SDU
University of Southern Denmark

A survey of experts to identify methods to detect problematic studies

stage 1 of the INveStigating ProblEmatic Clinical Trials in Systematic Reviews project

Wilkinson, Jack; Heal, Calvin; Antoniou, George A.; Flemyng, Ella; Avenell, Alison; Barbour, Virginia; Bordewijk, Esmee M.; Brown, Nicholas J.L.; Clarke, Mike; Dumville, Jo; Grohmann, Steph; Gurrin, Lyle C.; Hayden, Jill A.; Hunter, Kylie E.; Lam, Emily; Lasserson, Toby; Li, Tianjing; Lensen, Sarah; Liu, Jianping; Lundh, Andreas; Meyerowitz-Katz, Gideon; Mol, Ben W.; O'Connell, Neil E.; Parker, Lisa; Redman, Barbara; Seidler, Anna Lene; Sheldrick, Kyle; Sydenham, Emma; Dahly, Darren L.; van Wely, Madelon; Bero, Lisa; Kirkham, Jamie J.

Published in:
Journal of Clinical Epidemiology

DOI:
10.1016/j.jclinepi.2024.111512

Publication date:
2024

Document version:
Final published version

Document license:
CC BY

Citation for pulished version (APA):
Wilkinson, J., Heal, C., Antoniou, G. A., Flemyng, E., Avenell, A., Barbour, V., Bordewijk, E. M., Brown, N. J. L., Clarke, M., Dumville, J., Grohmann, S., Gurrin, L. C., Hayden, J. A., Hunter, K. E., Lam, E., Lasserson, T., Li, T., Lensen, S., Liu, J., ... Kirkham, J. J. (2024). A survey of experts to identify methods to detect problematic studies: stage 1 of the INveStigating ProblEmatic Clinical Trials in Systematic Reviews project. *Journal of Clinical Epidemiology*, *175*, Article 111512. https://doi.org/10.1016/j.jclinepi.2024.111512

Go to publication entry in University of Southern Denmark's Research Portal

Terms of use
This work is brought to you by the University of Southern Denmark.
Unless otherwise specified it has been shared according to the terms for self-archiving.
If no other license is stated, these terms apply:

- You may download this work for personal use only.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying this open access version

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim.
Please direct all enquiries to puresupport@bib.sdu.dk

Download date: 10. Jan. 2025

Journal of Clinical Epidemiology 175 (2024) 111512

Journal of Clinical Epidemiology

RESEARCH INTEGRITY SERIES

A survey of experts to identify methods to detect problematic studies: stage 1 of the INveStigating ProblEmatic Clinical Trials in Systematic Reviews project

Jack Wilkinson[a,*], Calvin Heal[a], George A. Antoniou[b,c], Ella Flemyng[d], Alison Avenell[e], Virginia Barbour[f], Esmee M. Bordewijk[g], Nicholas J.L. Brown[h], Mike Clarke[i], Jo Dumville[j,k], Steph Grohmann[d], Lyle C. Gurrin[l], Jill A. Hayden[m], Kylie E. Hunter[n], Emily Lam[o], Toby Lasserson[d], Tianjing Li[p], Sarah Lensen[q], Jianping Liu[r], Andreas Lundh[s,t], Gideon Meyerowitz-Katz[u], Ben W. Mol[v], Neil E. O'Connell[w], Lisa Parker[x], Barbara Redman[y], Anna Lene Seidler[n], Kyle Sheldrick[z], Emma Sydenham[aa], Darren L. Dahly[ab], Madelon van Wely[g], Lisa Bero[ac,1], Jamie J. Kirkham[a,1]

[a]*Centre for Biostatistics, The University of Manchester, Manchester Academic Health Science Centre, Manchester, UK*
[b]*Manchester Vascular Centre, Manchester University NHS Foundation Trust, Manchester, UK*
[c]*Division of Cardiovascular Sciences, School of Medical Sciences, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK*
[d]*Evidence Production and Methods Directorate, Cochrane Central Executive, London, UK*
[e]*Health Services Research Unit, University of Aberdeen, Aberdeen, UK*
[f]*Medical Journal of Australia, Sydney, Australia*
[g]*Department of Obstetrics and Gynaecology, Centre for Reproductive Medicine, Amsterdam University Medical Center, Amsterdam, Netherlands*
[h]*Department of Psychology, Linnaeus University, Växjö, Sweden*
[i]*Northern Ireland Methodology Hub, Queen's University Belfast, Belfast, UK*
[j]*Division of Nursing, Midwifery & Social Work, School of Health Sciences, The University of Manchester, Manchester, UK*
[k]*NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK*
[l]*School of Population and Global Health, The University of Melbourne, Melbourne, Australia*
[m]*Department of Community Health & Epidemiology, Dalhousie University, Halifax, Canada*
[n]*NHMRC Clinical Trials Centre, University of Sydney, Sydney, Australia*
[o]*Independent Lay Member, Unaffiliated, Cheshire, UK*
[p]*Department of Ophthalmology, University of Colorado Anschutz Medical Campus, Aurora, CO, USA*
[q]*Department of Obstetrics, Gynaecology and Newborth Health, Royal Women's Hospital, University of Melbourne, Melbourne, Australia*
[r]*Director, Centre for Evidence-Based Chinese Medicine, Beijing University of Chinese Medicine, Beijing, China*
[s]*Cochrane Denmark & Centre for Evidence-Based Medicine Odense, Department of Clinical Research, University of Southern Denmark, Odense, Denmark*
[t]*Department of Respiratory Medicine and Infectious Diseases, Copenhagen University Hospital - Bispebjerg and Frederiksberg, Copenhagen, Denmark*
[u]*School of Health and Society, University of Wollongong, Wollongong, Australia*
[v]*Department of Obstetrics and Gynaecology, Monash University, Melbourne, Australia*
[w]*Department of Health Sciences, Centre for Wellbeing Across the Lifecourse, Brunel University London, London, UK*
[x]*Charles Perkins Centre, Sydney Medical School, University of Sydney, Sydney, Australia*
[y]*New York University, New York, NY, USA*
[z]*Faculty of Medicine, University of New South Wales, Sydney, Australia*
[aa]*Cochrane Central Editorial Service, London, UK*
[ab]*HRB Clinical Research Facility, University College Cork, Cork, Ireland*
[ac]*University of Colorado Anschutz Medical Campus, Aurora, CO, USA*

