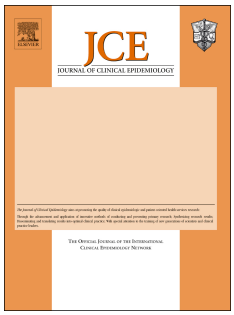


Assessing the feasibility and impact of clinical trial trustworthiness checks via an application to Cochrane Reviews: Stage 2 of the INSPECT-SR project

Jack Wilkinson, Calvin Heal, Georgios A. Antoniou, Ella Flemyng, Love Ahnström, Alessandra Alteri, Alison Avenell, Timothy Hugh Barker, David N. Borg, Nicholas JL. Brown, Rob Buhmann, Jose A. Calvache, Rickard Carlsson, Lesley-Anne Carter, Aidan G. Cashin, Sarah Cotterill, Kenneth Färnqvist, Michael C. Ferraro, Steph Grohmann, Lyle C. Gurrin, Jill A. Hayden, Kylie E. Hunter, Natalie Hyltse, Lukas Jung, Ashma Krishan, Silvy Laporte, Toby J. Lasserson, David RT. Laursen, Sarah Lensen, Wentao Li, Tianjing Li, Jianping Liu, Clara Locher, Zewen Lu, Andreas Lundh, Antonia Marsden, Gideon Meyerowitz-Katz, Ben W. Mol, Zachary Munn, Florian Naudet, David Nunan, Neil E. O'Connell, Natasha Olsson, Lisa Parker, Eleftheria Patetsini, Barbara Redman, Sarah Rhodes, Rachel Richardson, Martin Ringsten, Ewelina Rogozińska, Anna Lene Seidler, Kyle Sheldrick, Katie Stocking, Emma Sydenham, Hugh Thomas, Sofia Tsokani, Constant Vinatier, Colby J. Vorland, Rui Wang, Bassel H. Al Wattar, Florencia Weber, Stephanie Weibel, Madelon van Wely, Chang Xu, Lisa Bero, Jamie J. Kirkham

PII: S0895-4356(25)00157-X

DOI: https://doi.org/10.1016/j.jclinepi.2025.111824

Reference: JCE 111824

To appear in: *Journal of Clinical Epidemiology*

Received Date: 5 December 2024

Revised Date: 24 April 2025

Accepted Date: 5 May 2025

Please cite this article as: Wilkinson J, Heal C, Antoniou GA, Flemyng E, Ahnström L, Alteri A, Avenell A, Barker TH, Borg DN, Brown NJ, Buhmann R, Calvache JA, Carlsson R, Carter L-A, Cashin AG, Cotterill S, Färnqvist K, Ferraro MC, Grohmann S, Gurrin LC, Hayden JA, Hunter KE, Hyltse N, Jung L, Krishan A, Laporte S, Lasserson TJ, Laursen DR, Lensen S, Li W, Li T, Liu J, Locher C, Lu Z, Lundh A, Marsden A, Meyerowitz-Katz G, Mol BW, Munn Z, Naudet F, Nunan D, O'Connell NE, Olsson N, Parker L, Patetsini E, Redman B, Rhodes S, Richardson R, Ringsten M, Rogozińska E, Seidler AL, Sheldrick K, Stocking K, Sydenham E, Thomas H, Tsokani S, Vinatier C, Vorland CJ, Wang R, Al Wattar BH, Weber F, Weibel S, van Wely M, Xu C, Bero L, Kirkham JJ, Assessing the feasibility and impact of clinical trial

trustworthiness checks via an application to Cochrane Reviews: Stage 2 of the INSPECT-SR project, *Journal of Clinical Epidemiology* (2025), doi: https://doi.org/10.1016/j.jclinepi.2025.111824.

Assessing the feasibility and impact of clinical trial trustworthiness checks via an application to Cochrane Reviews: Stage 2 of the INSPECT-SR project

Jack Wilkinson[1*], Calvin Heal[1*], Georgios A Antoniou[2,3], Ella Flemyng[4], Love Ahnström[5], Alessandra Alteri[6], Alison Avenell[7], Timothy Hugh Barker[8], David N Borg[9,10], Nicholas JL Brown[11], Rob Buhmann[12], Jose A Calvache[13,14], Rickard Carlsson[11], Lesley-Anne Carter[1], Aidan G Cashin[15,16], Sarah Cotterill[1], Kenneth Färnqvist[17], Michael C Ferraro[15,16], Steph Grohmann[4], Lyle C Gurrin[18], Jill A Hayden[19], Kylie E Hunter[20], Natalie Hyltse[11], Lukas Jung[21], Ashma Krishan[1], Silvy Laporte[22], Toby J Lasserson[4], David RT Laursen[23,24], Sarah Lensen[25], Wentao Li[26], Tianjing Li[27], Jianping Liu[28], Clara Locher[29], Zewen Lu[1], Andreas Lundh[23,24,30], Antonia Marsden[1], Gideon Meyerowitz-Katz[31], Ben W Mol[25], Zachary Munn[26], Florian Naudet[29], David Nunan[32], Neil E O'Connell[33], Natasha Olsson[5], Lisa Parker[34], Eleftheria Patetsini[35], Barbara Redman[36], Sarah Rhodes[1], Rachel Richardson[37], Martin Ringsten[38], Ewelina Rogozińska[39], Anna Lene Seidler[20], Kyle Sheldrick[40], Katie Stocking[1], Emma Sydenham[41], Hugh Thomas[42], Sofia Tsokani[37,43] Constant Vinatier[29], Colby J Vorland[44], Rui Wang[20], Bassel H Al Wattar[45], Florencia Weber[46], Stephanie Weibel[46], Madelon van Wely[47], Chang Xu[48, 49] , Lisa Bero[50†], Jamie J Kirkham[1†]

*Joint first authorship

†Joint senior authorship

1. Centre for Biostatistics, The University of Manchester, Manchester Academic Health Science Centre, Manchester, UK

2. Manchester Vascular Centre, Manchester University NHS Foundation Trust, Manchester, UK

3. Division of Cardiovascular Sciences, School of Medical Sciences, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK

4. Evidence Production and Methods Directorate, Cochrane Central Executive, London, UK

5. Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

6. Obstetrics and Gynaecology Unit, IRCCS San Raffaele Scientific Institute, Milan, Italy

7. Health Services Research Unit, University of Aberdeen, Aberdeen, UK

8. Health Evidence Synthesis, Recommendations and Impact, School of Public Health, The University of Adelaide, SA, Australia

9. Australian Sports Commission, Australian Institute of Sport, Bruce, Australian Capital Territory, Australia

10. Queensland University of Technology, School of Exercise Nutrition Sciences, Brisbane, Queensland, Australia.

11. Department of Psychology, Linnaeus University, Växjö, Sweden

12. University of the Sunshine Coast, School of Health, Queensland, Australia

13. Department of Anesthesiology, Erasmus MC, The Netherlands

14. Department of Anesthesiology, Universidad del Cauca, Colombia

15. Centre for Pain IMPACT, Neuroscience Research Australia, Randwick, Australia

16. School of Health Sciences, Faculty of Medicine and Health, University of New South Wales Sydney, Australia

17. Department of Molecular medicine and surgery, Karolinska Institute, Stockholm, Sweden

18. School of Population and Global Health, The University of Melbourne

19. Department of Community Health & Epidemiology, Dalhousie University, Halifax, Canada

20. NHMRC Clinical Trials Centre, University of Sydney, Australia

21. Independent researcher, Heidelberg, Germany

22. Clinical Pharmacology Unit, INSERM U1059 Sainbiose, University Hospital of Saint-Etienne, France

23. Cochrane Denmark & Centre for Evidence-Based Medicine Odense (CEBMO), Department of Clinical Research, University of Southern Denmark, Odense, Denmark

24. Open Patient data Explorative Network (OPEN), Odense University Hospital, Odense, Denmark

25. Department of Obstetrics, Gynaecology and Newborn Health, Royal Women's Hospital, University of Melbourne, Melbourne, VIC, Australia

26. National Perinatal Epidemiology and Statistics Unit, Centre for Big Data Research in Health and School of Women's and Children's Health, The University of New South Wales, Sydney, Australia

27. Department of Ophthalmology and Department of Epidemiology, University of Colorado Anschutz Medical Campus, Colorado, USA

28. Centre for Evidence-Based Chinese Medicine, Beijing University of Chinese Medicine, Beijing, China.

29. Univ Rennes, CHU Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR_S 1085, Centre d'investigation clinique de Rennes (CIC1414), F-35000 Rennes, France

30. Department of Respiratory Medicine and Infectious Diseases, Copenhagen University Hospital, Bispebjerg and Frederiksberg, Denmark

