INSPECT-SR Guidance

INveStigating ProblEmatic Clinical Trials in Systematic Reviews

Jack Wilkinson Calvin Heal Ella Flemyng Georgios A. Antoniou
Tony Aburrow Zarko Alfirevic Alison Avenell Virginia Barbour
Vincenzo Berghella Dorothy V. M. Bishop Esmée M Bordewijk
Nicholas J. L. Brown Jana Christopher Mike Clarke Jamie Cummins
Darren Dahly Jane Dennis Patrick Dicker Jo Dumville Helen Frankish
Andrew Grey Steph Grohmann Lyle C. Gurrin Jill A. Hayden
James Heathers Kylie E Hunter Ian Hussey Lukas Jung Emily Lam
Toby J. Lasserson Sarah Lensen Tianjing Li Wentao Li Jianping Liu
Elizabeth Loder Andreas Lundh Gideon Meyerowitz-Katz Ben W. Mol
Florian Naudet Anna Noel-Storr Neil E. O'Connell Lisa Parker
Rita F. Redberg Barbara K. Redman Rachel Richardson Anna Lene Seidler
Kyle Sheldrick Emma Sydenham Madelon van Wely Colby J. Vorland
Rui Wang Stephanie Weibel Matthias Wjst Lisa Bero Jamie J. Kirkham

1.0

Table of contents

Introduction

INSPECT-SR (Wilkinson et al., 2025) is a tool for assessing the trustworthiness of randomised controlled trials (RCTs) in systematic reviews. It was developed for health RCTs but has utility in other fields too.

INSPECT-SR does not assess internal or external validity (which are covered by Risk of Bias tools and GRADE), nor does it cover conflicts of interest.

INSPECT-SR contains 21 checks across four domains to help the reviewer make a judgement about whether a study's data and findings can be trusted sufficiently to include it in a research synthesis.

Citation

INSPECT-SR, including this guidance document, should be cited using the reference for the INSPECT-SR manuscript:

Wilkinson JD, Heal C, Flemyng E, Antoniou GA, Aburrow T, Alfirevic Z, Avenell A, Barbour V, Berghella V, Bishop DVM, Bordewijk EM, Brown NJL, Christopher J, Clarke M, Dahly DL, Dennis J, Dicker P, Dumville J, Grey A, Grohmann S, Gurrin LC, Hayden JA, Heathers JAJ, Hunter KE, Hussey I, Jung L, Lam E, Lasserson TJ, Lensen S, Li T, Li W, Liu J, Loder E, Lundh A, Meyerowitz-Katz G, Mol BW, Naudet F, Noel-Storr A, O'Connell N, Parker L, Redberg RF, Redman BK, Richardson R, Seidler AL, Sheldrick K, Sydenham E, van Wely M, Vorland C, Wang R, Weibel S, Wjst M, Bero L, Kirkham JJ. (2025) INSPECT-SR: a tool for assessing trustworthiness of randomised controlled trials. medRxiv 2025.09.03.25334905; doi: 10.1101/2025.09.03.25334905

Contributors

This site is maintained by Ian Hussey and Jack Wilkinson. If you find issues etc, you can open an issue or pull request on the GitHub repository.

Each contributors role is listed below (e.g., contribution to the INSPECT-SR article or the guidance listed on this website).

Contributor	Article	Guidance
Jack Wilkinson		
Calvin Heal		
Ella Flemyng		
Georgios A. Antoniou		
Tony Aburrow		
Zarko Alfirevic		
Alison Avenell		
Virginia Barbour		
Vincenzo Berghella		

Contributor	Article	Guidance
Dorothy V. M. Bishop		
Esmée M Bordewijk		
Nicholas J. L. Brown		
Jana Christopher		
Mike Clarke		
Jamie Cummins		
Darren Dahly		
Jane Dennis		
Patrick Dicker		
Jo Dumville		
Helen Frankish		
Andrew Grey		
Steph Grohmann		
Lyle C. Gurrin		
Jill A. Hayden		
James Heathers		
Kylie E Hunter		
Ian Hussey		
Lukas Jung		
Emily Lam		
Toby J. Lasserson		
Sarah Lensen		
Tianjing Li		
Wentao Li		
Jianping Liu		
Elizabeth Loder		
Andreas Lundh		
Gideon Meyerowitz-Katz		
Ben W. Mol		
Florian Naudet		
Anna Noel-Storr		
Neil E. O'Connell		
Lisa Parker		
Rita F. Redberg		
Barbara K. Redman		
Rachel Richardson		
Anna Lene Seidler		
Kyle Sheldrick		
Emma Sydenham		
Madelon van Wely		
Colby J. Vorland		
Rui Wang		
Stephanie Weibel		
Matthias Wjst		
Lisa Bero		
Jamie J. Kirkham		

Motivation for INSPECT-SR

Systematic reviews exploring health interventions aim to include all eligible studies. This will often involve, but not always exclusively focus on, randomised controlled trials (RCTs). Systematic reviews appraise and synthesise this evidence to arrive at an overall conclusion about whether an intervention is beneficial and whether it causes harm. Problematic studies pose a threat to the evidence synthesis paradigm. These are defined by Cochrane as "any published or unpublished study where there are serious questions about the trustworthiness of the data or findings, regardless of whether the study has been formally retracted". Studies may be problematic because they include some false data or results, or they may be entirely fabricated. Research misconduct is just one possible explanation for false data. Another possibility would be the presence of catastrophic failures in the conduct of the trial, such as miscoding of participants' allocated treatment (e.g., inverting active intervention and placebo groups), failure to properly randomise participants, or severe errors in the analysis code. Whether they are the result of deliberate malpractice or honest error, these issues may not be immediately apparent to journal editors and peer reviewers. Consequently, problematic studies may be published, and subsequently included in systematic reviews. Studies are, of course, routinely appraised on the basis of their methodological validity during the systematic review process. However, these assessments are predicated on the lower-level assumption that the studies and the data they are based on are authentic, and also that the authors did not make any major errors during data collection, analysis or reporting. In fact, many reports of problematic studies describe sound methodology, and so are not currently flagged by critical appraisal tools.

This prompts the question of how we can identify problematic studies. The INSPECT-SR (INveStigating ProblEmatic Clinical Trials in Systematic Reviews) tool has been developed for this purpose, using empirical evidence and consensus methodology. The development process has been described in a protocol paper and in associated results papers. The tool can be used to assess the trustworthiness of RCTs.

