraisal checklist for survey studies

Amendments to the checklist: Screening questions were added. Question 1 from the original CEBMa checklist was removed as it is now covered by the screening questions. Question 11 was amended to acknowledge confounders rather than requiring them to be fully accounted for, as descriptive studies typically do not involve rigorous adjustment for confounders, focusing more on description than causality or association. Although confounders can subtly influence descriptive data, recognising their presence can enhance the rigor of the findings, even if controlling for them is not the primary goal.

Additionally, question 12 was removed as it is not consistent with the other checklist and is not related to methodological quality.

Questions specifically related to surveys – questions 6, 7, and 8 –, were revised to accommodate other data collection methods as well.

Due to the lack of instructions in the original checklist, detailed guidance was developed for each question.

Screening Question	Yes	No	Unclear	N/A
S1. Are there clear research questions, aims and/ or objectives?	☐	☐	☐	☐
S2. Do the collected data answer the research questions, aims and/ or objectives?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
1. Is the method of selection of the subjects (employees, teams, divisions, organisations) clearly described?	☐	☐	☐	☐
2. Could the way the sample was obtained introduce (selection) bias?	☐	☐	☐	☐
3. Was the sample of subjects representative with regards to the population to which the findings will be referred?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
4. Was the dataset/ number of observations sufficient enough to draw meaningful conclusions?	☐	☐	☐	☐
	☐	☐	☐	
6. Are the measurements (such as surveys, questionnaires, interviews, observations) likely to be valid and reliable?	☐	☐	☐	☐
7. Was the statistical significance assessed?	☐	☐	☐	
8. Are the confidence intervals given for the main results?	☐	☐	☐	☐

Guidance

Q1. Is the method of selection of the subjects (employees, teams, divisions, organisations) clearly described?

Look for a detailed explanation of how the participants or subjects were chosen for the study. Check if the inclusion and exclusion criteria are clearly stated and whether the selection process (such as random, convenience, purposive) is transparent and justified.

Q2. Could the way the sample was obtained introduce (selection) bias?

Assess whether the sampling method used could lead to selection bias. For example, if the sample was not randomly selected, consider whether certain groups were more likely to be included or excluded, which could skew the results.

Q3. Was the sample of subjects representative with regards to the population to which the findings will be referred?

Determine if the sample accurately reflects the broader population that the researchers intend to generalise the findings to. Consider factors like demographics (age, gender, etc.), geographic region, or other relevant characteristics. If not representative, consider how this might affect the generalisability of the results.

Q4. Was the dataset/ number of observations sufficient enough to draw meaningful conclusions?

If the sample size or number of observations is sufficient to provide reliable results, or if it is too small to detect significant effects, leading to potentially false conclusions. This can be done through sense check of the results: noting that novel studies or those with stringent inclusion and exclusion criteria may naturally have smaller samples. One-way authors might address this is by performing a power analysis before data collection to determine the required sample size. A power analysis helps to calculate the minimum sample size needed to detect a meaningful effect, given the effect size, significance level (usually 0.05), and desired power (typically 80% or 90%).

Q5. Was the dataset complete/all relevant observations recorded and free from missing data?

Determine the dataset is complete and whether missing data might compromise the study's findings. If there was missing data, assess whether appropriate techniques (e.g., imputation) were used to handle it.

Q6. Are the measurements (such as surveys, questionnaires, interviews, observations) likely to be valid and reliable?

For surveys, questionnaires, or interviews, check if the instruments have been validated in previous studies or through testing, ensuring they measure what they intend to (validity) and yield consistent results (reliability). Investigate whether the questions are clear, relevant to the study objectives, and appropriate for the target population.

For records, assess whether the data collection procedures were consistent across cases, and whether the records are accurate and up to date.

For observational studies, confirm that the criteria for observation were well-defined in advance and applied uniformly across all subjects and observers. Check if training or guidelines were given to observers to reduce variability between different observers or across time.

In all cases, consider if the study used methods to test or ensure reliability (e.g., test-retest reliability, inter-rater reliability) and whether the results from these methods are reported.

Q7. Was the statistical significance assessed?

Review the study to see if statistical tests were applied to determine whether the results are statistically significant. This usually involves reporting p-values, which indicate the likelihood that the results occurred by chance. The commonly used threshold for significance is $p < 0.05$.

Ensure the statistical tests used are appropriate for the type of data collected (e.g., t-tests, chi-square tests, ANOVA, regression models). For example, continuous data may require different statistical methods than categorical data

Consider whether the study addresses both statistical and practical significance. A result might be statistically significant but have little real-world impact, so the discussion should ideally reflect the practical importance of the findings.

Verify if the study has taken steps to avoid common errors such as p-hacking (multiple testing without correction) or reporting only statistically significant results while ignoring non-significant ones.

Are the confidence intervals given for the main results?

Check if the study reports confidence intervals (CIs) for key outcomes. Confidence intervals provide a range around the estimate (e.g., mean, odds ratio) that is likely to contain the true population parameter. Look for both the upper and lower bounds of the interval.

Confidence intervals can give more information than p-values alone, showing the precision of the estimate

7. Cross-Sectional Prevalence Studies

JBI critical appraisal checklist for cross-sectional prevalence studies

Amendments to the checklist: Screening questions added. The checklist was amended to split question 4 into two distinct parts: one concerning the population and another focusing on the setting. This change reflects the importance of context in our literature reviews and allows for a more targeted assessment of each component. Guidelines were amended to reflect this change.

Screening Question	Yes	No	Unclear	N/A
S1. Are there clear research questions, aims and/ or objectives?	☐	☐	☐	☐
S2. Do the collected data answer the research questions, aims and/ or objectives?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
1. Was the sample frame appropriate to address the target population?	☐	☐	☐	☐
2. Were study participants sampled in an appropriate way?	☐	☐	☐	☐
3. Was the sample size adequate?	☐			
4. Were the study subjects described in detail?	☐	☐	☐	☐
5. Was the setting described in detail?				
6. Was the data analysis conducted with sufficient coverage of the identified sample?	☐	☐	☐	☐
7. Were valid methods used for the identification of the condition?	☐	☐	☐	☐
8. Was the condition measured in a standard, reliable way for all participants?	☐	☐	☐	☐
9. Was there appropriate statistical analysis?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
10. Was the response rate adequate, and if not, was the low response rate managed appropriately?	☐	☐	☐	☐

Guidance

Q1. Was the sample frame appropriate to address the target population?

