A survey of experts to identify methods to detect problematic studies: Stage 1 of the INSPECT-SR Project

Jack Wilkinson[1], Calvin Heal[1], George A Antoniou[2,3], Ella Flemyng[4], Alison Avenell[5], Virginia Barbour[6], Esmee M Bordewijk[7], Nicholas J L Brown[8], Mike Clarke[9], Jo Dumville[10, 11], Steph Grohmann[4], Lyle C. Gurrin[15], Jill A Hayden[16], Kylie E Hunter[17], Emily Lam[18], Toby Lasserson[4], Tianjing Li[19], Sarah Lensen[20], Jianping Liu[21], Andreas Lundh[22, 23], Gideon Meyerowitz-Katz[24], Ben W Mol[25], Neil E O'Connell[26], Lisa Parker[27], Barbara Redman[28], Anna Lene Seidler[17], Kyle Sheldrick[29], Emma Sydenham[30], Darren L Dahly[31], Madelon van Wely[7], Lisa Bero[32]*, Jamie J Kirkham[1]*

*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] Health Services Research Unit, University of Aberdeen, Aberdeen, UK.

[6] Medical Journal of Australia, Sydney, Australia.

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

[8] Department of Psychology, Linnaeus University, Sweden.

[9] Northern Ireland Methodology Hub, Queen's University Belfast, UK.

[10] Division of Nursing, Midwifery & Social Work, School of Health Sciences, The University of Manchester, Manchester, UK.

[11] NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK.

[15] School of Population and Global Health, The University of Melbourne, Australia

[16] Department of Community Health & Epidemiology, Dalhousie University, Canada

[17] NHMRC Clinical Trials Centre, University of Sydney, Australia.

[18] Independent lay member, unaffiliated, UK.

[19] Department of Ophthalmology, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA.

[20] Department of Obstetrics, Gynaecology and Newborth Health, Royal Women's Hospital, University of Melbourne, Melbourne, Australia.

[21] Director, Centre for Evidence-Based Chinese Medicine, Beijing University of Chinese Medicine, Beijing, China.

[22] Cochrane Denmark & Centre for Evidence-Based Medicine Odense, Department of Clinical Research, University of Southern Denmark, Denmark.

[23] Department of Respiratory Medicine and Infectious Diseases, Copenhagen University Hospital Bispebjerg and Frederiksberg, Denmark.

[24] School of Health and Society, University of Wollongong, Australia.

[25] Department of Obstetrics and Gynaecology, Monash University, Melbourne, Australia.

[26] Department of Health Sciences, Centre for Wellbeing Across the Lifecourse, Brunel University London, UK.

[27] Charles Perkins Centre, Faculty Medicine & Health, University of Sydney, Sydney, Australia.

[28] New York University, New York, USA.

[29] Faculty of Medicine, University of New South Wales, Australia.

[30] Cochrane Central Editorial Service, London, UK.

[31] HRB Clinical Research Facility, University College Cork, Cork, Ireland

[32] University of Colorado Anschutz Medical Campus, Colorado, USA.

Abstract

Background

Randomised controlled trials (RCTs) inform healthcare decisions. Unfortunately, some published RCTs contain false data, and some appear to have been entirely fabricated. Systematic reviews are performed to identify and synthesise 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 INSPECT-SR project is developing a tool to identify problematic RCTs in systematic reviews of healthcare-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 categorised 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.

Conclusions
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.

Main text:

Background

Randomised 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 synthesising 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 (e.g., inverting active treatment and placebo conditions), failure in the computerised randomisation 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.

This prompts the question of how we can systematically detect problematic studies. The overall aim of the INSPECT-SR (INveStigating ProblEmatic Clinical Trials in Systematic Reviews) 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.

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.

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 (e.g. 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, (5-10)). 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.

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 randomised, to minimise 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.

Participants

People with expertise or experience of assessing potentially problematic studies, either prior to or post-publication, 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 personalised 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.

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 was

realistic based on previous research in similar populations e.g. (13) while representing a sufficient number of responses to obtain thorough feedback on the list of the checks.

Statistical analysis

We examined survey results, including participant characteristics, using descriptive statistics. Additional items suggested by respondents, and comments made on existing items, were summarised. 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.

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

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

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*	
Have you assessed potentially problematic studies as an independent researcher (post-publication)?	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 (pre-publication)?	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%)

Table 1: Characteristics of participants. Frequency (%)

*Multiple responses permitted

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.

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.

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 randomisation timestamp, the N+1th 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 sub-grouped 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).

Table 2. Novel suggestions for checks for problematic studies

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 (e.g. identifying or assuaging serious concerns), this would avoid the use of more burdensome or complex methods appearing later in the tool.

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. The number of items following the survey is shown in Figure 1, and the updated list is shown in the Supplementary Material (7, 9, 10, 14-42). 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.*

Discussion

We conducted an international survey of experts to elaborate an extensive list of potential checks for identifying problematic studies. 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 emphasised 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 (43, 44), 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. (45, 46)). 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.

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.

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.

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 asking experts for their professional opinion.

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.