Accepted 27 August 2024; Published online 31 August 2024

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.

[1] Joint senior authorship.

* Corresponding author. Centre for Biostatistics, University of Manchester, Room 1.307, Jean McFarlane Building, Oxford Road, Manchester, M13 9PL, UK.

E-mail address: jack.wilkinson@manchester.ac.uk (J. Wilkinson).

https://doi.org/10.1016/j.jclinepi.2024.111512

0895-4356/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Abstract

Background and Objective: Randomized controlled trials (RCTs) inform health-care decisions. Unfortunately, some published RCTs contain false data, and some appear to have been entirely fabricated. Systematic reviews are performed to identify and synthesize all RCTs which have been conducted on a given topic. This means that any of these 'problematic studies' are likely to be included, but there are no agreed methods for identifying them. The INveStigating ProblEmatic Clinical Trials in Systematic Reviews (INSPECT-SR) project is developing a tool to identify problematic RCTs in systematic reviews of health care-related interventions. The tool will guide the user through a series of 'checks' to determine a study's authenticity. The first objective in the development process is to assemble a comprehensive list of checks to consider for inclusion.

Methods: We assembled an initial list of checks for assessing the authenticity of research studies, with no restriction to RCTs, and categorized these into five domains: Inspecting results in the paper; Inspecting the research team; Inspecting conduct, governance, and transparency; Inspecting text and publication details; Inspecting the individual participant data. We implemented this list as an online survey, and invited people with expertise and experience of assessing potentially problematic studies to participate through professional networks and online forums. Participants were invited to provide feedback on the checks on the list, and were asked to describe any additional checks they knew of, which were not featured in the list.

Results: Extensive feedback on an initial list of 102 checks was provided by 71 participants based in 16 countries across five continents. Fourteen new checks were proposed across the five domains, and suggestions were made to reword checks on the initial list. An updated list of checks was constructed, comprising 116 checks. Many participants expressed a lack of familiarity with statistical checks, and emphasized the importance of feasibility of the tool.

Conclusion: A comprehensive list of trustworthiness checks has been produced. The checks will be evaluated to determine which should be included in the INSPECT-SR tool.

Plain Language Summary: Systematic reviews draw upon evidence from randomized controlled trials (RCTs) to find out whether treatments are safe and effective. The conclusions from systematic reviews are often very influential, and inform both health-care policy and individual treatment decisions. However, it is now clear that the results of many published RCTs are not genuine. In some cases, the entire study may have been fabricated. It is not usual for the veracity of RCTs to be questioned during the process of compiling a systematic review. As a consequence, these ''problematic studies'' go unnoticed, and are allowed to contribute to the conclusions of influential systematic reviews, thereby influencing patient care. This prompts the question of how these problematic studies could be identified. In this study, we created an extensive list of checks that could be performed to try to identify these studies. We started by assembling a list of checks identified in previous research, and conducting a survey of experts to ask whether they were aware of any additional methods, and to give feedback on the list. As a result, a list of 116 potential ''trustworthiness checks'' was created. In subsequent research, we will evaluate these checks to see which should be included in a tool, INveStigating ProblEmatic Clinical Trials in Systematic Reviews, which can be used to detect problematic studies.

© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Keywords: Research integrity; Fraud; Fabrication; Misconduct; Trustworthiness; Randomised controlled trials; Systematic reviews; Forensic analysis; Evidence synthesis; Critical appraisal

1. Background

Randomized controlled trials (RCTs) are performed to investigate whether treatments are safe and effective. Systematic reviews exploring health interventions aim to include all relevant RCTs, appraising and synthesizing this evidence to arrive at an overall conclusion about whether an intervention works 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'' [1,2]. Studies may be problematic because they include some false data or results, or 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 study, such as miscoding of patient conditions (eg, inverting active treatment and placebo conditions), failure in the computerized randomization service, 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 routinely appraised on the basis of their methodological validity during the systematic review process. However, these assessments are predicated on the 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 flagged by critical appraisal tools. At present, there are no agreed methods for identifying problematic RCTs, and it is typical for no assessment of authenticity to be undertaken at all. This means that there are no processes for preventing problematic RCTs from being included in systematic reviews, distorting the clinical evidence base, and potentially leading to harm.

What is new?

Key findings

- An extensive list of potential checks for assessing study trustworthiness was assembled using a survey of experts.

What this adds to what was known?