This question relies upon knowledge of the broader characteristics of the population of interest and the geographical area. If the study is of women with breast cancer, knowledge of at least the characteristics, demographics and medical history is needed. The term "target population" should not be taken to infer every individual from everywhere or with similar disease or exposure characteristics. Instead, give consideration to specific population characteristics in the study, including age range, gender, morbidities, medications, and other potentially influential factors. For example, a sample frame may not be appropriate to address the target population if a certain group has been used (such as those working for one organisation, or one profession) and the results then inferred to the target population (i.e. working adults). A sample frame may be appropriate when it includes almost all the members of the target population (i.e. a census, or a complete list of participants or complete registry data).

Q2. Were study participants recruited in an appropriate way?

Studies may report random sampling from a population, and the methods section should report how sampling was performed. Random probabilistic sampling from a defined subset of the population (sample frame) should be employed in most cases, however, random probabilistic sampling is not needed when everyone in the sampling frame will be included/ analysed. For example, reporting on all the data from a good census is appropriate as a good census will identify everybody. When using cluster sampling, such as a random sample of villages within a region, the methods need to be clearly stated as the precision of the final prevalence estimate incorporates the clustering effect. Convenience samples, such as a street survey or interviewing lots of people at public gatherings are not considered to provide a representative sample of the base population.

Q3. Was the sample size adequate?

The larger the sample, the narrower will be the confidence interval around the prevalence estimate, making the results more precise. An adequate sample size is important to ensure good precision of the final estimate. Ideally, we are looking for evidence that the authors conducted a sample size calculation to determine an adequate sample size. This will

estimate how many subjects are needed to produce a reliable estimate of the measure(s) of interest. For conditions with a low prevalence, a larger sample size is needed. Also consider sample sizes for subgroup (or characteristics) analyses, and whether these are appropriate. Sometimes, the study will be large enough (as in large national surveys) whereby a sample size calculation is not required. In these cases, sample size can be considered adequate.

When there is no sample size calculation and it is not a large national survey, the reviewers may consider conducting their own sample size analysis using the following formula: (Naing et al. 2006, Daniel 1999)

$$n = \frac{Z^2 P(1-P)}{d^2}$$

Where:

n= sample size

Z = Z statistic for a level of confidence

P = Expected prevalence or proportion (in proportion of one; if 20%, P = 0.2)

d = precision (in proportion of one; if 5%, d=0.05)

Ref:

Naing L, Winn T, Rusli BN. Practical issues in calculating the sample size for prevalence studies Archives of Orofacial Sciences. 2006;1:9-14.

Daniel WW. Biostatistics: A Foundation for Analysis in the Health Sciences.

Edition. 7th ed. New York: John Wiley & Sons. 1999.

Q4. Were the study subjects described in detail?

Certain diseases or conditions vary in prevalence across different populations (e.g. Women vs. Men, sociodemographic variables between countries). The study sample should be described in sufficient detail (e.g. includes age, gender, ethnicity, socioeconomic status, and any specific medical or psychological conditions relevant to the study, etc.) so that other researchers can determine if it is comparable to the population of interest to them.

Q5. Was the setting described in detail?

This question focuses on the thoroughness of the description of the physical, social, and cultural setting of the study. Reviewers should assess whether the environmental context is detailed enough to influence the study's outcomes. This includes specifics about, for example, the location (e.g., urban vs. rural), type of facility (e.g., hospital, community centre), and any relevant environmental factors (e.g., facility design, ambient conditions).

Note that the setting may not be relevant for some literature reviews. Reviewers should assess whether the setting could affect the prevalence or manifestation of the condition being studied. If the setting is relevant, check if the description is comprehensive. If the setting is not relevant, and the details of the setting are missing, mark this as 'Not Applicable'.

Q6. Was data analysis conducted with sufficient coverage of the identified sample?

Coverage bias can occur when not all subgroups of the identified sample respond at the same rate. For instance, you may have a very high response rate overall for your study, but the response rate for a certain subgroup (i.e. older adults) may be quite low.

Q7. Were valid methods used for the identification of the condition?

Here we are looking for measurement or classification bias. Many health problems are not easily diagnosed or defined and some measures may not be capable of including or excluding appropriate levels or stages of the health problem. If the outcomes were assessed based on existing definitions or diagnostic criteria, then the answer to this question is likely to be yes. If the outcomes were assessed using observer reported, or self-reported scales, the risk of over- or under-reporting is increased, and objectivity is compromised. Importantly, determine if the measurement tools used were validated instruments as this has a significant impact on outcome assessment validity.

Q8. Was the condition measured in a standard, reliable way for all participants?

Considerable judgment is required to determine the presence of some health outcomes. Having established the validity of the outcome measurement instrument (see item 6 of this scale), it is important to establish how the measurement was conducted. Were those involved in collecting data trained or educated in the use of the instrument/s? If there was more than one data collector, were they similar in terms of level of education, clinical or research experience, or level of responsibility in the piece of research being appraised? When there was more than one observer or collector, was there comparison of results from across the observers? Was the condition measured in the same way for all participants?

Q9. Was there appropriate statistical analysis?

Importantly, the numerator and denominator should be clearly reported, and percentages should be given with confidence intervals. The methods section should be detailed enough for reviewers to identify the analytical technique used and how specific variables were measured. Additionally, it is also important to assess the appropriateness of the analytical strategy in terms of the assumptions associated with the approach as differing methods of analysis are based on differing assumptions about the data and how it will respond.

Q10. Was the response rate adequate, and if not, was the low response rate managed appropriately?

A large number of dropouts, refusals or “not founds” amongst selected subjects may diminish a study’s validity, as can a low response rates for survey studies. The authors should clearly discuss the response rate and any reasons for non-response and compare persons in the study to those not in the study, particularly with regards to their socio-demographic characteristics. If reasons for non-response appear to be unrelated to the outcome measured and the characteristics of non-responders are comparable to those who do respond in the study (addressed in question 5, coverage bias), the researchers may be able to justify a more modest response rate. Ensure to answer this question with a focus on bias and not response rate only.

8. Expert Opinion

JBI critical appraisal checklist for text and opinion

Amendments to the checklist: Screening questions added.

Screening Question	Yes	No	Unclear	N/A
S1. Are there clear research questions, aims and/ or objectives?	☐	☐	☐	☐
S2. Do the collected data answer the research questions, aims and/ or objectives?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
1. Is the source of the opinion clearly identified?	☐	☐	☐	☐
2. Does the source of the opinion have standing in the field of expertise?	☐	☐	☐	☐
3. Are the interests of the relevant population the central focus of the opinion?	☐	☐		
4. Is the stated position the result of an analytical process, and is there logic in the opinion expressed?	☐	☐	☐	☐
5. Is there reference to the extant literature?	☐			
6. Is any incongruence with the literature/sources logically defended?	☐	☐	☐	☐

Guidance

Q1. Is the source of the opinion clearly identified?