31. School of Nursing, University of Wollongong, Australia

32. Centre for Evidence-Based Medicine, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

33. Dept of Health Sciences, Centre for Health and Wellbeing across the Lifecourse, Brunel University London, UK

34. Faculty of Medicine & Health, Charles Perkins Centre, University of Sydney, NSW, Australia

35. York Trials Unit, Department of Health Sciences, University of York, York, UK

36. New York University Grossman School of Medicine, New York, USA

37. Methods Support Unit, Cochrane, UK

38. Cochrane Sweden, Skåne University Hospital, Lund University, Lund, Sweden

39. Meta-Analysis Group, Institute of Clinical Trials and Methodology, MRC Clinical Trials Unit at UCL, London, UK

40. Faculty of Medicine, University of New South Wales, Australia

41. Cochrane Central Editorial Service, London, UK

42. The Lancet Gastroenterology & Hepatology, London, UK

43. Laboratory of Hygiene, Social & Preventive Medicine and Medical Statistics, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

44. Department of Epidemiology and Biostatistics, Indiana University School of Public Health-Bloomington, USA

45. Clinical Trials Unit, Anglia Ruskin University, Chelmsford, UK

46. University Hospital Würzburg, Department of Anaesthesiology, Intensive Care, Emergency and Pain Medicine, Würzburg, Germany

47. Centre for Reproductive Medicine, Department of Obstetrics and Gynaecology, Amsterdam University Medical Center, Netherlands

48. The Third Department of Hepatic Surgery, Eastern Hepatobiliary Surgery Hospital, Third Affiliated Hospital, Second Military Medical University, Naval Medical University, Shanghai, China

49. Proof of Concept Center, Eastern Hepatobiliary Surgery Hospital, Third Affiliated Hospital, Second Military Medical University, Naval Medical University, Shanghai, China

50. University of Colorado Anschutz Medical Campus, Colorado, USA

What is new

- An extensive list of potential checks for assessing study trustworthiness was assessed via an application to 95 randomised controlled trials (RCTs) in 50 Cochrane Reviews.
- Following application of the checks, assessors had concerns about the authenticity of 32% of the RCTs.
- If these RCTs were excluded, 22% of meta-analyses would have no remaining RCTs.
- However, the study showed that some checks were frequently infeasible, and others could be easily misunderstood or misinterpreted.
- The study restricted assessment to meta-analyses including five or fewer RCTs, which might distort the impact of applying the checks.

Abstract

Background

The aim of the INSPECT-SR project is to develop a tool to identify problematic RCTs in systematic reviews. In Stage 1 of the project, a list of potential trustworthiness checks was created. The checks on this list must be evaluated to determine which should be included in the INSPECT-SR tool.

Methods

We attempted to apply 72 trustworthiness checks to RCTs in 50 Cochrane Reviews. For each, we recorded whether the check was passed, failed or possibly failed, or whether it was not feasible to complete the check. Following application of the checks, we recorded whether we had concerns about the authenticity of each RCT. We repeated each meta-analysis after removing RCTs flagged by each check, and again after removing RCTs where we had concerns about authenticity, to estimate the impact of trustworthiness assessment. Trustworthiness assessments were compared to Risk of Bias and GRADE assessments in the reviews.

Results

95 RCTs were assessed. Following application of the checks, assessors had some or serious concerns about the authenticity of 25% and 6% of the RCTs, respectively. Removing RCTs with either some or serious concerns resulted in 22% of meta-analyses having no remaining RCTs. However, many checks proved difficult to understand or implement, which may have led to unwarranted scepticism in some instances. Furthermore, we restricted assessment to meta-analyses with no more than 5 RCTs (54% contained only 1 RCT), which will distort the impact on results. No relationship was identified between trustworthiness assessment and Risk of Bias or GRADE.

Conclusions

This study supports the case for routine trustworthiness assessment in systematic reviews, as problematic studies do not appear to be flagged by Risk of Bias assessment. The study produced evidence on the feasibility and impact of trustworthiness checks. These results will be used, in conjunction with those from a subsequent Delphi process, to determine which checks should be included in the INSPECT-SR tool.

Plain language summary

Systematic reviews collate evidence from randomised controlled trials (RCTs) to find out whether health interventions are safe and effective. However, it is now recognised that the

findings of some RCTs are not genuine, and some of these studies appear to have been fabricated. Various checks for these "problematic" RCTs have been proposed, but it is necessary to evaluate these checks to find out which are useful and which are feasible. We applied a comprehensive list of "trustworthiness checks" to 95 RCTs in 50 systematic reviews to learn more about them, and to see how often performing the checks would lead us to classify RCTs as being potentially inauthentic. We found that applying the checks led to concerns about the authenticity of around 1 in 3 RCTs. However, we found that many of the checks were difficult to perform and could have been misinterpreted. This might have led us to be overly sceptical in some cases. The findings from this study will be used, alongside other evidence, to decide which of these checks should be performed routinely to try to identify problematic RCTs, to stop them from being mistaken for genuine studies and potentially being used to inform healthcare decisions.

MAIN TEXT

Background

Systematic reviews of randomised controlled trials (RCTs) aim to include all trials that address the review question and meet the prespecified eligibility criteria. There is an understanding that RCTs included in a systematic review should be scrutinised for their internal validity, for example, using Risk of Bias tools (1, 2). These assessments require that the reviewer can trust what is written in a trial report to be an authentic account of what took place. However, this no longer appears to be tenable as a default assumption, as recent large-scale assessments have cast doubt on the veracity of many RCTs submitted to journals (3) or published in systematic reviews (4). Recent examples, such as ivermectin for COVID-19, illustrate how the failure to routinely interrogate the authenticity of eligible RCTs in systematic reviews allows fake studies to influence patient care (5).

Cochrane defines '*problematic studies*' as studies where there are 'serious questions about the trustworthiness of the data or findings' (6). Problematic studies could represent instances of academic misconduct such as research fraud, or could be the result of critical errors in trial processes. Cochrane policy, introduced in 2021, states that potentially problematic RCTs should not be included in a systematic review (6, 7). This prompts the question of what criteria could be used to identify problematic studies, which may appear to be high-quality on the basis of traditional Risk of Bias assessment (8). Cochrane's implementation guidance recognises that a number of methods for identifying problematic studies have been proposed, but does not recommend a method at this time.

The aim of the INSPECT-SR (INveStigating ProblEmatic Clinical Trials in Systematic Reviews) project is to develop a tool that can be used by systematic reviewers to assess the trustworthiness of RCTs (9). Several tools have recently been proposed for this purpose (10-14). However, none of these have involved a comprehensive evaluation and subsequent selection of potential trustworthiness checks. In Stage 1 of the development process, we

identified an extensive list of potential trustworthiness checks (15). A tool including all of these checks would not be practicable, and we anticipate that many of the checks will turn out to be infeasible or otherwise not useful. In Stages 2 (application to Cochrane Reviews) and 3 (Delphi survey), the checks on this list will be evaluated to determine which should be included in the final tool. These results will then feed into a series of consensus meetings (Stage 4) which will be used to develop a draft version of the INSPECT-SR tool. The draft tool will then be tested in the assessment of RCTs (Stage 5). Feedback from Stage 5 will be used to finalise the tool. The current study describes Stage 2 of the project, in which the identified checks were applied to RCTs included in a sample of Cochrane Reviews, in order to evaluate their feasibility and impact on review results, and to evaluate how often assessors had concerns about the authenticity of RCTs after applying the checks.

Methods

A protocol describing the INSPECT-SR project methods has previously been published (9). We undertook a large, collaborative project in which assessors applied a series of 72 trustworthiness checks to RCTs included in 50 Cochrane Reviews. The University of Manchester Ethics tool was used to determine that ethical approval was not required for this study (30th Sept 2022).

Description of trustworthiness checks

Prior to this exercise, a list of trustworthiness checks was assembled using a scoping review (16), qualitative study (17) and survey of experts (15). This list contained 116 checks arranged into five domains: *Inspecting results in the paper, Inspecting the research team and their work, Inspecting conduct, governance and transparency, Inspecting text and publication details,* and *Inspecting individual participant data*. In the current study we only considered the first four domains, as individual participant data are not generally available during systematic reviews and meta-analyses based on aggregate data; nor were they available to us. An extension to the INSPECT-SR tool based on the checks in the fifth domain, which can be applied when individual participant data are available, 'INSPECT-IPD', has been funded for development (Reference: NIHR30355). The first four domains included 76 checks (Tables 1 and 2). We made some modifications to the list in preparation for the current study, in consultation with the project expert advisory panel. This included refining the language of some items to improve clarity. To assist assessors in applying the checks, we drafted brief explanations for each check (S Tables 1 to 4). Four checks (checks 45, 66, 67, 72, Tables 1 and 2) were not assessed as they were not considered practicable in the context of the present study. Consequently, 72 checks were assessed.