Version and feedback

The current version number of the INSPECT-SR guidance document is visible at the top of this page. It is anticipated that the guidance document will be continually revised in response to feedback and ongoing monitoring of use of the tool. This will include the addition of more examples to illustrate both correct and improper use of the checks. Reviewers of the tool may provide anonymous feedback on their experience using this survey. This feedback will be monitored by the INSPECT-SR Study Management Group, and feedback will be used to update the guidance document. However, reviewers should not expect responses to questions posted on the feedback survey.

Overview of the INSPECT-SR tool

The INSPECT-SR tool guides the reviewer through a series of 21 checks in four domains to help them make a judgement about the trustworthiness of a study. In this context, trustworthiness does not encompass internal or external validity, as assessed using Risk of Bias tools and GRADE, nor does it include conflicts of interest. The four domains in the tool are:

Domain	Focus
Domain 1	Inspecting post-publication notices
Domain 2	Inspecting conduct, governance and transparency
Domain 3	Inspecting text and figures
Domain 4	Inspecting results in the study

The checks in each domain assist the systematic reviewer in identifying any domain-level concerns relating to trustworthiness. The reviewer may then use the domain-level judgement to arrive at an overall judgement about the trustworthiness of a trial. We emphasise that concerns about a trial's trustworthiness do not amount

to an accusation of misconduct. INSPECT-SR is not concerned with determining whether inauthentic data have arisen due to deliberate misconduct or honest error.

Application of the INSPECT-SR tool to assess an individual study

Reviewers are advised to use the checks in each domain to arrive at a domain-level judgement about trustworthiness. The tool does not use a prescriptive algorithm to produce a domain-level judgement from the checks in the domain. For each domain, the reviewer records a judgement of **"no concerns"**, **"some concerns"**, or **"serious concerns"**. The purpose of the tool is to identify potentially problematic studies, and a reviewer may decide they have sufficient concerns about a study without having completed all of the checks included in the tool. For efficiency, the reviewer could terminate the assessment of a study if they consider a judgement of "serious concerns" to be warranted at any point during the assessment. This conclusion could be reached before all domains have been assessed. Furthermore, the reviewer might decide that they have serious concerns in relation to a particular domain having completed only a subset of the checks in that domain. In this situation, it would not be necessary to complete the remaining checks in the domain. This differs from Risk of Bias tools, where the expectation is that all domains should be assessed for each study.

A response of "Yes" to an individual check should not generally automatically trigger a judgement of "serious concerns", but may do so if the check reveals a problem that is sufficient to compromise the trustworthiness of the study. Having made a judgement in relation to all four domains (or having reached a judgement of "serious concerns" in relation to any of the first three, if the reviewer has opted to terminate the assessment on this basis) the reviewer is required to make an overall judgement about the trustworthiness of the trial, again using the options "no concerns", "some concerns", or "serious concerns". It is expected that the overall judgement for a trial will typically be at least as severe as the judgement for the domain with the most severe rating. For example, where an assessment has been terminated following a rating of 'serious concerns' or a full assessment has been completed but one domain has been rated as "serious concerns", it is expected that the overall study judgement will also be one of "serious concerns". If the most severe domain-level judgement is "some concerns", then it is expected that the overall study-level judgement should be at least "some concerns", but the cumulative impact of judging there to be "some concerns" in several domains may be sufficient to warrant an overall judgement of "serious concerns" for the study. Reviewers should report the reasons for their domain and study-level judgements, to permit scrutiny. INSPECT-SR assessments should be reported in the appropriate section of the systematic review (e.g. in characteristics of included or excluded studies), to ensure transparency.

A trial may be reported across several publications, including conference abstracts, preprints, protocol papers, and secondary analysis papers. As for Risk of Bias assessment, it will typically be necessary to consider all of the publications describing the trial to obtain the information necessary to complete the checks. The final check in the tool explicitly asks the reviewer to consider whether there are any contradictions in information reported in different publications relating to the index trial.

Incorporating the INSPECT-SR tool into the systematic review process

As is expected for other aspects of a systematic review, such as Risk of Bias assessment and data extraction, it is recommended that two reviewers first apply INSPECT-SR independently before comparing their assessments and reaching agreement in relation to any areas of disagreement. To address some domains content knowledge is necessary (for example, to consider items relating to biological plausibility). For other domains, a level of statistical competence, if not expertise, will likely be needed to consider numerical data from the study. Some familiarity with clinical trial design, conduct and analysis will be useful in making judgements. We recommend that the INSPECT-SR tool is applied prior to Risk of Bias assessment of eligible randomised trials, regardless of which Risk of Bias tool is used for this purpose. This is because any trial judged to warrant "serious concerns", should not be included in the review, meaning there is no need to assess Risk of Bias for these trials.

Trials that receive a study-level judgement of "some concerns" should not be automatically excluded from the review or left out of the synthesis entirely, but it is recommended that these studies should be subjected to sensitivity analysis to determine their influence on the results and conclusions of the review. As for Risk of Bias assessment, there are several possibilities for how this could be operationalised, and systematic review teams should specify their approach at the protocol stage. Possibilities include: 1) restricting the primary analysis to trials with a study-level judgement of "no concerns", and including studies with a study-level judgement of "some concerns" in a sensitivity analysis; and 2) including studies with a study-level judgement of "some concerns" in the primary analysis with studies with "no concerns", and then performing a sensitivity analysis restricted to studies with a study-level judgement of "no concerns".

A judgement of "serious concerns" is intended to be used only when there are features of the study that, either alone or in combination, warrant serious doubts about the trustworthiness of the study. On reaching a judgement of "serious concerns", we recommend reviewing the check responses leading to this judgement to confirm that it is reasonable and justifiable as there may be an alternative explanation, such as journal word limits restricting what can be reported. "Serious concerns" should only be used if it is clear, beyond reasonable doubt, that there truly are serious concerns. As with other aspects of data extraction and critical appraisal of studies during the conduct of a systematic review, correspondence with study authors may be useful to clarify points of uncertainty in relation to trustworthiness assessment. If the individual participant data (IPD) can be accessed, it is possible to perform a more thorough trustworthiness assessment. If there are uncertainties remaining after application of INSPECT-SR, it is recommended to request the IPD from study authors in order to confirm or assuage concerns. An extension to INSPECT-SR that can be used for this purpose, INSPECT-IPD, is in development. Templates for reporting trustworthiness concerns to journals are available on Cochrane's website.