Is there a named author? Unnamed editorial pieces in journals or newspapers, or magasines give broader licence for comment, however authorship should be identifiable.

Q2. Does the source of opinion have standing in the field of expertise?

The qualifications, current appointment and current affiliations with specific groups need to be stated in the publication and the reviewer needs to be satisfied that the author(s) has some standing within the field.

Q3. Are the interests of the relevant population the central focus of the opinion?

The aim of this question is to establish the author's purpose in writing the paper by considering the intended audience. If the review topic is related to a clinical intervention, or aspect of health care delivery, a focus on health outcomes will be pertinent to the review. However, if for example the review is focused on addressing an issue of inter-professional behaviour or power relations, a focus on the relevant groups is desired and applicable. Therefore, this question should be answered in context with the purpose of the review.

Q4. Is the stated position the result of an analytical process, and is there logic in the opinion expressed?

In order to establish the clarity or otherwise of the rationale or basis for the opinion, give consideration to the direction of the main lines of argument. Questions to pose of each textual paper include: What are the main points in the conclusions or recommendations? What arguments does the author use to support the main points? Is the argument logical? Have important terms been clearly defined? Do the arguments support the main points?

Q5. Is there reference to the extant literature?

If there is reference to the extant literature, is it a non-biased, inclusive representation, or is it a non-critical description of content specifically supportive of the line of argument being put forward? These considerations will highlight the robustness of how cited literature was managed.

Q6. Is any incongruence with the literature/sources logically defended?

Is there any reference provided in the text to ascertain if the opinion expressed has wider support? Consider also if the author demonstrated awareness of alternate or dominant opinions in the literature and provided an informed defence of their position as it relates to other or similar discourses.

9. Grey Literature

AACODS (authority, accuracy, coverage, objectivity, date, significance) checklist for evaluation and critical appraisal of such grey literature.

Amendments to the checklist: Screening questions were added; however, for some grey literature reports, these questions may not apply. Summaries were reformulated into questions, and the section related to 'significance' was removed. Narrative-style guidance was developed for each question to provide further clarity.

Screening Question	Yes	No	Unclear	N/A
S1. Are there clear research questions, aims and/ or objectives?	☐	☐	☐	☐
S2. Do the collected data answer the research questions, aims and/ or objectives?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
1. Is the author responsible for the content reputable and authoritative in the field?	☐	☐	☐	☐
2. Is the organisation/group responsible for the content reputable and authoritative in the field?	☐	☐	☐	☐
3. Is the content accurate, reliable, and supported by credible references?	☐			
4. Have limitations been identified?	☐	☐	☐	☐
5. Is the content free from bias and balanced in its presentation?				
6. Is the content current?	☐	☐	☐	☐

Guidance

Q1: Is the author responsible for the content reputable and authoritative in the field?

Check if they are affiliated with a reputable organisation, hold professional qualifications, and are recognised as experts in other sources or cited by others (e.g., citations on Google Scholar). Additionally, determine if the author is a higher degree student under the supervision of an “expert.

Q2: Is the organisation/group responsible for the content reputable and authoritative in the field?

Verify if the organisation/group is well-known and authoritative in the field (e.g., WHO).

Q3: Is the content accurate, reliable, and supported by credible references?

Check for a clearly stated aim or brief and if it is met; verify the presence and adherence to a stated methodology; check if it has edited by a reputable authority, review if the content is supported by credible sources and documented references; check if it is representative of work in the field.

Q4: Have limitations been identified?

Identify any limitations in the content by first determining the scope of the document, including the specific contexts or conditions it addresses, such as regional applicability, specific facilities, technologies, etc. For example, consider if a guidance document excludes new or emerging technologies, evidence, etc, or if a case study focuses on a single, unrepresentative event. Assess whether a manufacturer's recommendation is highly specific to their product line, which may reduce its relevance if the review covers a range of products or systems. Evaluate whether the advice is designed for general use or tailored to particular circumstances, and review any stated limitations, such as disclaimers or conditions that affect the validity of the guidance. Consider any exclusions or gaps, such as scenarios or technologies not covered by the document. Determine how these limitations affect the relevance of the document to the review questions, ensuring that the evidence provided is suitable and applicable.

Q5: Is the content free from bias and balanced in its presentation?

Assess whether the author's perspective or stance on the topic is explicitly stated and easy to understand. Check that the information is presented fairly, considering different perspectives, and that there are no hidden biases influencing the content.

Q6: Is the content current?

Look for a clearly stated date related to the content – no easily discernible date is a strong concern. If no date is given, but can be closely ascertained, is there a valid reason for its absence? Check if the bibliography includes key contemporary materials.

10. Qualitative Studies

Critical Appraisal Skills Programme (CASP) checklist for qualitative research

Amendments to the checklist: Screening questions added. Question 1 and 2 of the CASP checklist were removed as they are covered in the screening questions. Question 3 of the CASP checklist was also removed as the study would be excluded prior to critical appraisal if deemed to not be appropriate study design. Question 10 was removed as it was not deemed relevant to complete the critical appraisal of the study.

Screening Question	Yes	No	Unclear	N/A
S1. Are there clear research questions, aims and/ or objectives?	☐	☐	☐	☐
S2. Do the collected data answer the research questions, aims and/ or objectives?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
1. Was the recruitment strategy appropriate to the aims of the research?	☐	☐	☐	☐
2. Was the data collected in a way that addressed the research issue?	☐	☐	☐	☐
3. Has the relationship between researcher and participants been adequately considered?	☐			
4. Have ethical issues been taken into consideration?	☐	☐	☐	☐
5. Was the data analysis sufficiently rigorous?				
6. Is there a clear statement of findings?	☐	☐	☐	☐

Guidance

Q1. Was the recruitment strategy appropriate to the aims of the research?

Consider:

- If the research has explained how the participants were selected
- If they explained why the participants they selected were the most appropriate to provide access to the type of knowledge sought by the study
- If there are any discussions around recruitment (e.g. why some people chose not to take part)

Q2. Was the data collected in a way that addressed the research issue?