- Checks were categorized into five conceptual domains, and feedback was obtained.
- The checks on this list will be evaluated in terms of usefulness and feasibility to determine which should be included in a tool (INSPECT-SR) for identifying problematic randomized controlled trials.

What is the implication and what should change now?

- Barriers to implementing checks were identified, including challenges in implementing statistical checks of study results.
- Feedback emphasized the importance of the tool being feasible to implement.

This prompts the question of how we can systematically detect problematic studies. The overall aim of the INveStigating ProblEmatic Clinical Trials in Systematic Reviews (INSPECT-SR) project is to develop and evaluate a tool for identifying problematic studies in the context of systematic reviews of RCTs of health interventions [3]. The INSPECT-SR tool will guide the user through a series of 'checks' for study trustworthiness. The development approach involves identifying a comprehensive list of checks for trustworthiness, and subjecting these to evaluation to determine which to include in the tool. The first objective in this process is generation of a comprehensive list of possible trustworthiness checks for evaluation in subsequent stages of the project. In addition to its use in the development of INSPECT-SR, we anticipate that this comprehensive list of trustworthiness checks will be a useful contribution to the research integrity literature.

The aim of Stage 1 of the INSPECT-SR process, reported here, was to assemble a comprehensive list of checks for potentially problematic studies, using a survey of experts and people with relevant experience. Specific objectives were to identify hitherto unidentified checks and to obtain feedback on previously identified ones.

2. Methods

The methods used in this study have been described in an online protocol (https://osf.io/6pmx5/) and in a protocol paper describing the INSPECT-SR project [3]. We give an overview here.

2.1. Assembling an initial list of checks for problematic studies

We assembled an initial list of trustworthiness checks of research studies, using several sources. Although our long-term goals in the INSPECT-SR project are to develop a tool for assessing RCTs in particular, at this stage we did not restrict the list to checks which had been proposed specifically in an RCT context. This was to ensure that we did not miss checks which could potentially be of use for assessing RCTs. However, some checks were considered as being out of scope (eg, they referred to purchasing of animals in animal studies, or related to risk of bias [4]). Excluded checks are shown in the Supplementary Material. We included checks which appeared in a recent scoping review [5] and qualitative study of experts [6]. We located and read the original studies or reports described by the scoping review to ensure that no checks were omitted. For example, the scoping review included the REAPPRAISED checklist [7] and we extracted the individual items from that checklist and included them in our list. We added additional checks which were known to the research team. For example, JW has a background in undertaking integrity investigations for journals and publishers, and he added checks used in this work. We started by including the checks from the papers included in the scoping review before adding any additional checks included in the qualitative study, and finally any additional checks known to the author team. If the same check was encountered multiple times during this process, it was added to the list only once. Some checks were considered redundant given other checks, and were excluded on this basis (see excluded checks in Supplementary Material, [5,6,8−10]). We defined five preliminary domains and categorized each check into one of these domains. The domains used were *Inspecting results in the paper*, *Inspecting the research team*, *Inspecting conduct, governance and transparency, Inspecting text and publication details*, and *Inspecting individual participant data*. The wording and categorization of the checks was reviewed by the project Expert Panel [3] and revised accordingly. The majority were rephrased as questions for consistency.

2.2. Online survey

The initial list of checks was implemented as an online survey in Qualtrics [11]. The survey can be viewed at https://osf.io/s34hx. Participants were informed about the motivation for the study and the content of the survey should they choose to participate. The survey then asked participants about their experience in assessing potentially problematic studies (with these questions being used to confirm eligibility), and presented participants with the list

of checks that could be used to assess potentially problematic studies. The checks were presented in their preliminary domains, and both the order of domains and the order of checks within each domain were randomized, to minimize the impact of potential sequence effects. Each check was presented alongside a free-text box, and participants were advised to comment on any aspect if they wished to do so. At the end of the list, participants were asked whether they were aware of any other checks which had not featured on the list, and were presented with a free text box to describe these. The survey was piloted by members of the research team and colleagues prior to launch. The survey opened on 14th November 2022 and closed on 25th January 2023. The survey was anonymous – we did not collect any identifying information in the survey. Ethical approval was not required for this study, since it involved asking experts for their professional opinion.

2.3. Participants

People with expertise or experience of assessing potentially problematic studies, either prior to or postpublication, were eligible to participate in the survey. This included editors of health journals, research integrity professionals, and researchers with experience of conducting research integrity investigations, or of undertaking related methodological research.

We implemented a multifaceted recruitment strategy. We promoted the project via conferences (International Clinical Trials Methodology Conference 2022, International Congress on Peer review and Scientific Publication 2022), social media (Twitter account of JW), and via a group of researchers and publishing representatives established to discuss problems posed by paper mills [12], inviting potential participants to contact JW. We identified and contacted individuals involved in relevant research integrity activities, including researchers, journal editors, and research integrity professionals. Additionally, the INSPECT-SR working group includes a Steering Group and an Expert Advisory Panel [3], and members of both of these were invited to participate if they met the eligibility criteria (the authors of the present article represent members of both groups). We invited eligible individuals by personalized email, and asked whether they could suggest any other potential participants. We aimed for a geographically diverse sample, and monitored responses to the question 'In which country do you primarily work?' as responses accrued. We made efforts to identify and invite potential participants based in nations which were not represented by reaching out to professional contacts in those regions and asking for suggestions for potential participants, and also by asking for suggestions from the organizers of recent and upcoming World Conferences on Research Integrity. We also identified international research integrity networks and contacted them to request details of the project to be shared with their members (African Research Integrity Network, Association for the Promotion of Research Integrity), again with a request for potential participants to contact JW.