Declaration of interests

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. 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. AA declares that The Health Services Research Unit, University of Aberdeen, is funded by the Health and Social Care Directorates of the Scottish Government. VB is EiC of the Medical Journal of Australia and on the Editorial Board of Research Integrity and Peer Review. NJLB 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. MC 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. EF, SG and TLa declare employment by Cochrane. EF additionally declares a role as Editorial Board member for Cochrane Synthesis and Methods. 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). TLi is funded by the National Eye Institute, National Institutes of Health (Grant #UG1 EY020522). SL is funded by NHMRC (APP1195189), and holds general or methodological editor positions for Cochrane Gynaecology and Fertility, Fertility and Sterility, and Human Reproduction. AL is on the editorial board of BMC Medical Ethics. BWM declares roles as Editor for Cochrane Gynaecology and Fertility and Sexually Transmitted Infections and for Fertility and Sterility. SL declares roles as Associate Editor for Human Reproduction, Methodological Editor for Fertility and Sterility, and Editor for Cochrane Gynaecology and Fertility. NOC is a member of the Cochrane Editorial Board and holds an ERA-NET Neuron Co-Fund grant for a separate project. ALS declares funding from Australian National Health and Medical Research Council Investigator Grants (GNT2009432). ES is a Sign-off Editor for the Cochrane Library. MvW is coordinating editor of Cochrane Gynaecology and Fertility and Cochrane Sexually Transmitted Infections, Methodological Editor of Human Reproduction Update and editorial Editor of Fertility and Sterility. All other authors have nothing to declare.

Acknowledgements

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

SUPPLEMENTARY MATERIAL

Section A. List of comments.
Section B. Updated list of checks for problematic studies
Section C. Items excluded from the survey.

Section A. List of comments. Note that identifying information has been redacted.

Inspecting the results in the paper

Is the number of participant withdrawals compatible with the disease, age and timeline?	Good [authors' note: respondent wrote this for almost every question]
	Withdrawal rates between papers by the same team are often interesting: simulation can expose very unlikely common rates etc.
	What is the standard? It opens the door for guessing?
	and local context e.g. I dont think certain cultures have many withdrawals
	In many fields this is impossible to tell.
	Unclear
Are subgroup means incompatible with those for the whole cohort?	In some circumstances, rounding can cause subgroup means to not add up.
	What "means"?
	Beware of Simpson's paradox
	Are subgroup means *and variances*...
	This one is suddenly phrased as incompatible rather than compatible.
Are the reported summary data compatible with the reported range?	What does this mean?
	This criteria is highly suggestive of an error, but I fear that it has poor specificity for scientific misconduct. It is suggestive of very poor reporting.
	important
Are the summary outcome data	Would be covered by tests of dispersion above
	how many duplicates in a table count too :)

identical or nearly identical across study groups?	I think this is irrelevant.
	important
Are there any discrepancies between data reported in figures, tables and text?	including those in supplementary files?
	Or data in graphs are not given as numerical data anywhere in tables or text?
	Please, may you consider also dicrepancies between results section and abstract?
	Threshold should be identified.
	Quite frequent, not very specific
	not sure if that's a red flag for me due to printing/human error
	This one feels vague
Are any baseline data implausible with respect to magnitude or variance?	Please add prompt or signalling questions.
	important
	I am wondering if dividing this section into questions that a) require and b) don't require subject matter expertise in the research topic would be helpful? Some of these could be assessed by researchers and methodologists without knowing anything about the topic, others require intimate knowledge of the area under study.
Are any outcome data, including estimated treatment effects, implausible?	or too large ?
	Who can judge a result is implausible without replicating the trial in the same circumstances, territory, patients characteristics etc.
	Why "estimated" not "reported" or "calculated"?
	first you need to define what's implausible; secondly, I feel that 'fake' trials wouldn't go for implausible effects
	Without a range of definitions for what is covered by implausible such as which reference point, this will be very difficult to assess.
	Though I agree this is very important, this criterion seems like it is best evaluated by someone with some subject matter expertise in the area. Is there a way to make this more general? Maybe something like "How closely do the outcome data agree with treatment effects from other studies?"
Are any of the baseline data excessively similar between randomized	similar or different, these two ideas could be combined, so drop the row below on difference. It's about under- or over-dispersion
	also observing distribution plots of baseline p values. Are reported p values consistent with reported data
	need to take rounding into account (how many digits)

groups?	Rounding has effect on this. Personal evaluation, unless the mean (SD) is reported to 4 decimal places, no method can account for rounding including Monte Carlo simulation.
	Though not recommended, p-values for demographics often reported. They must be uniform.
	What does "excessively similar". What if a trial is stratified?
	or excessively different
	I would specify "not explained by stratification of randomization on these covariates or randomization by minimization"
	I cannot find the reference quickly but XXXX use one of these for their papers, is that XXXX
	important, but needs accessible tool to perform that easily
	I love the inclusion of specific examples here - references would be helpful too for those not familiar
Are any of the baseline data excessively different between randomised groups?	Particularly, if the detected excessive difference is detected in prognostic covariates, this can indicate a manipulation.
	Is this a problem for a non-stratified variable?
	excessively similar more of a problem?
	In case of fully fabricated data, one can expect perfect data, fully compliant with what is expected in the ideal case. For the detection of fraud, poor inter-patient variability is a suggestive sign
Are there any discrepancies between the values for percentage and absolute change?	not completely clear to me
	What is "percentage change"? Do you mean "relative"?
	This item should be reformulated to make it more generic. "Are there any mathematical inconsistencies between summary statistics, such as an absolute difference of proportions inconsistent with the two given proportions"
	Subsitute "relative" for "percentage" for clarity and avoid limiting the scope
Are there any discrepancies between reported data and participant inclusion criteria?	Threshold should be identified.
Are the variances	What "variances"?