Consider:

- If the setting for the data collection was justified
- If it is clear how data were collected (e.g. focus group, semi-structured interview etc.)
- If the research has justified the methods chosen
- If the researcher has made the methods explicit (e.g. for interview method, is there an indication of how the interviews are conducted, or did they use a topic guide)
- If methods were modified during the study. If so, has the researcher explained how and why
- If the form of data is clear (e.g. tape recordings, video material, notes etc.)
- If the researcher has discussed saturation of data

Q3. Has the relationship between researcher and participants been adequately considered?

Consider:

- If the researcher critically examined their own role, potential bias and influence during (a) formulation of the research questions (b) data collection, including sample recruitment and choice of location
- How the researcher responded to events during the study and whether they considered the implications of any changes in the research design

Q4. Have ethical issues been taken into consideration?

Consider:

- If there are sufficient details of how the research was explained to participants for the reader to assess whether ethical standards were maintained
- If the researcher has discussed issues raised by the study (e.g. issues around informed consent or confidentiality or how they have handled the effects of the study on the participants during and after the study
- If approval has been sought from the ethics committee

Q5. Was the data analysis sufficiently rigorous?

Consider:

- If there is an in depth description of the analysis process
- If thematic analysis is used. If so, is it clear how the categories/themes were derived from the data
- Whether the researcher explains how the data presented were selected from the original sample to demonstrate the analysis process
- If sufficient data are presented to support the findings
- To what extent contradictory data are taken into account
- Whether the researcher critically examined their own role, potential bias and influence during analysis and selection of data for presentation

Q6. Is there a clear statement of findings?

Consider:

- If the findings are explicit
- If there is adequate discussion of the evidence both for and against the researchers arguments
- If the researcher has discussed the credibility of their findings (e.g. triangulation, respondent validation, more than one analyst)
- If the findings are discussed in relation to the original research question

11. Quasi-Experimental Studies

JBI critical appraisal checklist for quasi-experimental studies

Amendments to the checklist: Screening questions added. The checklist was amended to incorporate a question related to potential confounders.

Screening Question	Yes	No	Unclear	N/A
S1. Are there clear research questions, aims and/ or objectives?	☐	☐	☐	☐
S2. Do the collected data answer the research questions, aims and/ or objectives?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
	☐	☐	☐	☐
2. Were the participants included in any comparisons similar?	☐	☐	☐	☐
3. Were the participants included in any comparisons receiving similar treatment/ care, other than the exposure or intervention of interest?	☐			
4. Was there a control group?	☐	☐	☐	☐
5. Were there multiple measurements of the outcome both pre and post the intervention/ exposure?				
6. Was follow-up complete, and if not, were differences between groups in terms of their follow-up adequately described and analysed?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
7. Were the outcomes of participants included in any comparisons measured in the same way?	☐	☐	☐	☐
8. Were outcomes measured in a reliable way?	☐	☐	☐	☐
9. Was appropriate statistical analysis used?	☐	☐	☐	☐
10. Were potential confounders identified and appropriately controlled for in the analysis?	☐	☐	☐	☐

Guidance

Q1. Is it clear in the study what is the 'cause' and what is the 'effect' (i.e. there is no confusion about which variable comes first)?

Ambiguity with regards to the temporal relationship of variables constitutes a threat to the internal validity of a study exploring causal relationships. The 'cause' (the independent variable, that is, the treatment or intervention of interest) should occur in time before the explored 'effect' (the dependent variable, which is the effect or outcome of interest). Check if it is clear which variable is manipulated as a potential cause. Check if it is clear which variable is measured as the effect of the potential cause. Is it clear that the 'cause' was manipulated before the occurrence of the 'effect'?

Q2. Were the participants included in any comparisons similar?

The differences between participants included in compared groups constitute a threat to the internal validity of a study exploring causal relationships. If there are differences between participants included in compared groups there is a risk of selection bias. If there are differences between participants included in the compared groups maybe the 'effect' cannot be attributed to the potential 'cause', as maybe it is plausible that the 'effect' may be explained by the differences between participants, that is, by selection bias. Check the characteristics reported for participants. Are the participants from the compared groups similar with regards to the characteristics that may explain the effect even in the absence of the 'cause', for example, age, severity of the disease, stage of the disease, co-existing conditions and so on? *[NOTE: In one single group pre-test/post-test studies where the patients are the same (the same one group) in any pre-post comparisons, the answer to this question should be 'yes.']*

Q3. Were the participants included in any comparisons receiving similar treatment/care, other than the exposure or intervention of interest?

In order to attribute the 'effect' to the 'cause' (the exposure or intervention of interest), assuming that there is no selection bias, there should be no other difference between the groups in terms of treatments or care received, other than the manipulated 'cause' (the intervention of interest). If there are other exposures or treatments occurring in the same time with the 'cause', other than the intervention of interest, then potentially the 'effect' cannot be attributed to the intervention of interest, as it is plausible that the 'effect' may be explained by other exposures or treatments, other than the intervention of interest, occurring in the same time with the intervention of interest. Check the reported exposures or interventions received by the compared groups. Are there other exposures or treatments occurring in the same time with the intervention of interest? Is it plausible that the 'effect' may be explained by other exposures or treatments occurring in the same time with the intervention of interest?

Q4. Was there a control group?

Control groups offer the conditions to explore what would have happened with groups exposed to other different treatments, other than to the potential 'cause' (the intervention of interest). The comparison of the treated group (the group exposed to the examined 'cause', that is, the group receiving the intervention of interest) with such other groups strengthens the examination of the causal plausibility. The validity of causal inferences is strengthened in studies with at least one independent control group compared to studies without an independent control group. Check if there are independent, separate groups, used as control groups in the study. *[Note: The control group should be an independent, separate control group, not the pre-test group in a single group pre-test post-test design.]*

Q5. Were there multiple measurements of the outcome both pre and post the intervention/exposure?

In order to show that there is a change in the outcome (the 'effect') as a result of the intervention/treatment (the 'cause') it is necessary to compare the results of measurement before and after the intervention/treatment. If there is no measurement before the treatment and only measurement after the treatment is available it is not known if there is a change after the treatment compared to before the treatment. If is multiple measurements are collected before the intervention/treatment implemented then it is possible to explore the plausibility of alternative explanations other than the proposed 'cause' (the intervention of interest) for the observed 'effect', such as the naturally occurring changes in the absence of the 'cause', and changes of high (or low) scores towards less extreme values even in the absence of the 'cause' (sometimes called regression to the mean). If multiple measurements are collected after the intervention/treatment is implemented it is possible to explore the changes of the 'effect' in time in each group and to compare these changes

across the groups. Check if measurements were collected before the intervention of interest was implemented. Were there multiple pre-test measurements? Check if measurements were collected after the intervention of interest was implemented. Were there multiple post-test measurements?