Guidance on the use of individual checks in the INSPECT-SR tool

The following describes key points to consider in relation to individual checks in the INSPECT-SR tool. Illustrative examples are provided. There are four response options corresponding to each check: **"Yes"**, **"No"**, **"Unclear"**, **"Not Applicable"**. Checks have been worded such that a positive response ("Yes") corresponds to a potentially problematic feature of a study. Domain-level judgements do not follow from check responses in a deterministic or algorithmic fashion. For example, we have not specified a threshold corresponding to the number of checks that should be answered "Yes" to trigger domain-level judgements of "serious concerns". The purpose of the checks is to help the reviewer reach a domain-level judgement about whether or not they have concerns about trustworthiness, and to articulate a basis for that judgement. In some cases, a "Yes" response for a single check might be sufficient to trigger serious concerns for the domain and therefore for the study, depending on the nature and extent of the problem identified. However, a reviewer should not automatically assign a judgement of "serious concerns" on the basis that one or more checks were answered "Yes".

It is necessary to record and report some explanatory text detailing the reason for each check response in the appropriate section of the review. Depending on whether or not the trial is included or excluded, this should be in the characteristics of included or excluded studies tables. Some of the checks may be difficult to assess without topic expertise, and a comprehensive assessment is likely to require input from a review team with content and statistical method expertise; although the tool does not require any advanced statistical analyses, the ability to perform and interpret some basic statistical tests is required for some checks, and an understanding of key features of clinical trial processes, such as randomisation procedures, is useful to assess compatibility between reported methods and results. It is recommended to have two team members undertake the assessment independently, and to then discuss and agree on judgements. It might be useful to have team members with complementary skill sets undertake the assessment, as problems might be detected by a reviewer with a particular area of expertise. We use the term "index study" to refer to the study being assessed using the INSPECT-SR tool.

As the field evolves, freely available automated or AI-driven tools to facilitate (some of) these checks may become available. We recommend reviewers are cautious and critique these tools before use to ensure they are fit for purpose. Consider public details about the automation embedded within the tools, clear terms

and conditions for use, public and transparent evaluations, and clarity on the strengths, limitations, biases and generalisability. It is recommended that any automation or AI used in evidence synthesis adheres to the RAISE (Responsible AI in Evidence Synthesis) recommendations and guidance for responsible AI use in evidence synthesis. If reviewers would like to suggest automated or AI-driven tools that meet these standards and could be added to this guidance, please provide details using this survey.

Resources

- Editable template — a Word document template for recording and reporting INSPECT-SR assessments
- Feedback survey — provide anonymous feedback on your experience with the tool
- Retraction Watch database — search for retractions and post-publication notices
- PubPeer — post-publication peer review comments
- Cochrane implementation guidance — templates for reporting trustworthiness concerns to journals
- trustworthy.scientific.claims — A hub for forensic meta-science tool development

Domain 1: Inspecting post-publication notices

1.1. Does the study have an associated retraction?

- Reviewers should check whether the publication or publications describing the study have been retracted. Retractions highlight that there is a significant issue with the publication, such that the journal no longer stands by the article; once retracted, an article should no longer be considered part of the published record. The Committee for Publication Ethics provides guidelines on the reasons for retraction, though be aware that the content of retraction notices can be limited and not representative of all the journal's concerns relating to the reasons for retraction.
- In rare instances, an article might be removed or withdrawn, rather than retracted, meaning that the article itself should no longer be available. This may occur when there has been a breach of confidentiality, publication of libellous content, or copyright or Intellectual Property infringement (for example). Removed or withdrawn articles should be treated in the same manner as retracted articles when using INSPECT-SR.
- If a study report is retracted, it should be marked as such on the journal website, but not removed, and there should be a separate published retraction notice, also with a unique Digital Object Identifier (DOI). It is important to note that these processes may not be carried out in a systematic fashion across all journals nor indexed well by bibliographic databases.
- Checking for retractions should be performed by accessing the online version of the publication on the journal website, where online notices may be found which have not been indexed elsewhere, and by searching the Retraction Watch database, which is the largest and most reliable database of retractions. Be aware that there may be typographic errors in the Retraction Watch database, so DOI may be more useful for searching. Reviewers should confirm that they are searching the Retraction Watch database rather than the associated Retraction Watch blog (i.e. that they are not using the search function at https://retractionwatch.com/ — this does not perform a search of the Retraction Watch database and is not a suitable approach to conducting this check).
- It is recommended to repeat the search shortly before finishing the systematic review, to identify any retractions issued during the conduct of the review.
- Further guidance for searching for post-publication amendments, including retractions, is included in the Cochrane Handbook for Systematic Reviews of Interventions and its associated technical supplement. You may wish to consult with an Information Specialist if further assistance is needed.
- An answer of "yes" for this check would typically warrant a judgement of "serious concerns" for this domain and overall for the index study (i.e., the reviewer would not need to continue with other INSPECT-SR checks), regardless of the reason for retraction and particularly if it was for the main results paper. If the main results paper associated with a study has been retracted, a judgement of "serious concerns" will typically be warranted, regardless of the reason for the retraction. An exception would be where a study has been retracted and subsequently replaced by a new version (e.g. to correct an error). The replacement can then be assessed using INSPECT-SR.

Example of check 1.1

While assessing a clinical trial of a probiotic supplement for gestational diabetes, a systematic reviewer navigates to the journal website. The article has been replaced with a retraction notice, noting that an external statistical review had been performed on the basis of "significant concerns…about the integrity of the data" raised by a third party. The notice states "the main outcome of the external review was that the article's conclusions are unreliable". The reviewer answers "yes" for this check, and this is sufficient to assign a judgement of "serious concerns" for the domain and for the study.