Description of assessors

The INSPECT-SR working group includes a core management group and an expert advisory panel. Members of both were invited to act as assessors for the current study. We also invited additional collaborators who had expressed an interest in contributing to the development process. Collaborators were identified from a variety of sources. We invited attendees at presentations relating to the project to contact us to express an interest, and also invited individuals who had expressed an interest in the topic to JW using personalised emails and via social media. All assessors were considered to have sufficient expertise in research methods (specifically, to evaluate RCTs) to enable them to undertake the assessment. We did not require assessors to hold any particular qualification however. We did not require assessors to have specialist expertise relating to research integrity (for example, use of forensic statistical methods or investigation of misconduct cases), as a key objective was to learn about the usefulness and feasibility of the checks when applied by potential users of the INSPECT-SR tool (i.e. systematic reviewers, researchers, peer reviewers) who would not be expected to possess this specialist knowledge. Assessors who were considered to have made a substantial contribution to data acquisition and critical review of manuscript drafts, were given the option to co-author the manuscript.

Selection of Cochrane Reviews and RCTs

The sample size of 50 Cochrane reviews represented a number that was considered feasible to complete, while facilitating the evaluation of feasibility and impact of applying the checks across different topic areas. A preliminary pilot was conducted on a small number of RCTs to confirm this. The 50 reviews were purposefully selected from the Cochrane Library. To be eligible, a review could not be authored or co-authored by the assessor, and could not contain RCTs authored or co-authored by the assessor, to prevent any conflict of interest in conducting the assessment. As a feasibility requirement, we also required that the review contained at least one (meta-) analysis containing one to five RCTs. For brevity, we use the term 'meta-analysis' in this article to describe an analysis which produces a pooled average estimate and confidence interval for a treatment effect on an outcome based on the included studies, recognising that, when there is only one study, this involves reporting the estimate and confidence interval from that study. The RCTs in the first eligible meta-analysis in the review were subjected to the trustworthiness assessment, as a feasibility constraint. We also required that the review had not already undergone a trustworthiness assessment as part of the review process, since this could have resulted in the prior removal of problematic studies, distorting our assessment. Assessors were asked to suggest a topic with which they were broadly familiar. We attempted to match assessors to review topics, to replicate a typical scenario in which INSPECT-SR would be used (a systematic review would often be undertaken by someone with some relevant subject-matter knowledge). We then selected the most recent Cochrane Review relating to the topic suggested by the assessor that met the eligibility criteria. Assessors did not always have subject-matter knowledge relating to the review(s) they assessed, however. For example, some assessors were primarily methodologists, with limited clinical knowledge of the subject matter. For these people, we attempted to select

review topics to cover a broad range of health areas. We asked each assessor to record their familiarity with the review topic during data extraction (little or no familiarity, some familiarity, or high familiarity).

Data extraction and trustworthiness assessment

A bespoke data extraction form was produced, and was revised following piloting on a small number of RCTs, and can be accessed at https://osf.io/9pyw2/ . Assessors were informed of software that could be used to implement some of the statistical checks. Examples include the scrutiny package in R (18), online applications created to implement some checks e.g. applications for performing GRIMMER (Granularity-Related Inconsistency of Means Mapped to Error Repeats) and SPRITE (Sample Parameter Reconstruction via Iterative Techniques)(19, 20), or Microsoft Excel (21) for basic statistical checks, but it was not a requirement to use any particular software to undertake the assessment. For each Cochrane Review, the assessor extracted data and applied the list of checks to each RCT in the meta-analysis. An exception was check 26 - *Is there heterogeneity across studies in degree of imbalance in baseline characteristics (in meta-analysis)*, which was assessed only once per review. The assessor extracted the year of publication for each RCT, the summary data entered in the meta-analysis, and Risk of Bias and GRADE (Grading of Recommendations Assessment, Development and Evaluation) (22) assessments as presented in the Cochrane Review.

A second assessor performed a quality check of accuracy and completeness of this information following extraction. Any disagreements were resolved by discussion between assessors and a third team member (JW). The assessor attempted to apply each of the 72 checks to the trial, selecting one of four response options: not feasible; passed the check; possible fail; fail. For each check, assessors were asked to supply free text to explain their assessment. The country or countries in which the RCT was conducted was also recorded. After applying the checks, assessors recorded their answer to the question ''Do you have concerns about the authenticity of this study?" using one of four response options: no; some concerns; serious concerns; don't know. Assessors were asked if they had performed any additional checks (not included on the list) and if so, to describe both the checks and the results of applying them. There was space for the assessor to add any additional information, and to provide an estimate of how many hours it took them to assess the RCT. The intention had been for one assessor to assess all of the RCTs in the review, before checking by a second assessor. However, some assessors failed to complete the assessment of all RCTs in their allocated review, and so for several reviews the RCTs were split between two assessors, before being checked by a third assessor.

Statistical analysis

We summarised trial and Cochrane Review characteristics, and the responses for each check. We calculated how often assessors had concerns about study authenticity. We evaluated the impact of applying each check by comparing the analysis/ meta-analysis including all trials as per the review to a version in which any RCTs flagged by the check were removed, in terms of the numbers of trials, sample size, change in effect estimate, 95% confidence interval width, heterogeneity, and change in inference.

The first two of these metrics were assessed over all reviews, while the remainder were assessed separately for binary and continuous outcomes. We used the metafor package (23) in R to perform all meta-analyses, using odds ratios to summarise treatment effects with binary outcomes, and standardised mean differences to summarise treatment effects with continuous outcomes. Random effects meta-analyses using the DerSimonian and Laird (24) method were performed, as the most typical method employed in systematic reviews (25, 26).

We assessed potential redundancies among the checks by plotting the responses for each check for each RCT in an array. We made the post-hoc decision to undertake a hierarchical cluster analysis, using complete agglomeration based on Gower dissimilarity, as implemented in the cluster package in R (27). The purpose of this analysis was to identify possible clusters of checks that could potentially be combined. We used multinomial regression to assess the relationship between trustworthiness assessment and each Risk of Bias domain, and ordinal regression (proportional odds logistic regression) to consider the relationship between the GRADE assessment and the number of trials flagged for concerns. We used likelihood ratio tests for inference following regression model fits. We conducted an additional analysis which had not been specified in the protocol, where we evaluated the relationship between the assessment for each check and the overall assessment of the trial using the N-1 chi-squared test (28), to determine which checks were influential in reaching an overall assessment. The N-1 chi-squared test was used in anticipation of small expected counts (29). This analysis was performed in trials where the check was considered to be feasible, and the assessments were analysed as ‘passed’ vs ‘fail or possible fail’. We used a post-hoc significance threshold of 1% to highlight checks associated with the overall assessment, creating contingency tables (outcome of check vs overall assessment) for these checks to determine whether failing the check was associated with an assessor having overall concerns. We categorised the free-text responses to the question asking how long it took to complete the assessment in a post-hoc fashion (less than 90 minutes, 90 minutes to 3 hours, more than 3 hours). The dataset and analysis code for this study are available at https://osf.io/9pyw2/ .

Results

We included a total of 95 RCTs from 50 Cochrane Reviews. The reviews were from 24 different Cochrane Groups (S Table 5). Assessors considered themselves to have high familiarity with the review topic for 7/50 (14%) reviews, some familiarity for 20/50 (40%) of reviews, and little or no familiarity for 23/50 (46%). The characteristics of included Cochrane Reviews are shown

in Table 3. The median (IQR) number of RCTs per review was 1 (1 to 1.9). 27 (54%) contained only 1 RCT. The median (IQR) number of participants in the assessed RCTs was 71 (40 to 174). Fifteen of 95 (16%) were conducted in multiple countries, with the remaining 80 taking place in one of 21 different countries (S Table 6). Twenty-four (26%) RCTs took less than 90 minutes to assess, 29 (31%) took between 90 minutes and 3 hours, and 40 (42%) took more than 3 hours.

Responses to individual trustworthiness checks

Figure 1 and S Table 7 summarise the responses for each check, and S Figure 1 shows the study-level responses for each check. S Figure 2 shows how the checks are clustered in the dataset. Missing data for trustworthiness checks were infrequent, with only one check having missing data for as many as five RCTs (check 42). Check 26 is 'missing' for 10 RCTs, as it was only assessed once per review. A number of checks were considered to have 'failed' or 'possibly failed' often. The five checks most often receiving an assessment of 'failed' or 'possibly failed' were check 61 - *Are the data publically available?* (81%), check 30 - *Are contributorship statements present?* (69%), check 31 - *Are contributorship statements complete?* (57%), check 64 - *Has the study been prospectively registered?* (56%), check 49 – *Is a funding source reported?* (40%). Some statistical checks frequently resulted in responses of 'failed' or 'possibly failed'. Examples include check 12 - *Are differences in variances in baseline variables between randomised groups plausible?* (28%), check 11 – *Are statistical test [results] of outcomes correct?* (21%).

A number of checks were considered to be infeasible in most cases. The checks most frequently considered infeasible were check 40 - *Is the standard deviation of summary statistics in multiple studies by same authors plausible (when compared to simulated or bootstrapped data?)* (99%), check 62 - *Are additional patient data recorded in patient case records beyond what is reported in the paper? (*98%), check 38 - *Is the distribution of non-first digits in manuscripts from one author compatible with a genuine measurement process?* (90%), check 35 - *Is any duplicate reporting acknowledged or explained?* (89%), and check 29 - *Are withdrawal and loss to follow-up in multiple trials by the same author consistent with the expected (random) binomial distribution?* (84%).