in biological variables surprisingly consistent over time?	'surprisingly' is not an objective word.
	could we better define "surprisingly consistent?"
Are correct units reported?	Is this typographical?
	In my opinion, this has very poor specificity. This problem is mostly due to poor reporting.
	Love this
Are numbers of participants correct and consistent throughout the publication?	How would you know that they are "correct"?
	Including flowchart of how they arrived at that number. Could be cross-referenced to other studies that claim to use the same data.
	Often, there are errors at this level because authors do not explain well.
Are calculations of proportions and percentages correct?	Relates to missing data point below. Sometimes ITT is used as denominator, ignoring missing data.
	Is this typographical?
	This is predicated on knowing the denominator/ analysis set total, usually from Consort flowchart.
	Problematic but not very specific of misconduct in my opinion. This is suggestive of copy errors and statistics performed by non-statistician.
	important and easy to do
Are results internally consistent (for example, are there more births than pregnancies)?	Needs question on textual discrepancies, e.g. a trial for men that provides that gives results for women, or menopause dates in baseline data
	What about twins?
	Is this phrased in the "red flag" manner on purpose? I think if you ask "is it consistent", the example should be consistent, and not it is not. There should be more pregnancies than live births.
	important
	I see where this example is going, but a small amount of discrepancy would be OK due to stillbirths and multiple gestations.
Are non-first digits compatible	Not sure what is meant byt this.
	wording is unclear, I'm not sure what this means

with a genuine measurement process?	Is this Benford's law? Definitely worth using
	Is this, specifically, even digits?
	What does this mean?
	I think this phrasing confuses me quite a lot. Do you mean "are digits (other than the first) repeated in a manner that is to be expected?"
	This feels too vague to me - are you looking specifically for randomly distributed terminal digits as indicative of a low likelihood of bias in measurement? Making this a little clearer would help
Are the variances of integer data possible?	Needs explanation
	GRIM, GRIMMER I figure?
	Not something I've looked at.
	What integer data?
	important
	I would love a reference or tutorial on this
Are the means of integer data possible?	Why possible, not plausible? What integer data?
	Maybe this one could be combined with the previous question asking about variances??
Are data simulated from reported summary statistics plausible?	comparing tables across hunders (or thousands) of papers will be computationally challenging
	This is helpful but not absolute. Consider that the generated distributions from the age variable with a mean (SD) of 31.3 (3.5) represent any potential distribution of the variable with a mean between 31.25 and 31.35 and an SD between 3.45 and 3.55. This could result in billions of possible distributions, one of them represent the study arm. Even these kind of extensive simulation cannot be done in regular computers.
	What data will be "simulated"?
	Not sure what is meant here?
	I would remove the "simulated" word in this sentences
	This requires subject matter expertise that I don't have and woud love some references or information on this.
Are differences in variances in baseline variables between	What variances?
	Beware of Bartlett/F tests. They heavily rely on the normality assumption. If they are very significant (p<1e-6), they still may be useful. P-values too close to zero and too close to 1 are equally interesting (too much or not enough heterogeneity

randomised groups plausible(using summary data, e.g. F test, Bartlett)?	of variances)
	important, but needs accessible tool to perform that easily
	Love this
Are coefficients of variation plausible?	No something I've looked at.
	What coefficients?
	Redundant with "Are reported summary statistics plausible"
	some detail on how to determine this would be useful
Is the amount of missing data plausible?	and clearly reported ?
	This varies depending on how each centre counsels its patients.
	maybe reword to include - is there implausible lack of missing data
	Or investigated at all.
	Baseline characteristics, intermediate outcomes and final outcomes could have different proportions of missing data.
	Would this link to risk of bias rating - e.g., attrition
Are the results substantially divergent from the results of multiple other studies in meta-analysis	What meta-analysis?
	Will depend on the study inclusion criteria for the underpinning system review.
	I would reformulate : "implausibly divergent from the results [...] taking in account the methodology bias". Indeed, major divergence is expected with a very poor methodology without misconduct.
Is there heterogeneity across studies in degree of imbalance in baseline characteristics (in meta-analysis)	Why does this relate to imbalance, rather than the characteristics themselves?
	Is heterogeneity reasonably considered in the conclusion. I'm thinking about basing results on the prediction interval in random effect meta-analysis.
	Will depend on the study inclusion criteria for the underpinning system review.
	It would only diagnose an overall problem of the literature, that may be partly due to poor randomization schemes. Outliers in imbalance (very large or very small) are more interesting.
	I do not really understand this one. Is that a sign of fraud? Much less intuitive to me.
	This wouldn't be surprising, would it?