2.4. Sample size

We targeted a minimum sample size of 50 participants, and did not end recruitment once this target was met, first because our goal was to obtain feedback from as many experts as possible within the available timeframe, and second because we did not perform any inferential statistical analyses. The sample size was largely based on pragmatic considerations – we believed 50 participants were realistic based on previous research in similar populations, for example, [13] while representing a sufficient number of responses to obtain thorough feedback on the list of the checks.

2.5. Statistical analysis

We examined survey results, including participant characteristics, using descriptive statistics. Additional items suggested by respondents, and comments made on existing items, were summarized. The survey responses were used to add further items to the list, and to amend the wording of existing items, subject to review by Steering Group and Expert Advisory panel members.

3. Results

The initial list entered into the survey contained 102 checks (76 from papers referenced by the scoping review, 14 from the qualitative study, and 12 additional checks suggested by the author team). Figure 1 shows the distribution of the checks across the five domains. Eighty individuals accessed the survey. Nine individuals did not meet the eligibility criteria (insufficient experience in assessing problematic studies). Consequently, responses were obtained from 71 participants; 12 did not complete the survey. The study dataset is available at https://osf.io/6pmx5/.

3.1. Characteristics of participants

Table 1 shows the characteristics of participants. Responses were obtained from participants based in 16 countries across five continents, although the majority (55%) of participants was based in Europe (Table 1). The experience of the included participants is also outlined in Table 1. The majority had assessed potentially problematic studies as an independent researcher (85%) with around half having done so as a peer reviewer (49%). Most had been involved in methodological research into identifying problematic studies (58%), noting that this could have referred to involvement in the INSPECT-SR project. Fewer participants had investigated potentially problematic studies as a

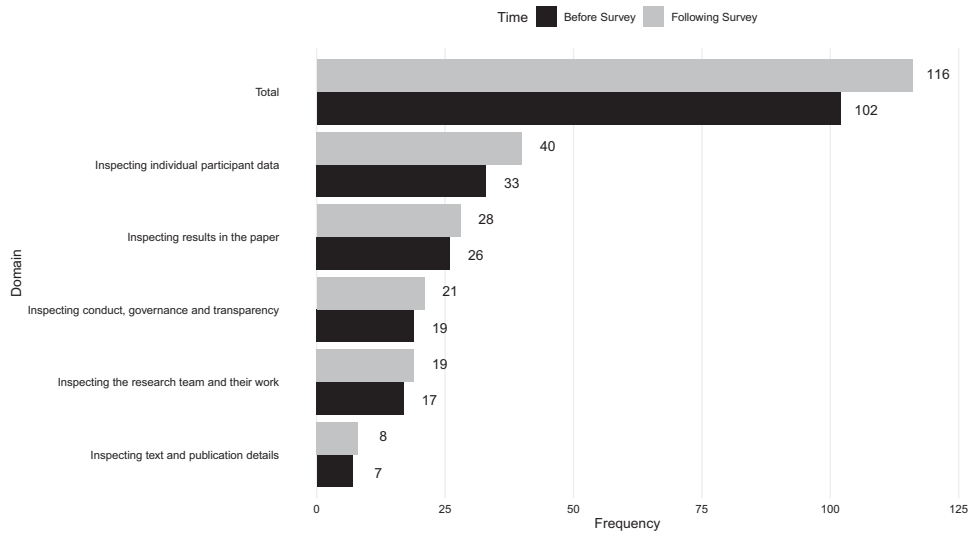

Figure 1. Number of checks in each domain before and after the survey.

journal editor (28%) or research integrity professional (27%).

3.2. Feedback on existing checks

The full list of comments by item on the list can be found in the Supplementary Material. Many suggestions revolved around specific wording changes to checks to clarify their purpose and differentiate them from each other. Feedback indicated that some checks were not well understood by participants. As an example, one check included in the domain *Inspecting individual participant data* was to 'make star plots for each group' [10,14]. This check received eight separate comments detailing participants' unfamiliarity with this concept. Similar comments were made in relation to many of the statistical checks included on the list, both in the aforementioned domain and also in the domain *Inspecting results in the paper*. Some comments indicated that the domain name *Inspecting the research team* did not clearly correspond to some of the checks contained in the domain, which referred to checking other work conducted by the research team of the index study.

3.3. Proposal of new checks

There were 38 suggestions of checks to add to the list. We were unable to interpret the meaning of four suggestions. Of the remainder, 19 suggestions, describing 14 distinct checks, were considered novel, that is, not sufficiently similar to existing checks to be considered a duplication (Table 2, with wordings edited for clarity). We categorized the proposed checks. We considered seven (50%) of the novel checks to fall within the *Inspecting individual participant data* domain. It was proposed that the country in which the study was conducted be included as a check. We have included this in Table 2 for completeness, and discuss the implications of this check in the discussion.

3.4. General feedback

Finally, participants were offered the chance to comment on the survey, or on the topic more generally. Redacted versions of these comments are included in the Supplementary Material. Redaction has been performed to conceal the identities of the participants and of the subjects of their comments. Desire for a practical, short tool was a common theme, with several participants suggesting it should be structured so that easier checks are performed first. If the outcome of these checks proved definitive (eg, identifying or assuaging serious concerns), this would avoid the use of more burdensome or complex methods appearing later in the tool.