Q6. Was follow up complete and if not, were differences between groups in terms of their follow up adequately described and analysed?

If there are differences with regards to the loss to follow up between the compared groups these differences represent a threat to the internal validity of a study exploring causal effects as these differences may provide a plausible alternative explanation for the observed 'effect' even in the absence of the 'cause' (the treatment or exposure of interest). Check if there were differences with regards to the loss to follow up between the compared groups. If follow up was incomplete (that is, there is incomplete information on all participants), examine the reported details about the strategies used in order to address incomplete follow up, such as descriptions of loss to follow up (absolute numbers; proportions; reasons for loss to follow up; patterns of loss to follow up) and impact analyses (the analyses of the impact of loss to follow up on results). Was there a description of the incomplete follow up (number of participants and the specific reasons for loss to follow up)? If there are differences between groups with regards to the loss to follow up, was there an analysis of patterns of loss to follow up? If there are differences between the groups with regards to the loss to follow up, was there an analysis of the impact of the loss to follow up on the results?

Q7. Were the outcomes of participants included in any comparisons measured in the same way?

If the outcome (the 'effect') is not measured in the same way in the compared groups there is a threat to the internal validity of a study exploring a causal relationship as the differences in outcome measurements may be confused with an effect of the treatment or intervention of interest (the 'cause'). Check if the outcomes were measured in the same way. Same instrument or scale used? Same measurement timing? Same measurement procedures and instructions?

Q8. Were outcomes measured in a reliable way?

Unreliability of outcome measurements is one threat that weakens the validity of inferences about the statistical relationship between the 'cause' and the 'effect' estimated in a study exploring causal effects. Unreliability of outcome measurements is one of different plausible explanations for errors of statistical inference with regards to the existence and the magnitude of the effect determined by the treatment

('cause'). Check the details about the reliability of measurement such as the number of raters, training of raters, the intra-rater reliability, and the inter-raters reliability within the

study (not to external sources). This question is about the reliability of the measurement performed in the study, it is not about the validity of the measurement instruments/scales used in the study. *[Note: Two other important threats that weaken the validity of inferences about the statistical relationship between the 'cause' and the 'effect' are low statistical power and the violation of the assumptions of statistical tests. These other threats are not explored within Question 8, these are explored within Question 9.]*

Q9. Was appropriate statistical analysis used?

Inappropriate statistical analysis may cause errors of statistical inference with regards to the existence and the magnitude of the effect determined by the treatment ('cause'). Low statistical power and the violation of the assumptions of statistical tests are two important threats that weakens the validity of inferences about the statistical relationship between the 'cause' and the 'effect'. Check the following aspects: if the assumptions of statistical tests were respected; if appropriate statistical power analysis was performed; if appropriate effect sizes were used; if appropriate statistical procedures or methods were used given the number and type of dependent and independent variables, the number of study groups, the nature of the relationship between the groups (independent or dependent groups), and the objectives of statistical analysis (association between variables; prediction; survival analysis etc.).

Q10. Were potential confounders identified and appropriately controlled for in the analysis?

For studies where participants are not randomly assigned, it is crucial to look at how the study deals with confounders. These are variables that could incorrectly influence the outcome of the study. Confounders are factors that might independently influence both the intervention and the outcome of the study.

Check if the study clearly identifies confounders. The study should list these variables and describe how it has adjusted for them to ensure that any conclusions drawn about the intervention's effect are valid. This could be through statistical controls like regression where confounders are included as covariates, or by using design methods such as stratification or matching, where participants are grouped based on confounder characteristics to equalise their effects across treatment groups.

12. Randomised Controlled Trial Studies

JBI critical appraisal checklist for randomised control trial studies

Amendments to the checklist: Screening questions added. The 'outcomes' section was removed from questions 7 through 12 to enhance the user-friendliness of the checklist.

Screening Question	Yes	No	Unclear	N/A
S1. Are there clear research questions, aims and/ or objectives?	☐	☐	☐	☐
S2. Do the collected data answer the research questions, aims and/ or objectives?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
1. Was true randomisation used for assignment of participants to treatment groups?	☐	☐	☐	☐
2. Was allocation to treatment groups concealed?	☐	☐	☐	☐
3. Were treatment groups similar at the baseline?	☐	☐	☐	☐
4. Were participants blind to treatment assignment?	☐	☐	☐	☐
5. Were those delivering the treatment blind to treatment assignment?	☐	☐	☐	☐
6. Were treatment groups treated identically other than the intervention of interest?	☐	☐	☐	☐
7. Were outcome assessors blind to treatment assignment?	☐	☐	☐	☐
8. Were outcomes measured in the same way for treatment groups?	☐	☐	☐	☐
9. Were outcomes measured in a reliable way?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
10. Was follow-up complete, and if not, were differences between groups in terms of their follow-up adequately described and analysed?	☐	☐	☐	☐
	☐	☐	☐	
12. Was appropriate statistical analysis used?	☐	☐	☐	☐
	☐	☐	☐	

Guidance

Q1. Was true randomisation used for assignment of participants to treatment groups?

- Category: Internal validity
- Domain: Bias related to selection and allocation
- Appraisal: Study level

If participants are not allocated to treatment and control groups by random assignment there is a risk that this assignment to groups can be influenced by the known characteristics of the participants themselves. These known characteristics of the participants may distort the comparability of the groups (i.e. does the intervention group contain more people over the age of 65 as compared to the control?). A true random assignment of participants to the groups means that a procedure is used that allocates the participants to groups purely based on chance, not influenced by any known characteristics of the participants. Reviewers should check the details about the randomisation procedure used for allocation of the participants to study groups. Was a true chance (random) procedure used? For example, was a list of random numbers used? Was a computer-generated list of random numbers used? Was a statistician, external to the research team consulted for the randomisation sequence generation? Additionally, reviewers should check that the authors are not stating they have used random approaches when they have instead used systematic approaches (such as allocating by days of the week).

Q2. Was allocation to groups concealed?

- Category: Internal validity
- Domain: Bias related to selection and allocation
- Appraisal: Study level

If those allocating participants to the compared groups are aware of which group is next in the allocation process, (i.e., the treatment or control group) there is a risk that they may deliberately and purposefully intervene in the allocation of patients. This may result in the preferential allocation of patients to the treatment group or to the control group. This may directly distort the results of the study, as participants no longer have an equal and random chance to belong to each group compared. Concealment of allocation refers to procedures that prevent those allocating patients from knowing before allocation which treatment or control is next in the allocation process. Reviewers should check the details about the procedure used for allocation concealment. Was an appropriate allocation concealment procedure used? For example, was central randomisation used? Were sequentially numbered, opaque and sealed envelopes used? Were coded drug packs used?