1.2. Does the study have an associated expression of concern or other relevant post-publication notice?

- Reviewers should check whether the publication or publications describing the study have associated expressions of concern or other post publication notices. Expressions of concern and other post publication notices, such as notifications, publisher notes, editor notes, etc., are not used and published as consistently as retractions. Expressions of concern are generally used when there are concerns raised about a publication but the evidence is inconclusive or the issue unresolved. Other notifications may be used to flag a potential issue or provide status updates.
- Because of this variability, the content and purpose of the notice should be carefully considered when making a judgement. Similarly to retractions, be aware that the content of expressions of concern or other notifications can also be limited and not representative of all the journal's concerns relating to the issue.
- If the notice indicates that there is an ongoing investigation then it is recommended that reviewers revisit the journal website to check for any updates prior to finishing the systematic review.
- Expressions of concern can be checked while looking at whether publications associated with the study have been retracted, by checking the journal website and Retraction Watch database.
- In addition to post-publication amendments and notices issued by journals or publishers, post-publication comments and critiques posted by researchers in the form of letters to the editor or posts on PubPeer (for example) relating to trustworthiness should also be considered. It is recommended to look for correspondence relating to the trial publication by citation searching in Web of Science or Scopus relating to the journal of trial publication. Searches of PubMed and Medline are recommended. It is also recommended to look for comments on PubPeer. These are readily accessed by downloading the PubPeer plugin, which will automatically flag a study with comments if you are examining it. The presence of critical correspondence or PubPeer comments should not automatically trigger concerns however, because some critiques of this nature may lack merit, or may not relate to trustworthiness of the study. We recommend that these comments should be carefully considered, as they might assist the reviewer in completing their assessment using the INSPECT-SR tool (for example, by directing attention to a problematic feature that can then be incorporated into the corresponding domain-level judgement).
- Further guidance for searching for post-publication amendments, including expressions of concern, is included in the Cochrane Handbook for Systematic Reviews of Interventions and its associated technical supplement. You may wish to consult with an Information Specialist if further assistance is needed.
- The answer to this check should contribute to a domain-level judgement.

Examples of check 1.2

1. While assessing a clinical trial of a probiotic supplement for gestational diabetes, a reviewer searches for the study on https://retractiondatabase.org/ by searching on the study DOI. The index study is included in the search results, indicating that it has an associated post-publication notice, labelled as

"Concerns/ Issues About data". Navigating to the article on the journal website reveals an "Editor's Note" reporting that the article is being investigated due to integrity concerns. The reviewer answers "yes" for this check, and this response contributes to the domain-level judgement.

2. While assessing a clinical trial of a weight-loss drug, the reviewer identifies a critical comment on PubPeer. On reviewing the comment, the reviewer learns that the criticism relates to the dose of drug used in the study. Because this is not relevant to the assessment of the study's trustworthiness, and because no other post-publication notices relating to the study were identified, the reviewer answers "no" for this check, and this response contributes to the domain-level judgement.

1.3. Do other studies by the research team highlight causes for concern?

- We suggest the reviewer searches the first, corresponding, and last author (at minimum) on the Retraction Watch database.
- It can be helpful to repeat the search with first names and last names switched, because journals and publishers may transpose names.
- A track record of problems relating to trustworthiness may introduce doubts about the index study.
- The reviewer should pay close attention to the content of any notices associated with the author. For example, a previous retraction due to an honest error may not warrant any concerns based on the author's track record.
- If the reviewer does perform these searches in relation to middle authors, the reviewer should consider whether a track record of integrity problems relating to middle authors on the index study are sufficient to introduce concerns about the trustworthiness of the index study. The reviewer should consider the contribution statement in the manuscript to assist with this decision.
- If comments relating to integrity issues on other studies from the author team are identified in other locations, not originating from the journal or publisher (for example, in a letter to the editor or PubPeer) we suggest that the reviewer considers the content of the comment as it may be useful in helping to identify problematic features of the index study.
- The answer to this check should contribute to a domain-level judgement.

Examples of check 1.3

1. A reviewer performs this check on a trial of a probiotic supplement for gestational diabetes by searching for the first and last author on http://retractiondatabase.org/. The search on the last author returns a large number of notices, including numerous retractions and expressions of concern relating to concerns over data integrity. The reviewer reads some of the associated notices to ensure that they are related to data integrity concerns. The reviewer answers "yes" for this check, and this response contributes to the domain-level judgement.
2. A reviewer performs this check on a trial of a dietary intervention for sleep apnea, by searching for the first and last author on http://retractiondatabase.org/. This identifies a publication describing a similar trial conducted concurrently by the author team that has been retracted. The retraction alludes to a lack of transparency on behalf of the authors but is not entirely clear about the motivation for the retraction. The reviewer answers "yes" for this check, and this response contributes to the domain-level judgement.

Domain 2: Inspecting conduct, governance and transparency

2.1. Are there concerns relating to ethical approval?

- The reviewer should look for details of the ethical approval for the study, which may be included in the study publication(s) or trial registration entry.
- The reviewer should look for a corresponding reference number, details of the panel/board granting approval, and for the date of approval. Where an ethical approval number is available, the reviewer may wish to search for that reference number online to check that it hasn't been taken from another, unrelated study.
- Ideally, the reviewer would check whether the panel/board granting approval exists and has the authority to grant ethical approvals, although this is likely to be difficult in most cases.
- Reporting ethical approval details in publications became more established in recent years, and therefore incomplete details in publications of older studies may not be an indication that the study is problematic.
- Partial reporting of these details could warrant a response of "Unclear".
- The answer to this check should contribute to a domain-level judgement.

Example of check 2.1

A trial manuscript reports that "The study was approved by XXXX University Hospital Research Ethics Committee". No approval number is provided, and there are no details about ethical approval on the trial registration page. An online search for the committee suggests that it does exist. Due to the fact that partial information is provided, the reviewer attempts to contact the committee, but receives no answer. They answer "unclear" for this check, and this response contributes to the domain-level judgement.

2.2. Are there concerns relating to the timing or absence of study registration?

- Absent or retrospective registration makes it difficult to determine whether the reported methods and results are an accurate reflection of a planned programme of work.
- This check speaks of concerns relating to the timing of study registration rather than to the absence of "prospective" (as opposed to "retrospective") registration. If registration occurs shortly after the commencement of participant recruitment (i.e. when only a small proportion of the target sample size has been recruited) it might not strictly be "prospective", but might not warrant concerns, for example. The implications of the timing of the registration should be considered in relation to the particular details of the index study.
- Be aware that study registration only became more established in recent years, and regulations and expectations can differ internationally.
- The answer to this check should contribute to a domain-level judgement.