3.5. Updated list of checks

Based on the responses to the survey, an updated list of possible checks for potentially problematic studies was developed, incorporating the new suggestions and updating the wording of items in response to feedback. One hundred and sixteen checks were included following the survey, as shown in Figure 1. The updated list is shown in the Supplementary Material [7,9,10,14–43]. Figure 2 shows the origin of checks included in the final list. In response to survey feedback, we changed the second domain name to *Inspecting the research team and their work.*

4. Discussion

We conducted a survey of experts to elaborate an extensive list of potential checks for identifying problematic

Table 1. Characteristics of participants. Frequency (%)

Characteristic	*N* (%)
Primary location of work	
Europe	39 (55%)
Australia/Oceania	15 (21%)
North America	10 (14%)
Africa	5 (7%)
South America	1 (1%)
Missing	1 (1%)
Experience[a]	
Have you assessed potentially problematic studies as an independent researcher (postpublication)?	60 (85%)
Have you conducted methodological research into the issue of identifying problematic studies?	41 (58%)
Have you assessed potentially problematic studies as a peer reviewer (prepublication)?	35 (49%)
Have you assessed potentially problematic studies as a journal editor?	20 (28%)
Have you assessed potentially problematic studies in any other capacity not listed here?	20 (28%)
Have you assessed potentially problematic studies as a research integrity professional?	19 (27%)
Have you assessed potentially problematic studies at the request of a journal or publisher?	17 (24%)
Have you assessed potentially problematic studies you have been involved in (e.g., possible misconduct by collaborators)?	10 (14%)

[a] Multiple responses permitted.

studies. We believe this is the most comprehensive collection of checks assembled to date, as we were able to supplement methods identified in a previous scoping review by returning to the original papers and extracting individual items (rather than tools aggregating items), including findings from a qualitative study, and performing a new survey of experts. The items on the list will be evaluated for their usefulness and feasibility to determine which checks should be included in the INSPECT-SR tool and any implications for the tool's structure [3]. It should be emphasized that a check's inclusion on the list does not amount to an endorsement by the research team. We anticipate that many of these checks will ultimately be found to be infeasible or simply not informative.

Participant responses highlighted a number of important considerations for the development of a tool for assessing potentially problematic studies. Despite representing a cohort of individuals with experience and expertise in problematic studies, many respondents expressed a lack of familiarity with items included on the list, particularly those relating to statistical methods. Given that the INSPECT-SR tool is intended for use by researchers without this level of expertise, our findings suggest that these checks would need to be accompanied by clear guidance to facilitate use and prevent misapplication and misinterpretation, similar to explanation and elaboration documents created to accompany reporting guidelines [44,45], or that application of these checks might need input from a statistician. This may also need to be accompanied by software to facilitate the implementation of more complex checks. In addition, this suggests that clear explanations would be needed to allow the checks to be evaluated as part of a subsequently planned consensus process [3]. Another clear theme among the survey responses related to the need for a tool to be feasible in terms of the time required to implement it. Some respondents expressed concern about the prospect of a tool involving too many checks; some had mistaken the list to represent the proposed tool, noting that it would not be workable. These concerns highlight the importance of evaluating not only the feasibility of individual items but also the practicality of the resulting tool. To this end, a draft version of the tool will be extensively tested in the production of new systematic reviews of RCTs, and revised accordingly. One proposal to increase the viability of the tool was to arrange the checks in a hierarchical format, with initial, less burdensome checks being performed first, potentially obviating more difficult checks should clear problems be apparent.

We included some checks which can only be applied when the underlying individual participant data are available in the survey. Often, these data will not be available to researchers, and so these checks will not be possible. This suggests that the core INSPECT-SR tool should not include checks requiring individual participant data. Accordingly, we will develop an extension to the core tool (working title INSPECT-IPD) which may be applied when the underlying dataset is available. Checks in the individual participant data domain were also unfamiliar to many participants, suggesting that the development of this extension would require input from subspecialists in forensic statistics.

One check which was proposed in response to the survey was to consider the country in which the study was performed. The introduction of this check would be contentious. From an empirical standpoint, while it is plausible that research misconduct would be more likely to occur in settings with limited research governance and oversight, robust evidence relating to the geographical variation in prevalence of problematic studies is relatively limited (with some exceptions, e.g., [46,47]). From an ethical standpoint, using the country of origin as an indicator of study provenance in its own right would discriminate against honest researchers based in these locations. This check will be subjected to evaluation as part of the development process.