Overall assessment and relationship to individual checks

Overall, responses to the question "Do you have concerns about the authenticity of this study?" were: no (60/95, 64%); some concerns (24/95, 25%); serious concerns (6/95, 6%). P-values from chi-squared tests looking at the outcome of each check against the overall trustworthiness assessment of the study are shown in S Table 8. Noting that these analyses were post-hoc and exploratory, 19 checks were associated with overall assessment using a 1% significance level. Contingency tables were inspected to examine the nature of these

associations (specifically to confirm that failing or possibly failing, rather than passing, a check was associated with the presence of concerns).

Of these 19, there were 11 checks for which failing (as opposed to passing) the check appeared to correlate with an assessment of overall concern: 1. *Are any baseline data implausible with respect to magnitude, frequency, or variance?* (p=0.00001); 2. *Is the number of participant withdrawals compatible with the disease, age and timeline?* (p=0.005); 8. *Are there any discrepancies between data reported in figures, tables and text?* (p = 0.00006), 9. *Are any outcome data, including estimated treatment effects, implausible?* (p = 0.000002), 19. *Are results internally consistent?* (p=0.00008), 37. *Does the statistics methods section use generic language, suggesting lack of expert statistical input?* (p=0.003), 51. *Is the reported staffing adequate for the study conduct as reported?* (p=0.009), 52. *Is the recruitment of participants plausible within the stated time frame for the research?* (p=0.0005), 53. *Is the recruitment of participants plausible considering the epidemiology of the disease in the area of the study location?* (p=0.0004), 56. *Are there any concerns about unethical practice?* (p=0.001), 64. *Has the study been prospectively registered?* (p=0.004).

Impact of applying the trustworthiness assessments on systematic review results

S Table 9 and S Table 10 show the impact of removing RCTs flagged by each check (considered individually) from meta-analysis, for binary and continuous outcomes respectively. In continuous outcome meta-analyses, removal of RCTs flagged by a check resulted in a median of 4% (IQR 0% to 12.5%, range 0% to 67%) of meta-analyses having no remaining trials. In binary outcome meta-analysis, the corresponding values were 4% (IQR 0% to 8%, range 0% to 73%). The sample size of reviews would be reduced to a median (of means) of 93% (IQR 87% to 97%, range 27% to 100%) of the original size. The median (of means) number of trials that would be removed from meta-analysis was 0.14 (IQR 0.06 to 0.24, range 0 to 1.52).

When RCTs were removed on the basis of the overall assessment (some or serious concerns), 33% of continuous outcome meta-analyses and 12% of binary outcome meta-analyses had no remaining trials. Amongst meta-analyses with at least one RCT remaining, for binary outcome meta-analyses, the mean ROR was 0.98; SE increased by 19% on average; none changed in terms of statistical significance (using a 5% significance threshold); and the mean ratio of confidence interval widths (width expressed as the ratio of upper to lower limit on OR scale) was 4.52. For continuous outcome meta-analyses with at least one RCT remaining, the average change in estimate was -0.02 SDs; SE (and, equivalently, CI width) increased by a mean of 5%; and none of the meta-analyses changed in terms of statistical significance.

Relationship between trustworthiness assessments, Risk of Bias and GRADE

We only investigated the relationship between overall trustworthiness assessment and risk of bias for reviews using the first version of the Cochrane RoB tool since there were only 10 reviews applying RoB 2. Multinomial regression did not indicate associations between any risk of bias domain and overall concern, with the exception of allocation concealment. However,

this was not in the expected direction, with concerns expressed more often for studies with unclear or low bias assessment compared to high bias assessment (p=0.01). The estimated relationship between number of trials flagged for concerns and GRADE assessment was imprecise (OR = 0.68, 95% CI = (0.39 to 1.17)).

New checks used by assessors

Assessors described eight checks which they used and which they felt were additional to the list of checks assessed in the study. Two of these – checking for trial registration, and checking the author list – were already covered by the primary list. Three others were variations of existing checks – checking the certification status of the ethical committee or institutional review board, looking at a related publication of a subgroup, and checking for consistency with the main article. Three were new: looking to see whether the authors exclusively worked together, checking whether the first author's department had participated in other RCTs, and looking into the reported funder.

Discussion

An extensive list of trustworthiness checks was assessed for their feasibility and impact by application to 95 RCTs in a sample of 50 Cochrane Reviews. The study allowed us to estimate how often each of the checks would be considered infeasible for routine use in systematic reviews, how often each would fail, and what the impact of applying the check would be on the estimates from meta-analysis. We found that, in the context of conducting a systematic review, the checks can be applied to identify problematic studies. Furthermore, the findings suggest that a substantial portion of meta-analyses would be left with no remaining RCTs if failed checks were used to identify and exclude problematic studies. Amongst those with remaining RCTs, there was a larger impact on precision than on the magnitude of effect estimates. The study also found that, following application of the checks, assessors frequently had concerns about the RCTs included in Cochrane Reviews, with "some concerns" being reported for 25% of studies, and '"serious concerns'" for a further 6%.

Feasibility of the checks

A number of checks were deemed generally infeasible. For example, assessments which involved taking an author-wide view have been successfully implemented in particular cases (e.g. (8, 30, 31)), but were not considered feasible by assessors in the context of the current study. One possible reason is that these checks require additional data collection to find out more about the authors of a study, their research team, or their other publications. In a previous survey of experts, the need for a trustworthiness tool to be practical and not too burdensome was emphasised (15), and therefore checks which require the identification and comparison of additional studies are unlikely to be palatable. Other checks that were deemed infeasible include checking for evidence of copied work, including copied sample characteristics and results tables. Unless the copying is identified between RCTs that both happen to be included in the review, it is difficult to see how this sort of check would be

practicable in the absence of automated solutions. Assessing the plausibility of various RCT features is likely to be difficult without recourse to domain expertise. Clearly, the results indicate that it would not be feasible to apply such a long list of checks routinely, as this took more than 3 hours for 42% of the trials.

Identification of problematic studies

Failed checks are potential indicators of a problematic trial. Some checks failed for most RCTs. For example, the study agreed with previous work suggesting that many (in the present study, most) RCTs are not prospectively registered (32-34), and few make the underlying data available (35). In relation to registration, assessors were much more likely to have concerns about authenticity for studies that were not prospectively registered compared to those that were. This could indicate that lack of registration was influential in reaching an overall judgement, or rather that studies with other problematic features were less likely to be prospectively registered. Prospective registration is routinely considered in relation to reporting bias, and an important question to be resolved in the INSPECT-SR development process is whether there is additional value in considering prospective registration in the assessment of trustworthiness.

Our findings also indicate that some checks may be prone to misinterpretation or misapplication, which was suggested by high failure rates. In particular, several statistical checks proved challenging. For example, 20% of RCTs were considered to have 'failed' or 'possibly failed' a check looking to see whether results of statistical tests of outcomes were correct. Some of these failures might be attributable to the rounding of continuous variables in published articles; p-values obtained from rounded summary statistics can differ from those obtained from analysis of the underlying data, meaning the question of assessing consistency cannot just be assessed reproducing the test and looking for an exact match (36). Another example was checking differences in baseline variance between groups, which '"failed'" or '"possibly failed'" for 28% of RCTs. Assessors were directed to use an F test here. However, this test has an inflated type 1 error rate for skewed variables (37), such that rejection of the test assumptions may have been frequently mistaken for rejection of the hypothesis of equality of variances. Instances such as these may have led to unwarranted skepticism about a study's authenticity in some instances. Although we did not detect associations between failure of these checks and concerns in post-hoc analyses, it remains possible that errors of this nature did sometimes occur, and may have influenced the overall assessment of a trial's authenticity.

Impact of identifying problematic studies on systematic reviews

The impact of removing RCTs flagged by these checks from meta-analyses may appear alarming; for example, removing RCTs for which assessors expressed concerns would result in 11 of 50 meta-analyses with no RCTs remaining. However, caution is needed here due to limitations introduced by our study design. We only assessed meta-analyses containing five

or fewer trials in this study. Consequently, more than half contained only a single RCT, although small numbers appear to be typical (38). This will exaggerate the number of reviews with no remaining studies following trustworthiness assessment. Moreover, several other metrics, such as the change in point estimate and associated uncertainty, could only be evaluated in meta-analyses with at least one trial remaining following assessment. Due to the fact that many meta-analyses only included one trial initially, this subset will omit many of the meta-analyses with any trials flagged at all, causing the impact of the checks on these metrics to be understated.

In line with our expectations, there did not appear to be a clear association between Risk of Bias domains and overall trustworthiness assessment, reinforcing the premise that these frameworks are evaluating different aspects of trials. Many problematic studies appear to frequently describe perfectly sound methods (8). We were unable to ascertain whether there is any link between GRADE and trustworthiness assessment, as our estimate of the relationship was too imprecise. We suggest that trustworthiness assessment should be performed prior to Risk of Bias and GRADE assessments, because the value of assessing the internal or external validity of a problematic study is doubtful.