Example of check 2.2

A clinical trial is registered on ClinicalTrials.gov with a registration number NCTXXXXXXXXX provided in the manuscript. Clicking on the Record History section of the record, we see that Version 1, the earliest version of the registration, was submitted on 6th July 2010. The trial manuscript states that participants were recruited between August 2008 and April 2010. Accordingly, this trial was retrospectively registered. As a result, it is impossible to know whether key features of the trial, such as sample size, outcomes, and eligibility criteria, were prospectively determined. The reviewer answers "yes" for this check, and this response contributes to the domain-level judgement.

2.3. Are there important inconsistencies between the publication and the registration documents?

- Where a trial has been registered (prospectively or otherwise), the reviewer may check for major inconsistencies between the publication(s) and registration entry.
- The reviewer should consider the first version of the registration page and any other versions posted before recruitment started, as well as the history of changes made to the registration page, rather than considering only the latest version.
- Unexplained discrepancies in planned sample size, interventions, study dates, or eligibility criteria could be grounds for concern. However, failure to achieve the planned sample size should not be considered a marker of untrustworthiness; many trials struggle with recruitment and considering adequacy of study sample size is not within the scope of INSPECT-SR.
- Disagreements between study publications and study registration documents posted after the trial start may be particularly concerning.
- The purpose of this check is not to investigate outcome reporting bias, which is covered by existing Risk of Bias frameworks. The focus should therefore be on other key aspects of the trial.
- Major unexplained discrepancies with a trial protocol, where available, could also lead to valid concerns about trustworthiness.
- If there is no study registration, reviewers should answer "Not applicable" for this check.
- The answer to this check should contribute to a domain-level judgement.

Example of check 2.3

A clinical trial has been retrospectively registered on the Iranian Registry of Clinical Trials, with the registration number IRCTXXXXXXXXXX. Despite being retrospectively registered, examination of the revisions made to the record indicates that the description of the control group differed on the registration record compared to the published manuscript, and was amended to match the manuscript over two years after publication. In addition, the recruitment date reported in the published manuscript (April 2016 to September 2016) differs compared to the recruitment dates described on the retrospective registration record (April 2016 to May 2016). The reviewer answers "yes" for this check, and this response contributes to the domain-level judgement.

2.4. Is the recruitment of participants implausible?

- The reviewer should consider the plausibility of recruiting a cohort of the reported size in the reported timeframe, taking care not to confuse the full study duration (which includes follow-up of participants) with the recruitment period. One challenge is that it is often difficult to know which has been reported in a publication.
- This check requires domain knowledge, for example of the prevalence of the condition under study at the time the trial was conducted and an idea of the number of cases likely to be available at the study site(s). In addition it is important to consider factors such as number of staff enrolling participants, and the time window of enrolment (e.g. during a daytime outpatient clinic 5 days a week).
- The numbers of participants screened and consenting to participate should be considered — inspection of a CONSORT diagram is likely to be useful here.
- Failure to achieve a target sample size should not be considered a marker of untrustworthiness; many trials struggle with recruitment and considering adequacy of study sample size is not within the scope of INSPECT-SR.
- The answer to this check should contribute to a domain-level judgement.

Examples of check 2.4

1. A randomized clinical trial investigates outcomes of endovascular repair of ruptured abdominal aortic aneurysm compared to open surgical repair in the catchment area of the authors' institution in Poland with a population of around 500,000. The recruitment period spanned from April 2010 to April 2012 (2 years), with the authors reporting a study population of 2,200 participants with ruptured abdominal aortic aneurysm. Since the incidence of ruptured abdominal aortic aneurysm is generally reported as 5.6 to 17.5 per 100,000 person-years in Western countries, the reported cohort size is judged implausible within the reported time frame, hence the reviewer answers "yes" for this check, and uses this response to inform the domain-level judgement.
2. A trial reported recruitment from April 2018 to September 2020, and reported complete follow-up of all participants for 6 months. This is contradicted by submission of the manuscript to a journal in January 2021. The reviewer answers "yes" for this check, and this response contributes to the domain-level judgement.

2.5. Are there concerns about the plausibility of conducting the study using the reported methods and resources?

- The reviewer should consider the plausibility of implementing the protocol as described, given the study setting and reported resources (staffing, funding).
- The reviewer should consider the number of additional participant visits required compared to usual clinical practice.
- This check requires domain knowledge, for example to understand the time and resources required to administer screening questionnaires and outcome assessments.
- The answer to this check should contribute to a domain-level judgement.

Examples of check 2.5

1. A single-centre trial of people with moderate cognitive impairment reports the use of monthly MRI scans and blood tests for one year to monitor their progress and the reimbursement of patients for their travel time. The trial recruited 312 participants over 18 months but reports that no funding was received for the trial. The reviewer regards both the number of MRI scans and the reimbursement of participants for more than 3700 visits to be implausible in these circumstances and answers "yes" for this check, and this response contributes to the domain-level judgement.
2. The trial investigates quality of life, measured using the Short Form-36 (SF-36) and Vascular Quality of Life questionnaire (VascuQoL), of patients with symptomatic peripheral arterial disease in the lower limbs with supervised exercise compared to endovascular treatment (balloon angioplasty or stenting). The study was conducted over a 6-month period by one researcher who was blinded to the intervention arm across 10 research institutions, and the authors state the questionnaires were delivered and completed in person by the researcher. The authors also state that supervised exercise programmes are not government funded and failed to report funding sources for the conduct of the study. The reviewer judges that it would be impractical for a single researcher to conduct the study at 10 different sites and that without any financial support, conduct of the study in the reported settings would be implausible, and therefore answers "yes" for this check, and this response contributes to the domain-level judgement.

Domain 3: Inspecting text and figures

3.1. Are there concerns relating to duplicated content, such as text or tables, or text that is incompatible with the study?

- Plagiarised text may be detected using plagiarism software, or may be noticed while examining multiple studies included in a review.
- Where software has been used in an attempt to conceal plagiarism (for example, by attempting to generate synonymous wording for the plagiarised text), this may produce 'tortured phrases' such as 'counterfeit consciousness' in place of 'artificial intelligence'. The reviewer should be alert to unusual phrases as a potential marker of plagiarism, while recognising that researchers may legitimately use software to assist with writing when writing in a language that is not their first language.
- Problematic studies may feature text that does not make sense in the context of the study. For example, if an author has unethically copied passages of text from a paper describing another study, they might have accidentally retained text describing features such as the study population or intervention from that study, which might not be consistent with the index study.
- While copying and editing of tables from other papers is known to be a feature of some problematic papers, we cannot currently recommend any software capable of reliably detecting this. The reviewer may however notice duplication of tables while looking across studies included in the systematic review, and any concerns of this nature can be reflected in the response to this check.
- If reviewers do not have access to plagiarism software, and do not identify any other anomalous text, they may wish to respond to this as 'Unclear'.
- Some cases of duplicated text are not necessarily problematic, for example, generic descriptions of methods or when authors recycle text ethically.
- The answer to this check should contribute to a domain-level judgement.