Inspecting the research team

Have the data been published elsewhere by the research team?	But excluding conference presentations?
	Does this need a qualifier, like 'without being acknowledged' - or does this make it too similar to the question above? Is publication of data elsewhere sufficient as a red flag?
	If for example, the data is re-published and the authors are transparent about this, would this item be used to identify the study as not problematic?
	Checking self-citation
	not always easy to ascertain
	While I agree this is important, it would be difficult to verify.
Is any duplicate reporting acknowledged or explained?	Not sure why this is additional to point below.
	Duplicates within the paper or between papers?
	golden dust...
	"In one review, we included a trial by XXX. We found this study has multiple publications which is fine, but one of these was retracted and was identical to another publication. This issue was not explained by the author team in their publications: [author note: details of publications provided]"
Are duplicate-reported data consistent between publications?	or " is there a high rate of identical and / or highly similar data across asrtic;es from the same team
	the meaning of this is unclear to me; do you mean expressly duplicated data that the authors acknowledge has been published in prior papers?
Are relevant methods consistent between publications?	Again this is too vague. Do you mean that it would be strange if a research team used one sort of standard treatment for one study but a different one in another, when if it was at the same hospital you would expect the standard care arm to be the same?
	'relevant' might need explanation?
	Between what publications?
	"of the same database ?"
	of whom?

Is there evidence of duplication of figures?	is this an authorship thing?
	this seems more about specifics than the team?
	be explicit: numbers or pictures?
	Figures as in graphs and also where images (radiological, histological etc) are presented.
	How does this compare with question above "Is there evidence of manipulation and duplication of the images".
	This seems like it should go under the "inspecting the results" section
	Or tables
Does consideration of other studies from members of the research team highlight causes for concern?	I usually take one study at a time and would not have the time to do this.
	How does this compare to the question above about retractions for other studies of the author team?
Is the distribution of non-first digits in manuscripts from one author compatible with a genuine measurement process?	Love this - but please add an explainer otherwise people will just start using similar thinking to 'P values in Table 1 = NS'
	Wording is unclear, I don't understand what the distribution of non-first digits would indicate.
	not clear why this would be limited to papers by single authors
	the validity of this method can be questioned
	I evaluated more than 2000 values in summary statistics for a group of authors with known integrity and I found then not compatible with genuine process. Perhaps, rounding has a role.
	why have you written 'non-first' rather than 'final' or 'last digits' in numerical data? (Presumably you don't mean to include middle digits of 3 digit numbers?). Why have you included the phrase 'manuscripts from one author' rather than 'in this paper'? Could consider adding 'or does it raise the suspicion of data falsification'?
	What does this mean?
	this isn't really inspecting the team?
	This would not be feasible when one author has a big number of publications.
	Same as above
	Add 'multiple' manuscripts here to clearly distinguish from earlier item?

	Unclear
	I love this one - though specifying terminal digit analysis may be more user-friendly for reviewers. And more specifically, I think looking for a signal of overrepresented 0's and 5's in the last digit place would be easiest to do.
Is the standard deviation of summary statistics in multiple studies by same authors plausible (when compared to simulated or bootstrapped data?)	I'm not sure what the "standard deviation of summary statistics" means. Does it refer to SDs that appear in the summary statistics, or the SD of the statistics themselves (as calculated by the reader)?
	Even within the same group, eligibility requirements and recruitment circumstances may vary. In developing nations, for instance, the features of recruiting patients from poor villages vary from those in other areas. In the village, women marry at a younger age and are more likely to be overweight due to their lack of full-time work and physical activity. Also as an illustration, in one of my trials, recruitment was primarily dependent on two referring satellites (one in the village and the other in the main city). The majority of women recruited from the clinic in the village were near to 22 years old, whereas those recruited from the clinic in the city were close to 30 or older. This gives the data on age a bimodal look, which is not quite common in RCTs. This will not be known until each scenario is explained individually. Although the inclusion criteria and authors are the same, the distribution of patients' age recruited for the other concurrent trial with the same inclusion criteria is different. Each trial should be evaluated in light of its unique circumstances.
	by 'summary statistics' do you mean 'baseline patient characteristics'? Could this be relevant even if identified in one study?
	Why in multiple studies?
	Be very cautious with this item... Summary statistics (e.g. mean age) are expected to be very different in the multiple studies of the same author if there are different inclusion criteria. If would restrict this criteria to "Several articles on very similar populations, with similar inclusion criteria, within the same centers, show a major inconsistency in some summary statistics that ought to be similar".
	As a reviewer, I would love to have some detail on how to do this, and some references showing appropriate use
Do all authors meet	how would you ever know?

criteria for authorship?	very hard to judge. I think contributor statements are often fictional, especially when there are bullies on the team
	how will you know that?
	This, and some other items, would/should be automated in the submission processess?
	Maybe unreasonable number of authors?
	Plausible number of authors too.
	Very difficult to gage, especially in this day and age of medical research
	not always possible to ascertain
	How would you go about checking this?
	as defined by ICJME, presumably?
Are contributorship statements present?	I'd combine this with the "complete" criterion as "present and complete"
	could be merged with first statement in this section
	Do you intend to word the questions consistently such that 'yes' and 'no' and consistently indicative of 'suspicious' eg if 'yes' = suspicious then this questions should be 'are contributorship statements lacking'
	Important!
	Is this more an issue for journal styles than problematic studies?
	see comments above
	Not relevant in my opinion. This is more dependent of the editor than of authors.
	journal specific section
Are contributorship statements complete?	As these are rarely required by journals, this might be a difficult domain to assess
	Why not ask this next to the earlier question on contributorship? Is it a reflection on the journal?
	This seems include the following "Are contributorship statements present?" - suggest merging them
	not sure if this is available in all journals
	How does this compare with the question above "Are contributorship statements present?" Could the two be merged?
Is authorship of related papers consistent?	I think this needs a bit of clarification. Do you mean if multiple papers arise from a single RCT, we would tend to expect high overlap of authors?