Q3. Were treatment groups similar at the baseline?

- Category: Internal validity
- Domain: Bias related to selection and allocation
- Appraisal: Study level

As with question 1, any differences between the known characteristics of participants included in compared groups constitutes a threat to internal validity. If differences in these characteristics do exist, then there is potential that the ‘effect’ cannot be attributed to the potential ‘cause’ (the examined intervention or treatment). This is because the ‘effect’ may be explained by the differences between participant characteristics and not due to the intervention/treatment of interest. Reviewers should check the characteristics reported for participants. Are the participants from the compared groups similar with regards to the characteristics that may explain the effect even in the absence of the ‘cause’, for example, age, severity of the disease, stage of the disease, co-existing conditions and so on? Reviewers should check the proportions of participants with specific relevant characteristics in the compared groups. [Note: Do NOT only consider the P-value for the statistical testing of the differences between groups with regards to the baseline characteristics.]

Q4. Were participants blind to treatment assignment?

- Category: Internal validity
- Domain: Bias related to administration of intervention/exposure
- Appraisal: Study level

Participants that are aware of their allocation to either the treatment or the control may behave, respond, or react differently to their assigned treatment (or control) than compared to participants that remain unaware of their allocation. Blinding of participants is a technique used to minimise this risk. Blinding refers to procedures that prevent participants from knowing which group they are allocated. If blinding has been followed, participants are not aware if they are in the group receiving the treatment of interest or if they are in any other group receiving the control interventions. Reviewers should check the details reported in the article about the blinding of participants with regards to treatment assignment. Was an appropriate blinding procedure used? For example, were identical capsules or syringes used? Were identical devices used? Be aware of different terms used, blinding is sometimes also called masking.

Q5. Were those delivering the treatment blind to treatment assignment?

- Category: Internal validity
- Domain: Bias related to administration of intervention/exposure
- Appraisal: Study level

Like question 4, those delivering the treatment that are aware of participant allocation to either treatment or control, may treat participants differently than compared to those that remain unaware of participant allocation. There is the risk that any potential change in behaviour may influence the implementation of the compared treatments and the results of the study may be distorted. Blinding of those delivering treatment is used to minimise this risk. When this level of blinding has been achieved, those delivering the treatment are not aware if they are treating the group receiving the treatment of interest or if they are treating any other group receiving the control interventions. Reviewers should check the details reported in the article about the blinding of those delivering treatment with regards to treatment assignment. Is there any information in the article about those delivering the treatment? Were those delivering the treatment unaware of the assignments of participants to the compared groups?

Q6. Were treatment groups treated identically other than the intervention of interest?

- Category: Internal validity
- Domain: Bias related to administration of intervention/exposure
- Appraisal: Study level

To attribute the ‘effect’ to the ‘cause’, (assuming no bias related to selection and allocation) there should be no other difference between the groups in terms of treatment or care received, other than the treatment or intervention controlled by the researchers. If there are other exposures or treatments occurring at the same time with the ‘cause’ (the treatment or intervention of interest), then the ‘effect’ can potentially not be attributed to the examined ‘cause’ (the investigated treatment). This is because it is plausible that the ‘effect’ may be

explained by these other exposures or treatments that occurred at the same time with the 'cause'. Reviewers should check the reported exposures or interventions received by the compared groups. Are there other exposures or treatments occurring at the same time with the 'cause'? Is it plausible that the 'effect' may be explained by other exposures or treatments occurring at the same time with the 'cause'? Is it clear that there is no other difference between the groups in terms of treatment or care received, other than the treatment or intervention of interest?

Q7. Were outcome assessors blind to treatment assignment?

- Category: Internal validity
- Domain: Bias related to assessment, detection and measurement of the outcome
- Appraisal: Outcome level

Like question 4 and 5, those assessing the outcomes that are aware of participant allocation to either treatment or control, may treat participants differently than compared to those that remain unaware of participant allocation. Therefore, there is a risk that the measurement of the outcomes between groups may be distorted, and the results of the study may themselves be distorted. Blinding of outcomes assessors is used in order to minimise this risk. Reviewers should check the details reported in the article about the blinding of outcomes assessors with regards to treatment assignment. Is there any information in the article about outcomes assessors? Were those assessing the treatment's effects on outcomes unaware of the assignments of participants to the compared groups?

Q8. Were outcomes measured in the same way for treatment groups?

- Category: Internal validity
- Domain: Bias related to assessment, detection and measurement of the outcome
- Appraisal: Outcome level

If the outcome is not measured in the same way in the compared groups, there is a threat to the internal validity of a study. Any differences in outcome measurements may be due to the method of measurement employed between the two groups, and not due to the intervention/treatment of interest. Reviewers should check if the outcomes were measured in the same way. Same instrument or scale used? Same measurement timing? Same measurement procedures and instructions?

Q9. Were outcomes measured in a reliable way?

- Category: Internal validity
- Domain: Bias related to assessment, detection and measurement of the outcome
- Appraisal: Outcome level

Unreliability of outcome measurements is one threat that weakens the validity of inferences about the statistical relationship between the 'cause' and the 'effect' estimated in a study

exploring causal effects. Unreliability of outcome measurements is one of the different plausible explanations for errors of statistical inference with regards to the existence and the magnitude of the effect determined by the treatment ('cause'). Reviewers should check the details about the reliability of the measurement used, such as the number of raters, training of raters, the intra-rater and the inter-raters reliability within the study (not as reported in external sources). This question is about the reliability of the measurement performed in the study, it is not about the validity of the measurement instruments/scales used in the study. Finally, some outcomes may not rely on instruments or scales (e.g. death) and reliability of the measurements may need to be assessed in the context of the study being reviewed. [Note: Two other important threats that weaken the validity of inferences about the statistical relationship between the 'cause' and the 'effect' are low statistical power and the violation of the assumptions of statistical tests. These other two threats are explored within Question 12).]

Q10. Was follow up complete and if not, were differences between groups in terms of their follow up adequately described and analysed?