Example of check 3.1

A trial of laparoscopic drilling contained an identical results table to an earlier trial with authors from the same university. The reviewer was able to identify this apparent duplication as both trials were eligible for inclusion in the same systematic review. The reviewer answers "yes" for this check, and this response contributes to the domain-level judgement.

3.2. Is there evidence of manipulation or duplication of figures?

- The reviewer should carefully examine figures for signs of improper manipulation or duplication.
- Known examples include duplication of plots across multiple panels of a multipanel figure, manual addition of false "error bars" to a plot, or shifting one survival curve to create a second curve.
- The answer to this check should contribute to a domain-level judgement.

Example of check 3.2

A published manuscript included a multi-panel bar chart, including six bar charts showing a variety of outcomes measured on two study groups at several timepoints. Although the outcomes were measured on a variety of different scales (such as duration in days, distance in mm, or scores derived from clinical questionnaires), all six bar charts were identical apart from differing axes and titles. The reviewer answers "yes" for the check, and this response contributes to the domain-level judgement.

Domain 4: Inspecting results in the study

4.1. Are there any unexplained discrepancies between reported data and participant eligibility criteria?

- The reviewer should check to see whether any results corresponding to participant characteristics are incompatible with the eligibility criteria.
- It is crucial to check whether an explanation is provided in the manuscript.
- The answer to this check should contribute to a domain-level judgement.

Example of check 4.1

A trial was conducted in postmenopausal women. In the published manuscript, sex was listed in the table of baseline characteristics, and there was a substantial portion of the participants in both study arms who were described as male sex, which does not appear to be consistent with the eligibility criteria. The reviewer answers “yes” for the check, and this response contributes to the domain-level judgement.

4.2. Are numbers of participants allocated to each group implausible given the allocation method?

- The reviewer should check whether the numbers of participants allocated to each group are plausible using the stated method of allocation.
- Simple randomisation usually results in unequal participant numbers allocated to study arms, although it is possible for equal allocation to occur.
- An observed imbalance might be incompatible with the stated method of allocation. For example, in a two-arm trial conducted in a single centre, if blocked randomisation has been implemented with a fixed block size and no stratification, it is not possible for an imbalance to occur that exceeds half the block size.
- The reviewer should be mindful of the possibility that the trial authors have described the allocation method incorrectly.
- A basic understanding of randomisation methods is necessary to perform this check.
- The answer to this check should contribute to a domain-level judgement.

Example of check 4.2

A trial manuscript reported that participants had been allocated in a single centre using blocked randomisation, with a fixed block size of 4 and no stratification. The manuscript reported an imbalance in the number of participants allocated to each arm, with 5 more participants being allocated to the control arm compared to the intervention arm. This is not consistent with the reported method of allocation, because the largest imbalance in group sizes at baseline using a fixed block size of 4 would be half the block size, which is 2. The reviewer answers "yes" for the check, and this response contributes to the domain-level judgement.

4.3. Are any baseline data implausible?

- The reviewer should consider the plausibility of the baseline characteristics.
- 'Plausibility' includes clinical or biological plausibility and numerical plausibility. Domain knowledge is necessary to judge clinical or biological plausibility.
- It is important to remember that participants in a clinical trial may not be representative of any particular patient population, and so characteristics of trial participants are not expected to be "typical". Even if a reasonably representative sample is achieved, random variation means that the characteristics of the sample may not match those of the target population, and this does not indicate a problem.
- Magnitude, frequency, variance, and repetition of values for distinct measurements within a table should be considered.
- Known examples in problematic studies include an excess of even or odd numbers, and an excess of multiples of 5.
- There are proposals to formally assess the degree of balance in baseline characteristics using one of several methods. These balance checks may be useful when applied appropriately by researchers with an understanding of the underlying methodology. However, these methods may malfunction if not used correctly, potentially leading to spurious concerns. As such, the routine use of these methods by non-experts is not recommended at present.
- The routine use of methods to assess digit distribution (for example, for conformity with Benford's Law) is not recommended. Benford's law is not expected to be valid for the majority of variables found in RCT baseline tables.
- The reviewer should consider whether unusual values could be due to reporting errors (for example, standard errors reported instead of standard deviations), which may not warrant concerns about trustworthiness but which would need to be corrected if the reviewer uses the data in a meta-analysis.
- The answer to this check should contribute to a domain-level judgement.

Example of check 4.3

A baseline table contains identical values for means, SDs or range limits for both study groups for nine of 11 reported baseline characteristics, reported to two decimal places. The reviewer judges that this is unlikely to be explained by the method of allocation used in the study (which was simple randomisation) and answers "yes" for this check, and this response contributes to the domain-level judgement.

4.4. Are there any discrepancies between results reported in figures, tables, and text?

- The reviewer should check for contradictions where the same results are reported in multiple places (figures, tables, main text, abstract).
- This includes checking for discrepancies between numbers of participants described and plotted in a figure.
- The answer to this check should contribute to a domain-level judgement.

Example of check 4.4

Mean outcome values are displayed using a bar chart in a published trial. The values shown in the bar chart are clearly not consistent with the corresponding values reported in the text. The reviewer answers "yes" for this check, and this response contributes to the domain-level judgement.

4.5. Are the numbers of participants lost to follow-up implausible?

- The reviewer should consider whether numbers of participants lost to follow-up are plausible. This may require domain knowledge, for example about the plausibility of little or no attrition given the context, condition, follow-up duration, and study protocol.
- The reviewer should also consider the role of incentives to minimise attrition in the study and whether they could explain low rates of attrition.
- It may be useful to consider what level of attrition was anticipated in the sample size calculation reported for the study. For example, if a substantial degree of attrition was anticipated, this may lead to concerns if there was actually little or no attrition in the study and no explanation is provided for this.
- Round, equal numbers of participants lost to follow-up, or numbers lost to follow-up resulting in a perfect match with the planned sample size, may be suggestive of problems, but are unlikely to be sufficient to warrant concerns unless other problematic features are also present.
- The answer to this check should contribute to a domain-level judgement.