Table 2. Novel suggestions for checks for problematic studies

Inspecting the results in the paper (2 checks proposed)
Are statistical tests internally consistent? (example: paper reports both *P* value and t statistic, but these are not consistent with each other)
Are important features missing from the paper?
Inspecting the research team (2 checks proposed)
Are withdrawal and loss to follow-up in multiple trials by the same author consistent with the expected (random) binomial distribution?
Given the nature of the study, does the author list make sense? - i.e., does a simple study have dozens of authors from different institutions and with diverse expertise.
Inspecting conduct, governance, and transparency (2 checks proposed)
In which country was the study conducted?
Is the procedure of the study aligned with local legislations?
Inspecting text and publication details (1 check proposed)
Was the time between submission to acceptance reasonable?
Inspecting individual participant data (7 checks proposed)
If authors provide an excel spreadsheet, then you could check the meta-data in the sheet, including things like when it was created, by whom, and the number of hours it's been opened. This will not be as useful if the excel is just an export from REDCap or similar.
Reorder rows by different column values: sometimes patterns become apparent, which the authors obscure by 'reshuffling' on another column value after fabricating data.
Check that when the dataset is ordered by participant ID or randomization timestamp, the N+1st participant has the same condition as the Nth 1/k of the time, where there are k conditions. If the condition assignment has been fabricated "by hand", the condition will often change too frequently as the faker tries to avoid "excessively long identical sequences.
Data fields missing from the IPD i.e., the paper reports data subgrouped by sex but sex is not available in the IPD.
Test whether a variable is a subset of a second variable within a data set.
The plausibility of the number of duplicated values (cases) across numeric variables within a data set.
An interaction test to assess the subgroup homogeneity to detect data manipulation to achieve implausible consistency (the *P* value of the Tarone-adjusted Breslow-Day test).

IPD, individual participant data.

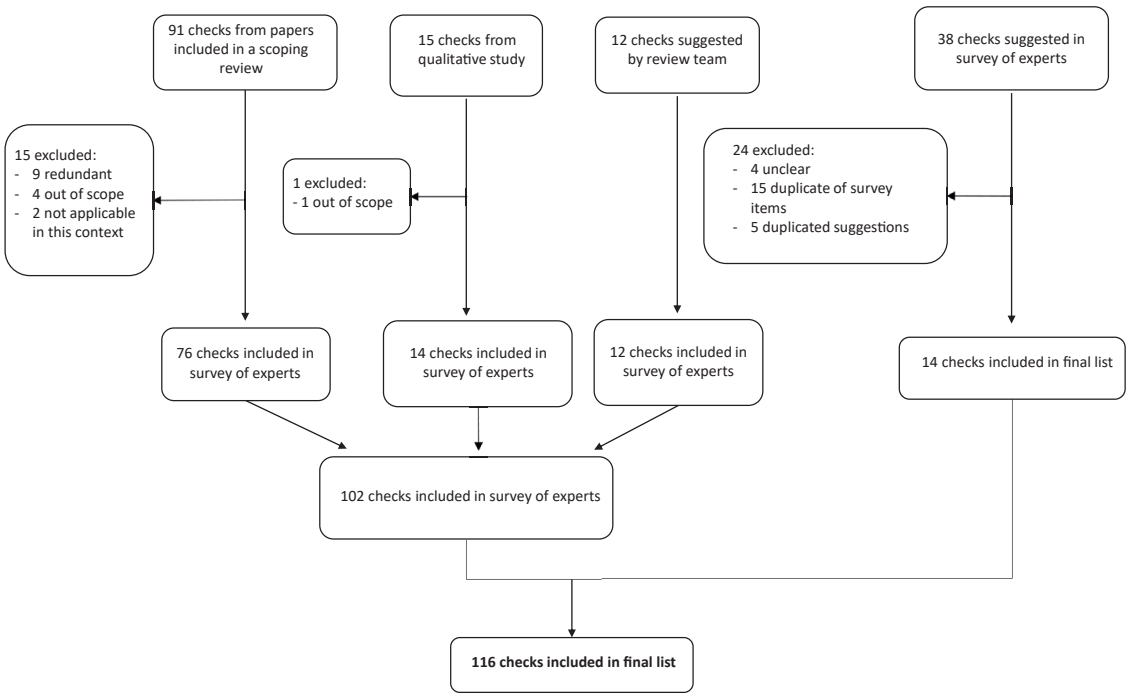

Figure 2. Flow chart showing origin of checks included in final list.

A considerable limitation of the present study is the failure to recruit many participants situated outside of Europe, Australia, and North America. Improving geographical representation in subsequent stages of the project will be necessary to ensure that the tool is both equitable and useful for the assessment of research globally. Some responses described concerns that some checks could not be reliably performed without knowledge of the local context. We also acknowledge that it is possible some checks have not been identified, and so we will ask participants in a subsequent Delphi exercise to propose any additional suggestions for evaluation to minimize the likelihood anything important is missed. We did not collect career stage or sex of the participants.

5. Conclusion

The items on the list will be evaluated via an application of the items on the list to RCTs in 50 Cochrane Systematic Reviews, an online Delphi survey, and consensus meetings, to produce a draft version of the INSPECT-SR tool. The draft version will then be subject to testing by users, and feedback from this testing will be used to improve and finalize the tool [3]. The final version will represent a feasible tool, backed by empirical evidence and broad expert consensus, for evaluating potentially problematic studies in health-related systematic reviews.

Ethics statement

The University of Manchester ethics decision tool was used on September 30, 2022. Ethical approval was not required for this study, since it involved asking experts for their professional opinion.