	Not clear to me.
	Is this considered a red flag?
	not sure if I understand the question
Can co-authors attest to the reliability of the paper?	How can judge? It opens the door for personal opinions and conflict of interest. every single publication should be judged separately. Authors may have serious honest errors that can be misclassified as integrity issues. If they progress and learn to do better research, they should be encouraged and their good publications should be seen as good.
	Can only obtain by asking them, and often there's no email for non-corresponding authors.
	when joined?
	Does this involve contacting the co-authors to ask?
Have other studies from the author team been retracted, or do they have expressions of concern, relevant post-publication amendment, or critical retraction	Each trial should stand by itself.
	Be careful about this. As a cautionary tale, see Peto R, et al. The trials of Dr Bernard Fisher: a European perspective on an American episode. Controlled Clinical Trials 1997;18:1-13.
	Post-publication amendments seem not a problem as most of them were due to unintentional errors, usually minor. PubPeer comments can be muddy. Anyone could post something, including those who have no expertise at all or have unverified intentions.
	definitely worth being aware of
	scite.ai can possibly help with this
	This is an example of a good criterion in my view because it provides specific places reviewers should look for problems (retraction watch, pubpeer)
Are the authors on staff of institutions they list?	This can be very difficult to ascertsin, especially years after publication of the article.
	Do authors have institutional email addresses?
	Not sure if this will be relevant for all countries. For example medical residents are not in thehospital's website
	this seems very useful.
	Does this discrimate against, for example, PPI?
	This seems like a typo? Do you mean, are they employed by whom they say they are employed?
	not always easy to ascertain

	How should we verify this?
Do any authors have a professorial title but no other publications on PubMed?	Why limit to professorial?
	interesting. I think only some journals include titles
	pubmed limited in coverage
	Not a check that I've done.
Does the statistics methods section use generic language, suggesting lack of expert statistical input?	Suggest things to look out for e.g. 'Begger's Test'
	will catch lots of bad practice rather than fraud, but that's still worthwhile
	Or might it suggest good, plain language writing?
	Nice. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0264360
	Is this misconduct or poor reporting? Moreover, this item should not be in the section "inspecting the research team".
	I do not think generic language suggests problems. It is more important to assess the content in statistics methods.
	may be penalising
	This is almost every paper!
	This is tricky because I wouldn't want simplicity to trigger suspicion for lack of specificity. Maybe rephrase to something like "Does the statical methods section use language that is too vague or general to permit replication?"

Inspecting conduct, governance and transparency

Is the volume of work reported by research group plausible, including that indicated by concurrent studies from the same group?	Consider adding a clarification after 'volume' (e.g. number of publications in a specified amount of time)
	This cannot be determined by guesswork and varies by country and facility. For some facilities, it is possible to enrol thouthands of individuals within a year, whereas in developed nations, a trial with 1,000 participants may require 25 participating sites and five years to complete. This criterion should not be applied to other communities (one-team-centrism [western-centrism] should not apply to other researchers from different communities). A one-by-one judgement is likely the most effective.
	important but seldom looked at

	Unless you have a platform trial situation.
	Reword: Is the volume of work and concurrent studies reported by the research group plausible?
	Maybe include suggestions for how to assess this - e.g. using the author field to search pubmed for first and last author?
Is the reported staffing adequate for the study conduct as reported?	Staffing might only be deduced from authorship of paper
	Do studies report staffing?
	Unlikely to be reported
	I do not understand this one. Do you mean that every role is filled such that it is plausible the study was conducted? For instance, a multicenter study with three authors seems unlikely (collaborators would want to be on the paper)
	I noticed that for some studies we identified for a review, many RCTs had only one author. This raises our concern about if the so-called trial is a true trial.
	This feels vague to me and I'm not sure how someone would be able to evaluate this for areas outside their subject matter expertise unless it was glaringly obvious (e.g., 3 people cannot conduct a multinational, multicenter RCT)
Is the recruitment of participants plausible within the stated time frame for the research?	Who can judge?
	sometimes time frame is not mentioned
	Brilliant one this
	this is what we looked at
	Are there any known benchmarks for this? Maybe we should develop some - could be an easy and cool paper.
Is the recruitment of participants plausible considering the epidemiology of the disease in the area of the study location?	Consider replacing disease with 'condition'. Not all trials focus on people with a disease.
	And size of research institution
	I think it will be hard to decide on the wording for a lot of these. Is it plausible? Yes. Is it likely? No. If there are lots of "unlikely" answers then they add together to a problem...
	maybe include details eg are there implausibly large numbers of participants recruited for a rare disease, or an implausibly small number recruited for a common disease and a well-resourced research study