Example of check 4.5

In a large multicentre trial of psychotherapy versus usual care for people with long-term depression, participants must attend the trial site every three months over an 18-month period to have their outcomes assessed. The trial manuscript reports that there was no loss to follow-up (all 524 participants at all study sites were retained in the study until the end of follow-up). The reviewer judges this to be very unusual for trials conducted in this population, where attrition rates are typically high even in trials with shorter follow-up durations. The reviewer answers "yes" for this check, and this response contributes to the domain-level judgement.

4.6. Are there any unexplained inconsistencies in the numbers of participants?

- The reviewer should check for unexplained inconsistencies in numbers of participants reported in different parts of the manuscript.
- Care should be taken not to regard as "unexplained" differences due to legitimate reasons such as loss to follow-up or exclusion of participants due to non-adherence.
- Checking the CONSORT diagram is recommended when undertaking this check.
- Large unexplained discrepancies with the planned sample size should be noted.
- The answer to this check should contribute to a domain-level judgement.

Examples of check 4.6

1. A manuscript reports that 100 participants were randomly allocated to treatment or control. This is reported in the text and in the table of baseline characteristics, where frequencies for (exhaustive) categorical variables sum to 100 participants. However, the text and results tables include outcome data for more than 150 participants. Noting the discrepancy, the reviewer answers "yes" for this check, and this response contributes to the domain-level judgement.
2. A manuscript presents a sample size calculation suggesting that 40 participants were to be recruited. The results section of the manuscript describes results for more than twice as many participants as this, with no explanation for the discrepancy. The reviewer answers "yes" for this check, and this response contributes to the domain-level judgement.

4.7. Are any outcome data, including estimated treatment effects, implausible?

- The reviewer should consider the plausibility of the outcome measurement values in each arm and estimated treatment effects.
- 'Plausibility' includes clinical or biological plausibility and statistical plausibility. Domain knowledge is necessary to judge clinical or biological plausibility.
- Magnitude, frequency, variance, and repetition of values for distinct measurements within a table should be considered.
- While the estimated treatment effect is being considered, the reviewer should be careful not to over-interpret the point estimate (typically the observed difference or ratio in a summary of the outcome measure for each of two study groups) without careful consideration of associated statistical measures of uncertainty (confidence intervals, p-values). A large point estimate for a treatment effect is not necessarily unusual if it is accompanied by wide confidence intervals.
- A significant result in a trial of a treatment for which there is no plausible mechanism of effect is not a cause for concern on its own. It is important to remember that a certain number of false positive results (Type 1 errors) are expected in trials of completely ineffective interventions.
- It may be useful to compare the estimated effects and CIs to those from other studies in a meta-analysis, to identify unexplained discrepancies. Meta-analysis may be conducted after trustworthiness assessment has been performed, and so this might not come to light until later in the review process. It might therefore be necessary to revisit the assessment should problems come to light when conducting meta-analysis.
- Duplication of estimated treatment effects between trials may also be judged to be implausible, particularly when this is evident across multiple outcome measures.
- The answer to this check should contribute to a domain-level judgement.

Examples of check 4.7

1. A meta-analysis contains two trials from the same author team. The reviewer notices that the point estimates corresponding to the treatment effects in the two trials are identical. The reviewer compares results for several of the other outcome measures between the two trials and notices that the point estimates are identical, or nearly identical, for all of them. The reviewer judges this to be implausible and answers "yes" for this check, and this response contributes to the domain-level judgement.

2. A meta-analysis contains three trials from the same author team. The results from these three trials are highly divergent from those from ten other trials in the meta-analysis. There is more than a 6-fold difference between the lowest of the lower confidence limits of the three studies compared to the upper confidence limit observed after pooling the remaining ten trials. The reviewer judges this to be implausible and answers "yes" for this check, and this response contributes to the domain-level judgement.

4.8. Are the means and variances of integer data impossible?

- This check only applies to variables that can only take integer values (e.g. 1, 2, 3, 4, …).
- For these variables, only certain values of the mean and standard deviation are possible for a given sample size.
- Consistency of reported means and standard deviations for a given sample size may be assessed using the GRIM and GRIMMER techniques.
- Online implementations of these checks are available at http://nickbrown.fr/GRIM (GRIM only), http://www.prepubmed.org/grimmer/, and https://errors.shinyapps.io/scrutiny/. The last of these offers a convenient interface for checking several sets of summary values at once.
- Percentages can be tested using GRIM (but not GRIMMER) if the sums were derived from integer data, such as the number of patients in a group.
- Measures of time, such as age in years or disease duration in months, may be subjected to GRIM/GRIMMER assessment only if recorded in whole units (e.g. years or months).
- The reviewer should be mindful of the possibility that inconsistencies may be explained by missing data resulting in a reduced number of participants (for example). This may nonetheless be problematic if the intent is to use the inconsistent result in the review, or if inconsistencies are sufficient in number to lead to doubt about the accuracy of results in the study in general.
- Consultation with a statistician may be useful to verify judgement.
- The answer to this check should contribute to a domain-level judgement.

Example of check 4.8

A trial manuscript reports the health of newborn infants using the Apgar score at 1 and 5 minutes. This outcome measure can only take integer values, with a score from 0 to 10. Applying GRIM and GRIMMER, the reviewer concludes that two of the reported mean values (8.96 and 9.96) are not compatible with the reported group size (30 participants), nor is the combination of a mean of 9.93 and a standard deviation of 0.18. The reviewer therefore answers "yes" to this question, since several impossible combinations of sample size and mean or standard deviation have been identified. This response contributes to the domain-level judgement.