Implications for development of INSPECT-SR and future directions

These observations have informed the development of the INSPECT-SR tool and accompanying guidance. The findings highlight the need for careful curation of the checks included in INSPECT-SR, and suggest that any statistical checks included in the tool would have to be accompanied by detailed guidance to enable their application, as well as to prevent misuse and misinterpretation. As technological solutions become available to facilitate some useful but difficult checks, they can become part of the tool implementation. As the role of automation, including artificial intelligence, is likely to expand in evidence synthesis, it will be important to examine how it might enable or hinder detection of problematic RCTs (39). For example, some checks, such as statistical checks, may be more amenable to automation than checks that require more content knowledge, such as the plausibility of participant recruitment or effect sizes.

Additional future directions informed by this study will be development of training for INSPECT-SR and tools that can be applied to individual patient data or observational study designs. Creating a searchable, open archive of trials that have been evaluated with INSPECT-SR will aid all systematic reviewers and users of trials. Lastly, although INSPECT-SR is being developed for use by systematic reviewers, adaptations of the tool could also be useful to journal editors or publishers who screen trials for research integrity problems.

Conclusion

The study appears to reinforce the need for routine trustworthiness assessment in RCTs, suggesting that problematic studies in systematic reviews may not be infrequent, and are not detected by Risk of Bias assessment. Only two of the studies judged to be concerning had associated retraction or expression of concern notices at the point of assessment, highlighting the need to evaluate other features in order to identify these untrustworthy trials. The time taken to complete the full barrage of checks for each RCT was long, and would likely not be practicable in the context of a typical systematic review. The goal of subsequent stages of the INSPECT-SR project will be to identify a subset of these checks that are both feasible and useful, and to implement these in the form of a tool that can be implemented by systematic reviewers. The results from this study will be used to select checks for this purpose, alongside a Delphi study of experts and potential users of the tool. Both sets of results will be presented to experts at a series of consensus meetings, which will be used to determine the content of a draft version of INSPECT-SR. The draft version of the tool will then be tested in the assessment of RCTs, and feedback will be used to finalise the tool in early 2025.

Declarations

JW, CH, GAA, LB, JJK declare funding from NIHR (NIHR203568) in relation to the current project. JW additionally declares Stats or Methodological Editor roles for BJOG, Fertility and Sterility, Reproduction and Fertility, Journal of Hypertension, and for Cochrane Gynaecology and Fertility. CH declares a Statistical Editor role for Cochrane Colorectal. GAA additionally declares a Statistical Reviewer role for the European Journal of Vascular and Endovascular Surgery. LB additionally declares a role as Academic Meta-Research Editor for PLoS Biology, and that The University of Colorado receives remuneration for service as Senior Research Integrity Editor, Cochrane. JJK additionally declares a Statistical Editor role for The BMJ. EF is employed by the Cochrane Collaboration and on the Editorial Board of Cochrane Evidence Synthesis and Methods. ES is a Senior (Sign-off) Editor for Cochrane and contributed to Cochrane's Policy for managing potentially problematic studies. SL is an editor for Cochrane Gynaecology and Fertility, Human Reproduction, and Fertility and Sterility. TJL is the Deputy Editor in Chief of The Cochrane Library and is an employee of The Cochrane Collaboration. DNB is an associate editor for Research Quarterly for Exercise and Sport and a section editor for Communications in Kinesiology. NEO is a member of the Cochrane Editorial Board and holds an ERA-NET Neuron Co-Fund grant for a separate project. RR declares acting as an author and editor on Cochrane reviews. KS is an editor for Cochrane Gynaecology and Fertility, and Fertility and Sterility. MvW declares to be co-ordinating editor for Cochrane Gynaecology and Fertility and Cochrane Sexually Transmitted Infections, methodological editor for Human Reproduction Update and Editorial Editor for Fertility & Sterility. HT is Deputy Editor of The Lancet Gastroenterology & Hepatology and is an employee of Elsevier. SL received funding from the French National Research Agency (ANR-23-CE36-0006-01). AK is an editorial board member for BJGP Open. TLi serves as the Principal Investigator on a grant from the National Eye Institute, National Institutes of Health that funds the work of Cochrane Eyes and Vision US Project. She also acts as a sign-off editor for The Cochrane Library. ZM is

supported by an NHMRC Investigator Grant 1195676. ZM is an associate Editor for BMC Medical Research Methodology and is on the Editorial Board for Clinical and Public Health Guidelines. RC is Editor-in-Chief at Meta-Psychology. CL is a work-package leader for the doctoral network MSCA-DN SHARE-CTD (HORIZON-MSCA-2022-DN-01 101120360), funded by the EU. CV received funding as part of the OSIRIS project (Open Science to Increase Reproducibility in Science); the OSIRIS (Open Science to Increase Reproducibility in Science) project has received funding from the EU (grant agreement No. 101094725). FN received funding from the French National Research Agency (ANR-23-CE36-0006-01), the French ministry of health and the French ministry of research. He is a work-package leader in the OSIRIS project (Open Science to Increase Reproducibility in Science). The OSIRIS project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101094725. He is a work-package leader for the doctoral network MSCA-DN SHARE-CTD (HORIZON-MSCA-2022-DN-01 101120360), funded by the EU. DN declares having led/co-authored/co-authoring Cochrane Reviews. He also declares having been part of the Cochrane Convenes initiative organised by Cochrane to consider the issue of misinformation, its impact on the health evidence ecosystem and solutions to address it. LJ is the creator of the scrutiny package in R. WL is supported by an NHMRC Investigator grant (GNT2016729). RW is supported by an NHMRC Investigator Grant (2009767) and acts as a Deputy Editor for Human Reproduction, and an editorial board member for BJOG and Cochrane Gynaecology and Fertility. EF, SGTLa and RR declare employment by Cochrane. TLa additionally declares authorship of a chapter in the Cochrane Handbook for Systematic Reviews of Interventions and that he is a developer of standards for Cochrane intervention reviews (MECIR). AL is on the editorial board of BMC Medical Ethics.

Ethical approval

The University of Manchester ethics decision tool was used on 30/09/22. Ethical approval was not required for this study, since it involved appraisal of published research.

Funding

This study/project is funded by the NIHR Research for Patient Benefit programme (NIHR203568). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

References

1. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.
2. Sterne JAC, Savovic J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898.
3. Carlisle JB. False individual patient data and zombie randomised controlled trials submitted to Anaesthesia. Anaesthesia. 2021;76(4):472-9.
4. Weeks J, Cuthbert A, Alfirevic Z. Trustworthiness assessment as an inclusion criterion for systematic reviews—What is the impact on results? Cochrane Evidence Synthesis and Methods. 2023;1(10):e12037.

5. Hill A, Mirchandani M, Pilkington V. Ivermectin for COVID-19: Addressing Potential Bias and Medical Fraud. Open Forum Infect Dis. 2022;9(2):ofab645.
6. Cochrane. Cochrane Policy for managing potentially problematic studies. Cochrane Database of Systematic Reviews: editorial policies Cochrane Library [Available from: https://www.cochranelibrary.com/cdsr/editorial-policies.
7. Boughton SL, Wilkinson J, Bero L. When beauty is but skin deep: dealing with problematic studies in systematic reviews. Cochrane Database Syst Rev. 2021;6(6):ED000152.
8. O'Connell NE, Moore RA, Stewart G, Fisher E, Hearn L, Eccleston C, et al. Investigating the veracity of a sample of divergent published trial data in spinal pain. Pain. 2023;164(1):72-83.
9. Wilkinson J, Heal C, Antoniou GA, Flemyng E, Alfirevic Z, Avenell A, et al. Protocol for the development of a tool (INSPECT-SR) to identify problematic randomised controlled trials in systematic reviews of health interventions. BMJ Open. 2024;14(3):e084164.
10. Mol BW, Lai S, Rahim A, Bordewijk EM, Wang R, van Eekelen R, et al. Checklist to assess Trustworthiness in RAndomised Controlled Trials (TRACT checklist): concept proposal and pilot. Res Integr Peer Rev. 2023;8(1):6.
11. Weibel S, Popp M, Reis S, Skoetz N, Garner P, Sydenham E. Identifying and managing problematic trials: A research integrity assessment tool for randomized controlled trials in evidence synthesis. Res Synth Methods. 2023;14(3):357-69.
12. Hunter KE, Aberoumand M, Libesman S, Sotiropoulos JX, Williams JG, Aagerup J, et al. The Individual Participant Data Integrity Tool for assessing the integrity of randomised trials. Res Synth Methods. 2024.
13. Abbott J, Acharya G, Aviram A, Barnhart K, Berghella V, Bradley CS, et al. Trustworthiness criteria for meta-analyses of randomized controlled studies: OBGYN journal guidelines. Elsevier; 2024. p. 101481.
14. Grey A, Bolland MJ, Avenell A, Klein AA, Gunsalus CK. Check for publication integrity before misconduct. Nature. 2020;577(7789):167-9.
15. Wilkinson J, Heal C, Antoniou GA, Flemyng E, Avenell A, Barbour V, et al. A survey of experts to identify methods to detect problematic studies: stage 1 of the INveStigating ProblEmatic Clinical Trials in Systematic Reviews project. J Clin Epidemiol. 2024;175:111512.
16. Bordewijk EM, Li W, van Eekelen R, Wang R, Showell M, Mol BW, et al. Methods to assess research misconduct in health-related research: A scoping review. J Clin Epidemiol. 2021;136:189-202.
17. Parker L, Boughton S, Lawrence R, Bero L. Experts identified warning signs of fraudulent research: a qualitative study to inform a screening tool. J Clin Epidemiol. 2022;151:1-17.
18. Jung L. scrutiny: Error Detection in Science. R package version 0.3.0. 2023 [Available from: https://CRAN.R-project.org/package=scrutiny.
19. Anaya J. The GRIMMER test: A method for testing the validity of reported measures of variability. PeerJ Preprints. 2016;4:e2400v1.
20. Heathers JA, Anaya J, van der Zee T, Brown NJ. Recovering data from summary statistics: Sample parameter reconstruction via iterative techniques (SPRITE). PeerJ Preprints; 2018. Report No.: 2167-9843.
21. Microsoft Corporation. Microsoft Excel. 2018.
22. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-6.
23. Viechtbauer W. Conducting meta-analyses in R with the metafor package. 2010.
24. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177-88.
25. DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. Contemp Clin Trials. 2015;45(Pt A):139-45.
26. Mheissen S, Khan H, Normando D, Vaiid N, Flores-Mir C. Do statistical heterogeneity methods impact the results of meta- analyses? A meta epidemiological study. PLoS One. 2024;19(3):e0298526.

27. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0. [Available from: https://cran.r-project.org/web/packages/cluster/index.html]. 2019.
28. Pearson ES. The choice of statistical tests illustrated on the interpretation of data classed in a 2× 2 table. Biometrika. 1947;34(1/2):139-67.
29. Campbell I. Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. Stat Med. 2007;26(19):3661-75.
30. Simonsohn U. Just post it: the lesson from two cases of fabricated data detected by statistics alone. Psychol Sci. 2013;24(10):1875-88.
31. Bolland MJ, Avenell A, Gamble GD, Grey A. Systematic review and statistical analysis of the integrity of 33 randomized controlled trials. Neurology. 2016;87(23):2391-402.
32. Hunter KE, Seidler AL, Askie LM. Prospective registration trends, reasons for retrospective registration and mechanisms to increase prospective registration compliance: descriptive analysis and survey. BMJ Open. 2018;8(3):e019983.
33. Harriman SL, Patel J. When are clinical trials registered? An analysis of prospective versus retrospective registration. Trials. 2016;17:187.
34. Azar M, Riehm KE, Saadat N, Sanchez T, Chiovitti M, Qi L, et al. Evaluation of Journal Registration Policies and Prospective Registration of Randomized Clinical Trials of Nonregulated Health Care Interventions. JAMA Intern Med. 2019;179(5):624-32.
35. Hamilton DG, Hong K, Fraser H, Rowhani-Farid A, Fidler F, Page MJ. Prevalence and predictors of data and code sharing in the medical and health sciences: systematic review with meta-analysis of individual participant data. BMJ. 2023;382:e075767.
36. Brown NJ, Heathers J. Rounded Input Variables, Exact Test Statistics (RIVETS). 2019.
37. BOX GEP. NON-NORMALITY AND TESTS ON VARIANCES. Biometrika. 1953;40(3-4):318-35.
38. Davey J, Turner RM, Clarke MJ, Higgins JP. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. BMC Med Res Methodol. 2011;11:160.
39. Thomas J, Flemyng E, Noel-Storr A, Moy W, Marshall I, Hajji R, et al. Responsible AI in Evidence Synthesis (RAISE): guidance and recommendations (Draft for consultation and revision) Open Science Framework2024 [Available from: https://osf.io/cn7x4.

Figure legends

Figure 1: Responses to trustworthiness checks in four domains

S Figure 1: Study-level responses to trustworthiness checks

S Figure 2: Dendrogram displaying hierarchical clustering of checks using complete agglomeration based on Gower dissimilarity

711

Inspecting results in the paper (28 checks)	Inspecting the research team and their work (19 checks)
1. Are any baseline data implausible with respect to magnitude, frequency, or variance? 2. Is the number of participant withdrawals compatible with the disease, age and timeline? 3. Are subgroup means incompatible with those for the whole cohort? 4. Are the reported summary data compatible with the reported range? 5. Are correct units reported? 6. Are calculations of proportions and percentages correct? 7. Are numbers of participants correct and consistent throughout the publication? 8. Are there any discrepancies between data reported in figures, tables and text? 9. Are any outcome data, including estimated treatment effects, implausible? 10. Are baseline statistical tests correct? 11. Are statistical tests of outcomes correct? 12. Are differences in variances in baseline variables between randomised groups plausible? 13. Are any of the baseline data excessively similar between randomized groups? 14. Are any of the baseline data excessively different between randomised groups? 15. Are the summary outcome data identical or nearly identical across study groups? 16. Are there any discrepancies between the values for percentage and absolute change? 17. Are there any discrepancies between reported data and participant inclusion criteria? 18. Are the variances in biological variables surprisingly consistent over time? 19. Are results internally consistent? 20. Are coefficients of variation unusually similar when calculated across variables reported in the paper? 21. Is the amount of missing data plausible? 22. Are the results substantially divergent from the results of multiple other studies in meta-analysis? 23. Are non-first digits compatible with a genuine measurement process? 24. Are the variances of integer data possible? 25. Are the means of integer data possible? 26. Is there heterogeneity across studies in degree of imbalance in baseline characteristics (in meta-analysis) – only once per review 27. Are integer data simulated from reported summary statistics plausible? 28. Are important features missing from the paper?	29. Check whether withdrawal and loss to follow-up in multiple trials by the same author are consistent with the expected (random) binomial distribution 30. Are contributorship statements present? 31. Are contributorship statements complete? 32. Have the data been published elsewhere by the research team in an illegitimate fashion? 33. Are duplicate-reported data consistent between publications? 34. Are relevant methods consistent between publications? 35. Is any duplicate reporting acknowledged or explained? 36. Is there evidence of duplication of figures? 37. Does the statistics methods section use generic language, suggesting lack of expert statistical input? 38. Is the distribution of non-first digits in manuscripts from one author compatible with a genuine measurement process? 39. Does consideration of other studies from members of the research team highlight causes for concern (including expressions of concern, relevant post-publication amendment, or critical retraction)? 40. Is the standard deviation of summary statistics in multiple studies by same authors plausible (when compared to simulated or bootstrapped data?) 41. Do all authors meet criteria for authorship? 42. Is authorship of related papers consistent? 43. Are the authors on staff of institutions they list? 44. Do any authors have a professorial title but no other publications on PubMed? 45. Can co-authors attest to the reliability of the paper? 46. Given the nature of the study, does the author list make sense? - e.g.does a simple study have dozens of authors from different institutions and with diverse expertise? 47. In which country was the study conducted?

712 Table 1: Trustworthiness checks in the first and second domains of the assessed list

713

714

715

716

717

718

719

Inspecting conduct, governance, and transparency (22 checks)	Inspecting text and publication details (7 checks)
48. Is the grant funding number identical to the number in unrelated studies? 49. Is a funding source reported? 50. Is the volume of work reported by research group plausible, including that indicated by concurrent studies from the same group? 51. Is the reported staffing adequate for the study conduct as reported? 52. Is the recruitment of participants plausible within the stated time frame for the research? 53. Is the recruitment of participants plausible considering the epidemiology of the disease in the area of the study location? 54. Is the interval between study completion and manuscript submission plausible? 55. Is there evidence that the work has been approved by a specific, recognized committee? (ethics) 56. Are there any concerns about unethical practice? 57. Could the study plausibly be completed as described? 58. Are the study methods plausible, at the location specified? 59. Are the locations where the research took place specified, and is this information plausible? 60. Do the authors agree to share individual participant data? 61. Are the data publically available? 62. Are additional patient data recorded in patient case records beyond what is reported in the paper? 63. Does the trial registration number refer to other studies? 64. Has the study been prospectively registered? 65. Are details such as dates and study methods in the publication consistent with those in the registration documents? 66. Do authors cooperate with requests for information? 67. Do authors provide satisfactory responses to requests? 68. Was the time between submission to acceptance reasonable? 69. Is the procedure of the study aligned with local legislations?	70. Are there typographical errors? 71. Has the study been retracted or does it have an expression of concern, a relevant post-publication amendment, a critical Retraction Watch or PubPeer comment or has been previously excluded from a systematic review? 72. Is there evidence of copied work, such as duplicated or partially duplicated tables? 73. Is there evidence of text reuse (cutting and pasting text between papers), including text that is inconsistent with the study? 74. Is there evidence of automatically-generated text? 75. Was the study published in journal from a list of predatory/ low quality journals? 76. Is there evidence of manipulation or duplication of images?