	Developing a rule-of-thumb would be cool for this. For example, for a disease with an estimated prevalence of X per 100,000 people, a recruitment rate above Y is incompatible with that prevalence.
Is the interval between study completion and manuscript submission plausible?	Important point
	Important!
	another investigated aspect
Could the study plausibly be completed as described?	This needs an explanatory note. Does this refer to studies being conducted rather than completed?
	too vague?
	This looks vague.
	I think this might be a rather subjective question
	This feels too vague
Are the study methods plausible, at the location specified?	Please give an example or further prompt/signalling questions.
	Who can judge except the authors or their institute?
	Subjective unless you have experience in the setting
	Maybe a few examples between parentheses would help : (number and nature of examinations and interventions performed)
	another subjective question, picking only on extreme cases
Are the locations where the research took place specified, and is this information plausible?	What is this? Are the locations plausible? What does that men?
Is a funding source reported?	or funding sources, would it have provided sufficient funds for conducting the research
	Do we know that this is a risk of fraud though? Not having funding?
	The way this is posed could lead to ambiguity. Would it be better to distiguish cases in which funding source is not reported (because there was no funding source) and those in which funding is not discussed at all?
	I noticed that regarding the true or false of Chinese trials, a report mentioned that an RCT with a clear funding source statement could be true.
	Unlike some of the other categories, this one is an objective yes/no - easy for reviewers to assess.

Has the study been prospectively registered?	Needs a separate question about retrospective registration
	And Perhaps if the results are reported withnin a certain timeframe from reporting of study is complete (e.g. legal limits)
	ID of registration
	Would this call RECOVERY into question?
	Consider adding a year limit when prospective registration becomes mandatory (e.g. after 2010)
	Only applies to trials.
	in our case study was "prospectively reported2 on NCTN yet questionable
	Reword: Was the study prospectively measured?
	I wonder whether you could also ask for the reporting of the results on the trial registry (over and above registration)
	Adequate prospective registration is not happening for a lot of legitimate trials, so I am not sure this is relevant.
	This may require going to the individual registries to verify the date
Are details such as dates and study methods in the publication consistent with those in the registration documents? -	The 'historical versions' facility in ClinicalTrials.gov is very useful to trace the timeline of changes.
	might be one of the 1st things to check?
	study locations, heirarchy of outcomes, randomization methods, sample size
	This could be a multipart question.
	another investigated aspect
Is there evidence that the work has been approved by a specific, recognized committee? (ethics)	Evidence other than their own claims?
	recognised by whom?
	not sure if lack of this info would be an immediate red flag... some journals give very little space to describe methods
	Another specific yes/no question, which is great - but I wonder if the wording could be improved, to something like "Have the authors indicated that the work has been reviewed and approved by an instituitonal review board or ethics committee?"
Are there any concerns about unethical practice?	Do not make this yes no - suggest using 'Describe any concerns....'
	concerns by whom?
	vague question

	This feels like it could encompass problems more broadly than just research misconduct.
Is the grant funding number identical to the number in unrelated studies?	good, this should be expanded to ethics too, see where multiple unrelated studies have identical issues
	Again a flip in the phrasing (from "is it okay" to "is it not okay"), because I think you mean that this means they just copied in a random different grant number, right. Perhaps "is the grant funding number unique or related to similar studies?"
	there are typos in NCTN numbers not to mention grant numbers...
	This needs some nuance
	How would a reviewer know this?
Are the data publically available?	Please clarify if you mean data available outside the study publication.
	I sometimes wonder if a weird excuse not to provide the data is a bit of a flag too
	URL/Domain
	They almost never are so I dont see this being helpful at distinguishing between risky and other studies?
	Is this a good or a bad thing?
	"publicly"
	Consider clarifying what "data" means here, de-identified IPD or results published in trial registries.
	This is related to good conduct (open science), not misconduct. However, publically available data may help to detect some misconduct.
	This looks irrelevant.
	even with legit trials you might not have publically available data :(
	Or if not publicly assessible, then at least in a secure repository with proper links etc
Do the authors agree to share individual participant data?	Separate question about whether data is already provided needed
	Is this a good or a bad thing? Made up IPD can be shared more easily than real IPD
	Not specific of misconduct in my opinion. Authors rarely accept data sharing.
	answer as above

	This may not always be possible
Are additional patient data recorded in patient case records beyond what is reported in the paper?	Don't understand this question
	not clearly worded
	How could you possibly know this?
	Is this a good or a bad thing?
	Not sure this is relevant. Can you give an example here?
	Just a yes is not so informative. Is a list available with additional patient data ....
	not sure how would I be able to check that without direct contact with trialists
	Unclear
Do authors cooperate with requests for information?	there are legitimate reasons for not cooperating?
	Does the request originate from individuals who have no conflicts of interest or who do not focus their work on a certain geographical region?
	Is this a good or bad thing. Genuinely busy researchers might not have time?
	we had a person who replied to our email but the reply wasn't really helpful
Do authors provide satisfactory responses to requests?	What is the threshold of satisfactory response?
	Is this a good or bad thing. Genuinely busy researchers might not have time?
	answer above
	Do requests for study data merit a separate item?