Tools for this check

- scrutiny web app — R package implementing GRIM and GRIMMER tests, with a convenient interface for checking multiple values at once
- Nick Brown's GRIM calculator — online GRIM checker
- PrePubMed GRIMMER — online GRIMMER checker

4.9. Are there errors in statistical results?

- The reviewer should check whether results of statistical analyses are consistent with reported summary data. For example, where a t-test has been used, the reviewer can check whether the reported p-value is consistent with the reported group means and standard deviations (e.g., using the online calculator at GraphPad).
- Caution is needed however, as p-values based on tests of continuous variables will not generally be reproducible from rounded summary data. The reviewer should consider whether the p-value is consistent with, rather than exactly reproducible from, the reported summary data.
- Checking consistency of results of non-parametric analyses of continuous measures is typically not possible using the reported summary data.
- For categorical data analysed by chi-squared test or similar, where frequencies are reported, the reviewer can attempt to reproduce the p-value from the reported data, without concerns relating to rounding of summary data.
- The study authors may have used variations of the reported statistical tests. For example, variations of the chi-squared test and t-test are commonly used (e.g. use of Yates' correction, or unequal variances t-tests) and where a discrepancy is found, the reviewer should consider whether this could explain the issue.
- Online calculators can be used to perform this check, for example OpenEpi.
- Following the logic set out above, for continuous data, it may be indicative of problems if the statistical results can all be reproduced exactly from rounded summary data, as this may imply that no underlying dataset was analysed in producing the results.
- Checking a large number of statistical results in a manuscript might not be practicable unless the reviewer has ample time to perform the assessment. If this is the case, it is recommended to check a selection of results from baseline and results tables.
- The reviewer should also be mindful of the possibility that some discrepancies could be caused by undisclosed adjustment for covariates (for example).
- This check, like all others in INSPECT-SR, is a study-level assessment and not an outcome-level assessment. While it is recommended to include the review outcomes when performing this check, it is not generally recommended to restrict the check to these variables. Rather, it is advisable to consider a selection of variables (at least), including a sample of results from a baseline table where possible. The reviewer should not be reassured if they identify no errors in the review outcomes, despite finding errors elsewhere. If a trial has been fabricated, it is possible that the fabricator might focus more attention on the key outcomes of the study than on other incidental variables.
- The answer to this check should contribute to a domain-level judgement.

Examples of check 4.9

1. A manuscript reports results of a t-test for two groups of 30 participants. In group 1, there is a reported mean of 20 and a standard deviation of 4. In group 2, there is a reported mean of 21 and a standard deviation of 2. The p-value is reported as p=0.02. If we try to reproduce the result using the summary

data, we get a p-value of p=0.23, which may appear to contradict the reported result. However, the reported summary data is rounded. We can find the smallest p-value that would be consistent with the reported data by using values that would be rounded to those reported in the paper, while making the difference in means as large as possible and the standard deviations as small as possible. In this case, the actual group means could be 19.5 and 21.449, and the standard deviations 3.5 and 1.5. The p-value in this case would be 0.006, which is clearly smaller than the reported value. The summary data are therefore consistent with the reported p-value. If we wanted to see how large the p-value could be while remaining consistent with the summary data, we would make the means as similar as possible and the standard deviations as large as possible, while ensuring that the values would round to the reported summary data. In this example, the reviewer should answer "no" if they do not identify any errors in statistical results elsewhere.

2. A manuscript reports "sex" as a binary baseline characteristic in Table 1, showing the frequencies of male and female participants in each of the two study groups. This is a 2×2 table, and a chi-squared test could be performed if we wanted to make a comparison between the study groups. This would result in a single p-value. However, in the manuscript, two different p-values are presented; one for male participants and one for female participants. This does not make sense. Moreover, the reviewer performs a chi-squared test, in addition to several plausible alternative tests, and neither of the reported p-values match any of the p-values obtained from these checks. The reviewer answers "yes" for this check, and this response contributes to the domain-level judgement.

Tools for this check

- GraphPad t-test calculator — online t-test calculator
- OpenEpi — open source epidemiologic statistics

4.10. Are any other contradictions implied by the data?

- There may be other instances of contradictory results that would not be detected using the other checks in this domain. Several examples have been identified, but appear to occur too infrequently to warrant routine checking. Should anomalies of this nature be observed, they may be recorded in response to this check. Specifically:
  - Subgroup counts or means could conflict with results for the overall cohort.
  - A mean value (for example) could fall outside of a reported range.
  - Some combinations of outcome are not possible. For example, it is not possible to have more birth events (with one "birth event" defined as the birth of at least one child) than pregnancies.
- The answer to this check should contribute to a domain-level judgement.

Example of check 4.10

A manuscript reports a mean quality of life score for the intervention group overall, and also separately for participants aged 65 or over, or under 65. There are 41 participants in the intervention group overall, with 30 of these being aged less than 65 years and 11 being 65 years or older. The overall mean score is reported as 7.3. The mean score for the 30 participants younger than 65 is reported as 6.2, and the mean score for the 11 participants aged 65 years or older is 7.4. We may attempt to reproduce the overall mean score from the subgroup-specific means, by calculating (6.2×30 + 7.4×11) / 41 = 6.52, which is smaller than the reported value of 7.3. We must confirm that the discrepancy cannot be explained by rounding, by considering the values for the subgroup means that would make the recalculated value as large as possible (e.g. by setting the subgroup means to 6.2499 and 7.499 and recalculating the overall value). This returns a value of 6.57, which is clearly not compatible with the reported value of 7.3, even if the latter value has been rounded up. The reviewer records a response of "yes" for this check, and this response contributes to the domain-level judgement.

4.11. Are there inconsistencies in descriptions of methods and results across publications describing the study?

- The reviewer should check for major unexplained discrepancies between publications associated with the study, such as a conference abstract and a main results paper.
- Conflicting results, group sizes, or descriptions of methods could warrant concerns.
- The answer to this check should contribute to a domain-level judgement.

Example of check 4.11

A reviewer compares a published trial manuscript to a conference abstract reporting the completed trial. The description of the study methods, including eligibility criteria and study dates, are consistent between the manuscript and abstract. However, the sample size reported in the manuscript is considerably larger than that reported in the conference abstract. No explanation for this discrepancy is included in the manuscript. In light of the unexplained discrepancy, the reviewer answers "yes" for this check, and this response contributes to the domain-level judgement.

Appendices

Editable template

An editable Word document template for INSPECT-SR is available for download. The template mirrors the structure of this guidance, providing one page per check with space to record notes, judgements, and justifications for each item.

Download template

License

© Jack Wilkinson (2026)

Text and figures are licensed under a Creative Commons Attribution 4.0 (CC BY 4.0) license.

Code is licensed under the MIT License.

You are free to copy, share, adapt, and reuse the contents of this book — text, figures, and code — for any purpose, including commercial use, provided you cite it.