720 Table 2: Trustworthiness checks in the third and fourth domains of the assessed list

721

722

723

724

725

726

Review-level summary (n = 50)

Number of RCTs in assessed meta-analysis	1	27 (54%)
	2	10 (20%)
	3	8 (16%)
	4	2 (4%)
	5	2 (4%)
	6*	1 (2%)
Number of participants in assessed meta-analysis	Median (IQR)	147 (53 to 341)
Outcome Type	Binary	26 (52%)
	Continuous	24 (48%)
Year of publication	2023	27 (54%)
	2022	10 (20%)
	2021	3 (6%)
	2020	5 (10%)
	2019	4 (8%)
	2014	1 (2%)
GRADE assessment	High	3 (6%)
	Moderate	7 (14%)
	Low	25 (50%)
	Very low	15 (30%)

727 Table 3: Characteristics for 50 Cochrane Reviews assessed in the study. Frequency (%) or median (1st quartile to 3rd quartile)

728 *Assessed in error, included in analysis.

729

730

Inspecting results in the paper (28 checks)	Inspecting the research team and their work (19 checks)
1. Are any baseline data implausible with respect to magnitude, frequency, or variance? 2. Is the number of participant withdrawals compatible with the disease, age and timeline? 3. Are subgroup means incompatible with those for the whole cohort? 4. Are the reported summary data compatible with the reported range? 5. Are correct units reported? 6. Are calculations of proportions and percentages correct? 7. Are numbers of participants correct and consistent throughout the publication? 8. Are there any discrepancies between data reported in figures, tables and text? 9. Are any outcome data, including estimated treatment effects, implausible? 10. Are baseline statistical tests correct? 11. Are statistical tests of outcomes correct? 12. Are differences in variances in baseline variables between randomised groups plausible? 13. Are any of the baseline data excessively similar between randomized groups? 14. Are any of the baseline data excessively different between randomised groups? 15. Are the summary outcome data identical or nearly identical across study groups? 16. Are there any discrepancies between the values for percentage and absolute change? 17. Are there any discrepancies between reported data and participant inclusion criteria? 18. Are the variances in biological variables surprisingly consistent over time? 19. Are results internally consistent? 20. Are coefficients of variation unusually similar when calculated across variables reported in the paper? 21. Is the amount of missing data plausible? 22. Are the results substantially divergent from the results of multiple other studies in meta-analysis? 23. Are non-first digits compatible with a genuine measurement process? 24. Are the variances of integer data possible? 25. Are the means of integer data possible? 26. Is there heterogeneity across studies in degree of imbalance in baseline characteristics (in meta-analysis) – only once per review 27. Are integer data simulated from reported summary statistics plausible? 28. Are important features missing from the paper?	29. Check whether withdrawal and loss to follow-up in multiple trials by the same author are consistent with the expected (random) binomial distribution 30. Are contributorship statements present? 31. Are contributorship statements complete? 32. Have the data been published elsewhere by the research team in an illegitimate fashion? 33. Are duplicate-reported data consistent between publications? 34. Are relevant methods consistent between publications? 35. Is any duplicate reporting acknowledged or explained? 36. Is there evidence of duplication of figures? 37. Does the statistics methods section use generic language, suggesting lack of expert statistical input? 38. Is the distribution of non-first digits in manuscripts from one author compatible with a genuine measurement process? 39. Does consideration of other studies from members of the research team highlight causes for concern (including expressions of concern, relevant post-publication amendment, or critical retraction)? 40. Is the standard deviation of summary statistics in multiple studies by same authors plausible (when compared to simulated or bootstrapped data?) 41. Do all authors meet criteria for authorship? 42. Is authorship of related papers consistent? 43. Are the authors on staff of institutions they list? 44. Do any authors have a professorial title but no other publications on PubMed? 45. Can co-authors attest to the reliability of the paper? 46. Given the nature of the study, does the author list make sense? - e.g.does a simple study have dozens of authors from different institutions and with diverse expertise? 47. In which country was the study conducted?

Table 1: Trustworthiness checks in the first and second domains of the assessed list

Inspecting conduct, governance, and transparency (22 checks)	Inspecting text and publication details (7 checks)
48. Is the grant funding number identical to the number in unrelated studies? 49. Is a funding source reported? 50. Is the volume of work reported by research group plausible, including that indicated by concurrent studies from the same group? 51. Is the reported staffing adequate for the study conduct as reported? 52. Is the recruitment of participants plausible within the stated time frame for the research? 53. Is the recruitment of participants plausible considering the epidemiology of the disease in the area of the study location? 54. Is the interval between study completion and manuscript submission plausible? 55. Is there evidence that the work has been approved by a specific, recognized committee? (ethics) 56. Are there any concerns about unethical practice? 57. Could the study plausibly be completed as described? 58. Are the study methods plausible, at the location specified? 59. Are the locations where the research took place specified, and is this information plausible? 60. Do the authors agree to share individual participant data? 61. Are the data publically available? 62. Are additional patient data recorded in patient case records beyond what is reported in the paper? 63. Does the trial registration number refer to other studies? 64. Has the study been prospectively registered? 65. Are details such as dates and study methods in the publication consistent with those in the registration documents? 66. Do authors cooperate with requests for information? 67. Do authors provide satisfactory responses to requests? 68. Was the time between submission to acceptance reasonable? 69. Is the procedure of the study aligned with local legislations?	70. Are there typographical errors? 71. Has the study been retracted or does it have an expression of concern, a relevant post-publication amendment, a critical Retraction Watch or PubPeer comment or has been previously excluded from a systematic review? 72. Is there evidence of copied work, such as duplicated or partially duplicated tables? 73. Is there evidence of text reuse (cutting and pasting text between papers), including text that is inconsistent with the study? 74. Is there evidence of automatically-generated text? 75. Was the study published in journal from a list of predatory/ low quality journals? 76. Is there evidence of manipulation or duplication of images?

Table 2: Trustworthiness checks in the third and fourth domains of the assessed list

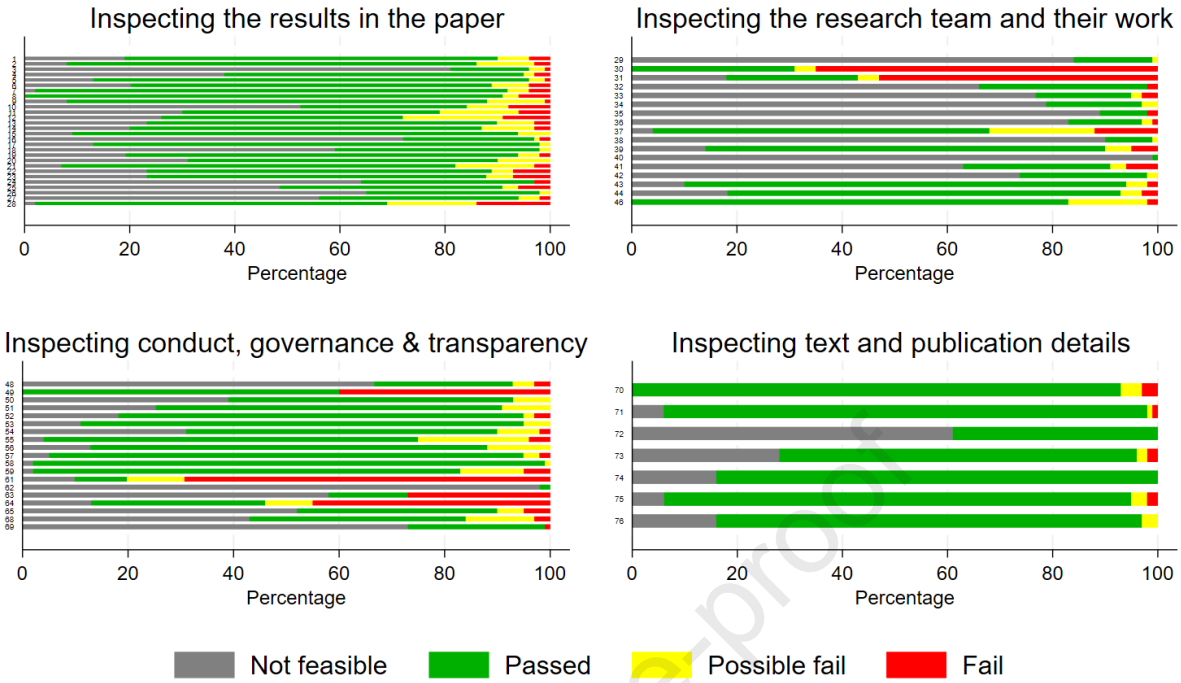
Review-level summary (n = 50)