Inspecting text and publication details

Is there evidence of copied work, such as duplicated or partially duplicated tables?	There should be a threshold before judgment.
	This involves assessment of other relevant papers. A bit of work.
	Or implausible distributions, e.g., over dispersion
	This may be beyond copied work: identical or similar (+/- 1 or 2) values in tables
	difficult to do / not feasible due to logistics (N included trials x number of authors x their other publications)
	meaning, from the same manuscript or from other manuscripts?
Is there evidence of text reuse (cutting and pasting text between papers), including text that is inconsistent with the study?	or image reuse
	suggest 'text or font that is inconsistent with the study'
	Some Cochrane reviews use standard text. Would this make them potentially "problematic"?
	I am more worried about numbers than words
	I have developed software that does this with a given set of PDFs. Working on performance improvements.
	Text reuse of methods of related apers may be appropriate
	Could this be automated? Plagiarism checks?
Are there typographical errors?	I think should be 'an unusually large number' of typos
	many publishers are increasingly limiting amount of copy editing offered, so I would not consider this a factor (unless in combination with a lot of other things, in which case typos are usually a moot issue anyway)
	clarify, not really seen this as a sign of fraud
	Could unfairly flag work from non-English speakers, but I agree that

	problematic papers are often sloppy
	not a reliable indicator
	how is this relevant for the integrity of the study?
	Typographical errors should not be a concern.
	I consider badly written manuscripts as indicate of a poorly conducted study.
	How might this identify studies that you define as "problematic" given that it is as much a role for the journal as the authors, and will bias against people who are writing in languages other than their main one
	Not relevant in my opinion, unless poor reporting is considered as scientific misconduct.
	This looks irrelevant.
	I don't see this as a red flag
	This would include many papers!
Is there evidence of automatically-generated text?	clarify please, is this tortured phrases?
	Not something I've seen / considered.
	Some Cochrane reviews use standard text. Would this make them potentially "problematic"?
	Specific examples? e.g. tortured synonyms
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?	amendments the same as corrections?
	PubPeer is potential good source of information but after validation and replication the comment provided
	I usually check PubPeer.
	Many studies are "excluded" from a systematic review for reasons that have nothing to do with being "problematic"
	Even though it is implied, you could also add letter to the editor as a specific point.
	I would add the words "due to concerns in scientific integrity" after "or has been previously excluded from a systematic review"
	important and possible to do
	see scite.ai's Reference Check feature.
	This issue is important to consider as I noted above.
Was the study	Please sign post to lists and explain what you mean by low quality journal.

published in journal from a list of predatory/ low quality journals?	Do you mean journal with substandard policies relating to editorial or peer review?
	...list of a journals with predatory practices ?
	predatory journal list/low quality is a criterium that often selects on global north publications as higher quality. what might be a more specific way of doing this?
	We may need to define low quality journals
	Or not PubMed/ Scopus indexed
	What list of predatory journals and how to define low quality journals?
	important aspect to check
	Journal quality could be a little disputed. It would be good to provide a list of what these journals might be.
	For some Chinese studies, this journal related factor could be considered in judging if an RCT is a true trial.
Is there evidence of manipulation or duplication of images?	"Duplication"?
	This question is a duplicate of a previous one

Inspecting individual participant data

Compare leading digits in individual participant data to Benford's Law	Be careful. Benford's Law (at least, the famous bit referring to the leading digits) often does not apply to scientific data because the range is (naturally) too constrained
	It is not useful for truncated data, e.g. age 18 to 40 in infertility trial. Clinical trials most probably has truncated data.
	What are "leading digits"?
	Not relevant and dangerous in my opinion. Benford's law is found when mixing statistics coming from very different horizons, but each variable will contain data that is very far from Benford's law, and it is expected that even when mixing all variables of a study they still do not follow Benford's law, although, that depends on the exact mix of variables available. For instance, the human weight has a first digit that is mainly between 5 and 9, very far from Benford's law. Overall, major departures from Benford's law are expected. Moreover, most variables have well known distributions (e.g. the human weight). Just check that distributions are consistent with the known distribution (mean, variance, skewness)!
	never heard of
Compare non-first digits in individual participant data to expected	Second digits may not be expected to be random. So check uniformity of 3rd digits (for minimum 3-digit data).
	What does this mean
	I think same as before? Perhaps harmonize the phrasing of those questions
	we don't do this
	This one feels like a repeat of an earlier question about terminal digit analysis
Compare distribution of leading digits in individual participant data between randomised arms	What are "leading digits"?
	Same as below (Benford's) but including a comparison? Unclear what that would add
	we don't do this
	I don't have the subject matter expertise to know what this is getting at - what would distribution of leading digits indicate? I'm guessing they should be similar between randomized arms?

Compare distribution of non-first digits in individual participant data between randomised arms	Can provide a false impression. For instance, if we are collecting FSH in a multicenter trial and one centre's device is miscalibrated and giving overreading, or under reading, this can produce contradictory results for non-first digits for authentic data. In addition, rounding will influence the analysis of non-first digits.
	What does this mean?
Comparing multiple versions of a spreadsheet containing study data for consistency	During data collection process, transcription errors are so common.
	What multiple versions?
	yes, basic consistency checks
Examining spreadsheet for formulae used to fabricate data	Has this happened? Are people really that stupid? That is insane.
	we don't do this formally but would be picked up when converting data to stata dta
Change global format to 'general' in Excel - calculated values display long strings of numbers to right of decimal place, fabricated values may not	It is amenable for transcription error and it is operator dependant.
	What does this mean?
	Interesting!
	we don't do this
Can the results in the paper be reproduced from the underlying dataset?	lack of reproducibility can come from many factors, not necessarily problematic
	Important. But how far do you go?
	This seems like a critically important question, but requires reviewers to essentially replicate the authors' entire analysis. Outside of special circumstances like reviewing cases of suspected fraud, I bet this is rarely (if ever) done.
Statistical test to compare variances in baseline variables between groups using IPD (Levene, Brown-Forsythe)	What variances?
	never heard of
Identifying inliers using singular value decomposition	I don't know what this is
	i don't know what this is!?
	What does this mean?
	i don't understand this one