CRediT authorship contribution statement

Jack Wilkinson: Writing − review & editing, Writing − original draft, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Calvin Heal:** Writing − review & editing, Writing − original draft, Formal analysis, Data curation. **George A. Antoniou:** Writing − review & editing, Methodology, Funding acquisition. **Ella Flemyng:** Writing − review & editing, Methodology. **Alison Avenell:** Writing − review & editing, Methodology, Conceptualization. **Virginia Barbour:** Writing − review & editing, Methodology. **Esmee M. Bordewijk:** Writing − review & editing, Methodology. **Nicholas J.L. Brown:** Writing − review & editing, Methodology. **Mike Clarke:** Writing − review & editing, Methodology. **Jo Dumville:** Writing − review & editing, Methodology, Conceptualization. **Steph Grohmann:** Writing − review & editing. **Lyle C. Gurrin:** Writing − review & editing, Methodology. **Jill A. Hayden:** Writing − review & editing, Methodology. **Kylie E. Hunter:** Writing − review & editing, Methodology. **Emily Lam:** Writing − review & editing, Conceptualization. **Toby Lasserson:** Writing − review & editing, Methodology, Conceptualization. **Tianjing Li:** Writing − review & editing, Methodology. **Sarah Lensen:** Writing − review & editing, Methodology, Conceptualization. **Jianping Liu:** Writing − review & editing. **Andreas Lundh:** Writing − review & editing, Methodology. **Gideon Meyerowitz-Katz:** Writing − review & editing, Methodology. **Ben W. Mol:** Writing − review & editing, Methodology. **Neil E. O'Connell:** Writing − review & editing, Methodology. **Lisa Parker:** Writing − review & editing, Methodology, Conceptualization. **Barbara Redman:** Writing − review & editing, Methodology. **Anna Lene Seidler:** Writing − review & editing, Methodology. **Kyle Sheldrick:** Writing − review & editing, Methodology. **Emma Sydenham:** Writing − review & editing, Methodology, Conceptualization. **Darren L. Dahly:** Writing − review & editing, Visualization, Methodology. **Madelon van Wely:** Writing − review & editing, Methodology, Conceptualization. **Lisa Bero:** Writing − review & editing, Methodology, Funding acquisition, Conceptualization. **Jamie J. Kirkham:** Writing − review & editing, Methodology, Investigation, Funding acquisition, Conceptualization.

Data availability

The study dataset is available at https://osf.io/6pmx5/.

Declaration of competing interest

J.W., CH, GAA., L.B., and J.J.K. declare funding from NIHR (NIHR203568) in relation to the current project. J.W. additionally declares Stats or Methodological Editor roles for BJOG, Fertility and Sterility, Reproduction and Fertility, Journal of Hypertension, and for Cochrane Gynecology and Fertility. C.H. declares a Statistical Editor role for Cochrane Colorectal. L.B. 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. J.J.K. additionally declares a Statistical Editor role for The BMJ. A.A. declares that The Health Services Research Unit, University of Aberdeen, is funded by the Health and Social Care Directorates of the Scottish Government. V.B. is EiC of the Medical Journal of Australia and on the Editorial Board of Research Integrity and Peer Review. N.J.L.B. declares roles as Editorial Board member for International Review of Social Pyschology/Revue Internationale de Psychologie Sociale, Statistical Advisory Board member for Mental Health Science, and Advisory Board member for Meta-Psychology. M.C. declares that he is Co-ordinating Editor for the Cochrane Methodology Review Group,

Editor in Chief, Journal of Evidence-Based Medicine, and Coordinating Editor, James Lind Library. E.F., S.G., and T.La. declare employment by Cochrane. E.F. additionally declares a role as Editorial Board member for Cochrane Synthesis and Methods. T.La. 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). T.Li. is funded by the National Eye Institute, National Institutes of Health (Grant #UG1 EY020522). S.L. is funded by NHMRC (APP1195189), and holds general or methodological editor positions for Cochrane Gynecology and Fertility, Fertility and Sterility, and Human Reproduction. A.L. is on the editorial board of BMC Medical Ethics. B.W.M. declares roles as Editor for Cochrane Gynecology and Fertility and Sexually Transmitted Infections, and for Fertility and Sterility. S.L. declares roles as Associate Editor for Human Reproduction, Methodological Editor for Fertility and Sterility, and Editor for Cochrane Gynecology and Fertility. N.O.C. is a member of the Cochrane Editorial Board and holds an ERA-NET Neuron Cofund grant for a separate project. A.L.S. declares funding from Australian National Health and Medical Research Council Investigator Grants (GNT2009432). E.S. is a Sign-off Editor for the Cochrane Library. M.v.W. is coordinating editor of Cochrane Gynecology and Fertility and Cochrane Sexually Transmitted Infections, Methodological Editor of Human Reproduction Update and editorial Editor of Fertility and Sterility. There are no competing interests for any other author.

Acknowledgments

The authors would like to thank Richard Stevens for helpful comments during the planning of this study.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2024.111512.

References

[1] Cochrane. Cochrane Policy for managing potentially problematic studies. Cochrane Database Syst Rev: editorial policies Cochrane Library. Available at: https://www.cochranelibrary.com/cdsr/editorial-policies. Accessed September 24, 2024.

[2] 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:ED000152.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Grey A, Bolland MJ, Avenell A, Klein AA, Gunsalus CK. Check for publication integrity before misconduct. Nature 2020;577:167−9.

[8] Kirkwood AA, Cox T, Hackshaw A. Application of methods for central statistical monitoring in clinical trials. Clin Trials 2013;10:783−806.

[9] Dahlberg JE, Davidian NMJS, ethics e. Scientific forensics: how the Office of Research Integrity can assist institutional investigations of research misconduct during oversight review. Sci Eng Ethics 2010;16:713−35.

[10] Buyse M, George SL, Evans S, Geller NL, Ranstam J, Scherrer B, et al. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. Stat Med 1999;18:3435−51.