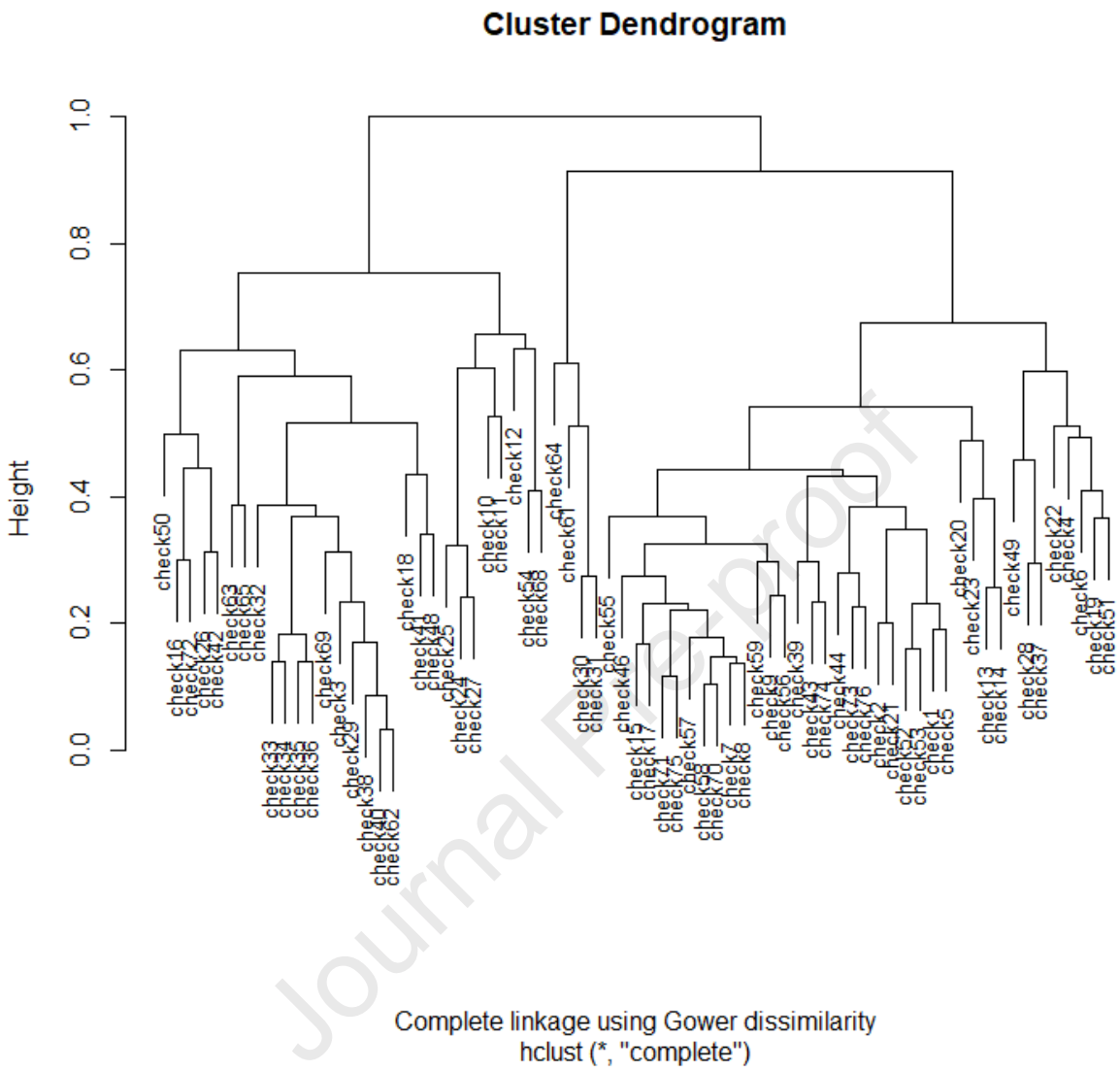
Number of RCTs in assessed meta-analysis	1	27 (54%)
	2	10 (20%)
	3	8 (16%)
	4	2 (4%)
	5	2 (4%)
	6*	1 (2%)
Number of participants in assessed meta-analysis	Median (IQR)	147 (53 to 341)

Outcome Type	Binary	26 (52%)
	Continuous	24 (48%)
Year of publication	2023	27 (54%)
	2022	10 (20%)
	2021	3 (6%)
	2020	5 (10%)
	2019	4 (8%)
	2014	1 (2%)
GRADE assessment	High	3 (6%)
	Moderate	7 (14%)
	Low	25 (50%)
	Very low	15 (30%)

Table 3: Characteristics for 50 Cochrane Reviews assessed in the study. Frequency (%) or median (1st quartile to 3rd quartile)

*Assessed in error, included in analysis.





Declarations

JW, CH, GAA, LB, JJK declare funding from NIHR (NIHR203568) in relation to the current project. JW additionally declares Stats or Methodological Editor roles for BJOG, Fertility and Sterility, Reproduction and Fertility, Journal of Hypertension, and for Cochrane Gynaecology and Fertility. CH declares a Statistical Editor role for Cochrane Colorectal. GAA additionally declares a Statistical Reviewer role for the European Journal of Vascular and Endovascular Surgery. LB additionally declares a role as Academic Meta-Research Editor for PLoS Biology, and that The University of Colorado receives remuneration for service as Senior Research Integrity Editor, Cochrane. JJK additionally declares a Statistical Editor role for The BMJ. EF is employed by the Cochrane Collaboration and on the Editorial Board of Cochrane Evidence Synthesis and Methods. ES is a Senior (Sign-off) Editor for Cochrane and contributed to Cochrane's Policy for managing potentially problematic studies. SL is an editor for Cochrane Gynaecology and Fertility, Human Reproduction, and Fertility and Sterility. TJL is the Deputy Editor in Chief of The Cochrane Library and is an employee of The Cochrane Collaboration. DNB is an associate editor for Research Quarterly for Exercise and Sport and a section editor for Communications in Kinesiology. NEO is a member of the Cochrane Editorial Board and holds an ERA-NET Neuron Co-Fund grant for a separate project. RR declares acting as an author and editor on Cochrane reviews. KS is an editor for Cochrane Gynaecology and Fertility, and Fertility and Sterility.MvW declares to be co-ordinating editor for Cochrane Gynaecology and Fertility and Cochrane Sexually Transmitted Infections, methodological editor for Human Reproduction Update and Editorial Editor for Fertility & Sterility. HT is Deputy Editor of The Lancet Gastroenterology & Hepatology and is an employee of Elsevier. SL received funding from the French National Research Agency (ANR-23-CE36-0006-01). AK is an editorial board member for BJGP Open. TLi serves as the Principal Investigator on a grant from the National Eye Institute, National Institutes of Health that funds the work of Cochrane Eyes and Vision US Project. She also acts as a sign-off editor for The Cochrane Library. ZM is supported by an NHMRC Investigator Grant 1195676. ZM is an associate Editor for BMC Medical Research Methodology and is on the Editorial Board for Clinical and Public Health Guidelines. RC is Editor-in-Chief at Meta-Psychology. CL is a work-package leader for the doctoral network MSCA-DN SHARE-CTD (HORIZON-MSCA-2022-DN-01 101120360), funded by the EU. CV received funding as part of the OSIRIS project (Open Science to Increase Reproducibility in Science); the OSIRIS (Open Science to Increase Reproducibility in Science) project has received funding from the EU (grant agreement No. 101094725). FN received funding from the French National Research Agency (ANR-23-CE36-0006-01), the French ministry of health and the French ministry of research. He is a work-package leader in the OSIRIS project (Open Science to Increase Reproducibility in Science). The OSIRIS project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No. 101094725. He is a work-package leader for the doctoral network MSCA-DN SHARE-CTD (HORIZON-MSCA-2022-DN-01 101120360), funded by the EU. DN declares having led/co-authored/co-authoring Cochrane Reviews. He also declares having been part of the Cochrane Convenes initiative organised by Cochrane to consider the issue of misinformation, its impact on the health evidence ecosystem and solutions to address it. LJ is the creator of the scrutiny package in R. WL is supported by an NHMRC Investigator grant (GNT2016729). RW is supported by an NHMRC Investigator Grant (2009767) and acts as a Deputy Editor for Human Reproduction, and an editorial board member for BJOG and Cochrane Gynaecology and Fertility. EF, SGTLa and RR declare employment by Cochrane. TLa additionally declares authorship of a chapter in the Cochrane Handbook for Systematic Reviews of Interventions and that he is a developer of standards for Cochrane intervention reviews (MECIR). AL is on the editorial board of BMC Medical Ethics.

What is new

- An extensive list of potential checks for assessing study trustworthiness was assessed via an application to 95 randomised controlled trials (RCTs) in 50 Cochrane Reviews.
- Following application of the checks, assessors had concerns about the authenticity of 32% of the RCTs.
- If these RCTs were excluded, 22% of meta-analyses would have no remaining RCTs.
- However, the study showed that some checks were frequently infeasible, and others could be easily misunderstood or misinterpreted.
- The study restricted assessment to meta-analyses including five or fewer RCTs, which might distort the impact of applying the checks.