- Category: Internal validity
- Domain: Bias related to participant retention
- Appraisal: Result level

For this question, follow up refers to the period from the moment of randomisation to any point in which the groups are compared during the trial. This question asks if there is complete knowledge (measurements, observations etc.) for the entire duration of the trial for all randomly allocated participants. If there is incomplete follow up from all randomly allocated participants, this is known as post-assignment attrition. As RCTs are not perfect, there is almost always post-assignment attrition, and the focus of this question is on the appropriate exploration of post-assignment attrition. If differences do exist with regards to the post-assignment attrition between the compared groups of an RCT, then there is a threat to the internal validity of that study. This is because these differences may provide a plausible alternative explanation for the observed 'effect' even in the absence of the 'cause' (the treatment or intervention of interest). It is important to note that with regards post-assignment attrition, it is not enough to know the number of participants and the proportions of participants with incomplete data; the reasons for loss to follow up are essential in the analysis of risk of bias.

Reviewers should check if there were differences with regards to the loss to follow up between the compared groups. If follow up was incomplete (incomplete information on all participants), examine the reported details about the strategies used to address incomplete follow up. This can include descriptions of loss to follow up (absolute numbers; proportions; reasons for loss to follow up) and impact analyses (the analyses of the impact of loss to follow up on results). Was there a description of the incomplete follow up including the

number of participants and the specific reasons for loss to follow up? Even if follow up was incomplete, but balanced between groups, if the reasons for loss to follow up are different (e.g., side effects caused by the intervention of interest), these may impose a risk of bias if not appropriately explored in the analysis. If there are differences between groups with regards to the loss to follow up (numbers/proportions and reasons), was there an analysis of patterns of loss to follow up? If there are differences between the groups with regards to the loss to follow up, was there an analysis of the impact of the loss to follow up on the results? [Note: Question 10 is NOT about intention-to-treat (ITT) analysis; question 11 is about ITT analysis.]

Q11. Were participants analysed in the groups to which they were randomised?

- Category: Statistical conclusion validity
- Appraisal: Result level

This question is about the intention-to-treat (ITT) analysis. There are different statistical analysis strategies available for the analysis of data from RCTs, such as intention-to-treat analysis (known also as intent to treat; abbreviated, ITT), per-protocol analysis, and as-treated analysis. In the ITT analysis the participants are analysed in the groups to which they were randomised. This means that regardless of whether participants received the intervention or control as assigned, were complaint with their planned assignment or participated for the entire study duration, they are still included in the analysis. The ITT analysis compares the outcomes for participants from the initial groups created by the initial random allocation of participants to those groups. Reviewers should check if an ITT analysis was reported; check the details of the ITT. Were participants analysed in the groups to which they were initially randomised, regardless of whether they participated in those groups, and regardless of whether they received the planned interventions?

[Note: The ITT analysis is a type of statistical analysis recommended in the Consolidated Standards of Reporting Trials (CONSORT) statement on best practices in trials reporting, and it is considered a marker of good methodological quality of the analysis of results of a randomised trial. The ITT is estimating the effect of offering the intervention, that is, the effect of instructing the participants to use or take the intervention; the ITT it is not estimating the effect of receiving the intervention of interest.]

Q12. Was appropriate statistical analysis used?

- Category: Statistical conclusion validity
- Appraisal: Result level

Inappropriate statistical analysis may cause errors of statistical inference with regards to the existence and the magnitude of the effect determined by the treatment (‘cause’). Low statistical power and the violation of the assumptions of statistical tests are two important threats that weaken the validity of inferences about the statistical relationship between the

'cause' and the 'effect'. Reviewers should check the following aspects: were the assumptions of the statistical tests were respected; if appropriate statistical power analysis was performed; if appropriate effect sizes were used; if appropriate statistical methods were used given the nature of the data and the objectives of statistical analysis (association between variables; prediction; survival analysis etc.).

Q13. Was the trial design appropriate and any deviations from the standard RCT design (individual randomisation, parallel groups) accounted for in the conduct and analysis of the trial?

- Category: Statistical conclusion validity
- Appraisal: Study level

The typical, parallel group RCT may not always be appropriate depending on the nature of the question being asked. Therefore, some additional RCT designs may have been employed that each come with their own additional considerations.

Crossover trials should only be conducted in people with a chronic, stable condition, where the intervention produces a short-term effect (i.e. relief in symptoms). Crossover trials should ensure there is an appropriate period of washout between treatments. This may also be considered under question 6.

Cluster RCTs randomise groups individuals or groups (e.g. communities, wards etc.) forming 'clusters.' When we are assessing outcomes on an individual level in cluster trials, there are unit-of-analysis issues, as individuals within a cluster are correlated. This should be considered by the study authors when conducting analysis, and ideally authors will report the intra-cluster correlation coefficient. This may also be considered under question 12.

Stepped wedge RCTs may be appropriate to establish when and how a beneficial intervention may be best implemented within a defined setting, or due to logistical, practical, or financial considerations in the roll out of a new treatment/intervention. Data analysis in these trials should be conducted appropriately, considering the effects of time. This may also be considered under question 12.

13. Systematic Literature Reviews and Meta-Analysis

JBI critical appraisal checklist for systematic literature reviews

Amendments to the checklist: Screening questions added. A question on protocol registration was incorporated, enhancing review process transparency and accountability. Additionally, the descriptions for Questions 1 and 2 were revised to allow for greater flexibility in the structuring of review questions and enhance user friendliness of the checklist. This modification facilitates the inclusion of diverse types of evidence and research inquiries, extending beyond the traditional PICO framework.

Screening Question	Yes	No	Unclear	N/A
S1. Are there clear research questions, aims and/ or objectives?	☐	☐	☐	☐
S2. Do the collected data answer the research questions, aims and/ or objectives?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
1. Is the review question clearly and explicitly stated?	☐	☐	☐	☐
2. Were the inclusion criteria appropriate for the review question?	☐	☐	☐	☐
3. Was the search strategy appropriate?	☐			
4. Were the sources and resources used to search for studies adequate?	☐	☐	☐	☐
5. Were the criteria for appraising studies appropriate?				
6. Was critical appraisal conducted by two or more reviewers independently?	☐	☐	☐	☐
7. Were there methods to minimise errors in data extraction?	☐	☐	☐	☐
8. Were the methods used to combine studies appropriate?	☐	☐	☐	☐

Checklist Question	Yes	No	Unclear	N/A
9. Was the likelihood of publication bias assessed?	☐	☐	☐	☐
10. Were recommendations for policy and/or practice supported by the reported data?	☐	☐	☐	☐
	☐	☐	☐	
12. Is the review protocol registered, and is this registration appropriately documented and accessible?	☐	☐	☐	☐

Guidance

Q1. Is the review question clearly and explicitly stated?