[11] Qualtrics. Qualtrics XM provo, Utah, USA2005. [cited 2024 January]. Available at: https://www.qualtrics.com. Accessed September 24, 2024.

[12] Byrne JA, Christopher J. Digital magic, or the dark arts of the 21(st) century-how can journals and peer reviewers detect manuscripts and publications from paper mills? FEBS Lett 2020;594(4):583−9.

[13] Blanco D, Hren D, Kirkham JJ, Cobo E, Schroter S. A survey exploring biomedical editors' perceptions of editorial interventions to improve adherence to reporting guidelines. F1000Res 2019;8:1682.

[14] Taylor RN, McEntegart DJ, Stillman EC. Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. Drug Inform J 2002;36(1):115−25.

[15] Nuijten MB, Hartgerink CH, Van Assen MA, Epskamp S, Wicherts JM. The prevalence of statistical reporting errors in psychology. Behav Res Methods 2016;48:1205−26.

[16] Li W, van Wely M, Gurrin L, Mol BWJF. Sterility. Integrity of randomized controlled trials: challenges and solutions. Fertil Steril 2020;113(6):1113−9.

[17] Carlisle JB. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. Anaesthesia 2017;72(8):944−52.

[18] Barnett AJF. Automated detection of over-and under-dispersion in baseline tables in randomised controlled trials. F1000Res 2022;11:783.

[19] Mosimann J, Dahlberg J, Davidian N, Krueger JJ. Terminal digits and the examination of questioned data. Acc Res 2002;9(2):75−92.

[20] Anaya JJPP. The GRIMMER test: a method for testing the validity of reported measures of variability. PeerJ Preprints 2016;4:e2400v1.

[21] Brown NJ, Heathers JAJSP, Science P. The GRIM test: a simple technique detects numerous anomalies in the reporting of results in. psychology 2017;8(4):363−9.

[22] 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;6:e26968v1.

[23] Snedecor G, William GJS, Cochran WG. Statistical methods. Hoboken, NJ: Wiley-Blackwell; 1989.

[24] Bartlett MSJPotRSoLSA-M, Sciences P. Properties of sufficiency and statistical tests. Proc Royal Soc London. Series A Math Phys Sci 1937;160(901):268−82.

[25] Bolland MJ, Gamble GD, Avenell A, Cooper DJ, Grey A. Participant withdrawals were unusually distributed in randomized trials with integrity concerns: a statistical investigation. J Clin Epidemiol 2021;131:22−9.

[26] 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.
[27] Clark L, Fairhurst C, Cook E, Torgerson DJ. Important outcome predictors showed greater baseline heterogeneity than age in two systematic reviews. J Clin Epidemiol 2015;68:175−81.
[28] Bordewijk EM, Wang R, Askie LM, Gurrin LC, Thornton JG, van Wely M, et al. Data integrity of 35 randomised controlled trials in women' health. Eur J Obstet Gynecol Reprod Biol 2020;249:72−83.
[29] Simonsohn UJ. Just post it: the lesson from two cases of fabricated data detected by statistics alone. Psychol Sci 2013;24(10):1875−88.
[30] Errami M, Wren JD, Hicks JM, Garner HR. eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. Nucleic Acids Res 2007;35(suppl_2):W12−5.
[31] Errami M, Sun Z, George AC, Long TC, Skinner MA, Wren JD, et al. Identifying duplicate content using statistically improbable phrases. Bioinformatics 2010;26(11):1453−7.
[32] Garner H, Pulverer B, Marusić A, Petrovechi M, Loadsman J, Zhang Y, et al. How to stop plagiarism. Nature 2012;481(7382):21−3.
[33] Higgins JR, Lin F-C, Evans JP. Review p. Plagiarism in submitted manuscripts: incidence, characteristics and optimization of screening—case study in a major specialty medical. journal 2016;1(1):1−8.
[34] Taylor DB. Journal club: plagiarism in manuscripts submitted to the AJR: development of an optimal screening algorithm and management pathways. Am J Roentgenol 2017;208(4):712−20.
[35] Bohannon J. Hoax-detecting software spots fake papers. Am Assoc Adv Sci 2015;348(6230):18−9.
[36] Al-Marzouki S, Evans S, Marshall T, Roberts I. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. BMJ 2005;331:267−70.
[37] Schultz BB. Levene test for relative variation. Syst Zool 1985;34(4):449−56.
[38] Brown MB, Forsythe AB. Robust tests for the equality of variances. J Am Stat Assoc 1974;69(346):364−7.
[39] Greenacre M, Ayhan HÖ. Identifying inliers. Working Papers 763, Barcelona School of Economics.
[40] Carlisle JB. False individual patient data and zombie randomised controlled trials submitted to Anaesthesia. Anaesthesia 2021;76(4):472−9.
[41] Barton D, David FJB. Multiple runs. Biometrika 1957;44(1/2):168−78.
[42] Wu X, Carlsson MJPs. Detecting data fabrication in clinical trials from cluster analysis perspective. Pharm Stat 2011;10(3):257−64.
[43] Barnett A. Automated detection of over- and under-dispersion in baseline tables in randomised controlled trials [version 2; peer review: 2 approved]. F1000Res 2023;11(783).
[44] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.
[45] Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ Br Med J (Clin Res Ed) 2010;340:c869.
[46] Woodhead M. 80% of China's clinical trial data are fraudulent, investigation finds. BMJ 2016;355:i5396.
[47] Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One 2009;4:e5738.