The review question is an essential step in the systematic review process. A well-articulated question defines the scope of the review and aids in the development of the search strategy to locate the relevant evidence. The review question should be explicitly stated and formulated around the elements of an established question framework such as PICO (Population, Intervention, Comparator, Outcome), PCC (Population, Concept, Context), PEO (Population, Exposure, Outcome), SPIDER (Sample, Phenomenon of Interest, Design, Evaluation, Research type), etc. An explicitly stated question, formulated around its PICO (Population, Intervention, Comparator, Outcome) elements aids both the review team in the conduct of the review and the reader in determining if the review has achieved its objectives. Ideally the review question should be articulated in a published protocol; however, this will not always be the case with many reviews that are located.

Q2. Were the inclusion criteria appropriate for the review question?

The inclusion criteria should be identifiable from and match the review question. The necessary elements of the question framework used should be explicit and clearly defined. The inclusion criteria should be detailed and the included reviews should clearly be eligible when matched against the stated inclusion criteria. Appraisers of meta-analyses will find that inclusion criteria may include criteria around the ability to conduct statistical analyses which would not be the norm for a systematic review. The types of included studies should be relevant to the review question, for example, an umbrella review aiming to summarise a range of effective non-pharmacological interventions for aggressive behaviours amongst elderly patients with dementia will limit itself to including systematic reviews and meta-analyses that synthesise quantitative studies assessing the various interventions; qualitative or economic reviews would not be included.

Q3. Was the search strategy appropriate?

A systematic review should provide evidence of the search strategy that has been used to locate the evidence. This may be found in the methods section of the review report in some cases, or as an appendix that may be provided as supplementary information to the review publication. A systematic review should present a clear search strategy that addresses each of the identifiable PICO components of the review question. Some reviews may also provide a description of the approach to searching and how the terms that were ultimately used were derived, though due to limits on word counts in journals this may be more the norm in online only publications. There should be evidence of logical and relevant keywords and terms and also evidence that Subject

Headings and Indexing terms have been used in the conduct of the search. Limits on the search should also be considered and their potential impact; for example, if a date limit was used, was this appropriate and/or justified? If only English language studies were included, will such a language bias have an impact on the review? The response to these considerations will depend, in part, on the review question.

Q4. Were the sources and resources used to search for studies adequate?

A systematic review should attempt to identify “all” the available evidence and as such there should be evidence of a comprehensive search strategy. Multiple electronic databases should be searched including major bibliographic citation databases such as MEDLINE and CINAHL. Ideally, other databases that are relevant to the review question should also be searched, for example, a systematic review with a question about a physical therapy intervention should also look to search the PEDro database, whilst a review focusing on an educational intervention should also search the ERIC. Reviews of effectiveness should aim to search trial registries. A comprehensive search is the ideal way to minimise publication bias, as a result, a well conducted systematic review should also attempt to search for grey literature, or “unpublished” studies; this may involve searching websites relevant to the review question, or thesis repositories.

Q5. Is the criteria for appraising studies appropriate?

The systematic review should present a clear statement that critical appraisal was conducted and provide the details of the items that were used to assess the included studies. This may be presented in the methods of the review, as an appendix of supplementary information, or as a reference to a source that can be located. The tools or instruments used should be appropriate for the review question asked and the type of research conducted. For example, a systematic review of effectiveness should present a tool or instrument that addresses aspects of validity for experimental studies and randomised controlled trials such as randomisation and blinding – if the review includes observational research to answer the same question a different tool would be more appropriate. Similarly, a review assessing diagnostic test accuracy may refer to the recognised QUADAS[1] tool.

Q6. Was critical appraisal conducted by two or more reviewers independently?

Critical appraisal or some similar assessment of the quality of the literature included in a systematic review is essential. A key characteristic to minimise bias or systematic error in the conduct of a systematic review is to have the critical appraisal of the included studies completed independently and in duplicate by members of the review team. The systematic review should present a clear statement that critical appraisal was conducted by at least two reviewers working independently from each other and conferring where necessary to reach decision regarding study quality and eligibility on the basis of quality.

Q7. Were there methods to minimise errors in data extraction?

Efforts made by review authors during data extraction can also minimise bias or systematic errors in the conduct of a systematic review. Strategies to minimise bias may include conducting all data extraction in duplicate and independently, using specific tools or instruments to guide data extraction and some evidence of piloting or training around their use.

Q8. Were the methods used to combine studies appropriate?

A synthesis of the evidence is a key feature of a systematic review. The synthesis that is presented should be appropriate for the review question and the stated type of systematic review and evidence it refers to. If a meta-analysis has been conducted this needs to be reviewed carefully.

Was it appropriate to combine the studies? Have the reviewers assessed heterogeneity statistically and provided some explanation for heterogeneity that may be present? Often, where heterogeneous studies are included in the systematic review, narrative synthesis will be an appropriate method for presenting the results of multiple studies. If a qualitative review, are the methods that have been used to synthesise findings congruent with the stated methodology of the review? Is there adequate descriptive and explanatory information to support the final synthesised findings that have been constructed from the findings sourced from the original research?

Q9. Was the likelihood of publication bias assessed?

As mentioned, a comprehensive search strategy is the best means by which a review author may alleviate the impact of publication bias on the results of the review. Reviews may also present statistical tests such as Egger's test or funnel plots to also assess the potential presence of publication bias and its potential impact on the results of the review. This question will not be applicable to systematic reviews of qualitative evidence.

Q10. Were recommendations for policy and/or practice supported by the reported data?

Whilst the first nine (9) questions specifically look to identify potential bias in the conduct of a systematic review, the final questions are more indictors of review quality rather than validity. Ideally a review should present recommendations for policy and practice. Where these

recommendations are made there should be a clear link to the results of the review. Is there evidence that the strength of the findings and the quality of the research been considered in the formulation of review recommendations?

Q11. Were the specific directives for new research appropriate?

The systematic review process is recognised for its ability to identify where gaps in the research, or knowledge base, around a particular topic exist. Most systematic review authors will provide some indication, often in the discussion section of the report, of where future research direction should lie. Where evidence is scarce or sample sizes that support overall estimates of effect are small and effect estimates are imprecise, repeating similar research to those identified by the review may be necessary and appropriate. In other instances, the case for new research questions to investigate the topic may be warranted.

Q12. Is the review protocol registered, and is this registration appropriately documented and accessible?

The protocol of the systematic review should be registered with a recognised registry like PROSPERO or an equivalent platform. This should be indicated in the literature review and be accessible. Additionally, the registration should include enough and comprehensive details about review question(s), objectives, and methodology.