	never heard of
	Unclear
Identifying inliers using Mahalanobis distance	I don't know what this is
	not familiar with this
	What does this mean?
	never heard of
	Unclear
Colour code values in Excel spreadsheet to highlight outlying values, patterns and repetition	no
	Why?
	we don't do this
	Unclear
Plot column values in order provided by author and by group (check for repetition, patterns, differences in patterns between groups)	column and row
	Same as before?
	we check for patterns
	An example of this would be useful
Plot differences between consecutive column values in order provided by author and by group (check for repetition, patterns, differences in patterns between groups)	What does this mean?
	we check for patterns and repetitions
Probability of column sequences via simulation and resampling	I don't know what this is
	Useful but has different application and it is operator dependant, particularly if we are going to consider the cumulative sequence.
	What does this mean?
	we don't do this
Test of runs of the same value (e.g. resampling, Wald-Wolfowitz)	What does this mean?
	never heard of
Checks of sequences in decimal places (after deleting integer)	I don't know what this is
	And checks of sequence of integers after deleting decimals.
	What does this mean?

	This is unclear to me? You mean any continuous values?
	we don't do this
Examine relationships between variables for biological plausibility (e.g. by plotting against each other)	this - and several others below - is not phrased as a question
	probably a good idea but we don't do it
Statistical test to compare multivariate correlations between variables between treatment groups (several methods)	What does this mean?
	add the words, "shows unexplained heterogeneity"
	we don't do this
	some example methods and references would help
Plot correlation coefficients for each pair of variables by group using greyscale/ heatmap	Or plot ORs for binary data, e.g. we found an OR of about 30000 for some in a vitamin D paper, where it seemed columns of data had been copied with only 1 or 2 changed
	what pairs?
	never heard of
Compare kurtosis of baseline variables between groups	I don't know what this is
	Also, evidence of truncated distributions?
	What does this mean?
	Not very relevant in my opinion. Randomization should lead to the same kurtosis, but kurtosis sampling fluctuations are known to be extreme, so that very large differences are expected due to randomness.
	we don't do this
Consider whether distribution of variable follows simple but implausible model (such as Normal)	Is a normal distribution implausible for health data?
	Indeed, non-significant normality tests on large samples (>2000) may be indicative of a database generated by software, especially if it is found on several variables.
	You mean it follows it 'too' well?
	yes, basic check
Check repeated measures for interpolation and duplication	I don't know what this is
	never heard of
	I don't know what this one means
Inspect recruitment over	Will you get dates?

calendar time (compare between groups)	we perform thorough checks on dates (T2E data)
Inspect time between participant visits	Why?
	Unclear to me.
	yes, basic check on dates
	What would one be looking for here?
Check visit dates (plausibility of visits on Sundays)	Or Friday in Islamic societies, Saturday in Israel.
	I would change to 'plausibility of visit dates on holidays/weekends'
	as long as dates are collected in local time, or can be set to local time
	Other countries have different weekends, e.g., Egypt weekend is Friday.
	Not a problem in some countries or for some conditions
	Or Fridays, depending on location
	and public holidays
	Rephrase, perhaps: "Are there irregularities in the dates of visit? (for instance, often on Sundays)"
	we check days of randomisation
Plot Chernoff faces for each group	I was not aware that these had any application for diagnosing problematic data
	I don't know what this is
	doesn't seem that helpful
	What does this mean?
	never heard of
	Unclear
	very cool!
Make Star Plots for each group	I am unfamiliar with star plots
	I don't know what this is
	no
	What does this mean?
	again, not familiar with this
	Not sure what that is
	never heard of
	Unclear

	this section contains many items that feel more like a laundry-list than other sections. Maybe organizing it with subsections of what type of fraud or fabrication we are trying to detect would be useful, or what kind of data are available to work with? That might help users choose between the various methods.
Apply neighbourhood clustering method of Wu and Carlsson	I don't know what this is
	not familiar with this
	What does this mean?
	never heard of
	Unclear
	Specifying the objective of doing this (and all of these, actually) would be useful - e.g., "to detect X"
Are data internally consistent?	not sure what this means
	Maybe give an example, for instance dose of drug infused between minimum and maximum rates during a reported time vs the total dose infused.
	This looks vague.
	yup that's what we do in our IPD checks
	How? What checks would be helpful here?
	Unclear
	this question is very general - could we specify a bit more/
Are only a small number of baseline characteristics collected in IPD	not sure this is a sign of fraud, happy to be corrected
	Is this a good or a bad thing?
	if the database analyzed is supposed to be the complete database.
	This looks unimportant.
	Small is less than 4?
	in small trials you will have a small number of baseline characteristics... also you might not have access to the entire dataset due to limited budget (in one IPDMA we could afford to access "buy" only a subset of variables)
Calculate autocorrelation between column values, overall and by group	What does this mean?

Check whether the randomisation sequence consistent with the description in the paper	Also, given knowledge of block sizes, are the ITT group numbers identical, or larger than permitted by blocks?
	yes, basic check

Do you know of any methods not listed here, or do you have any to suggest? Please describe them here if so.

no
No
"Reported information on consent process of trial participants. Willingness of triallists to share such information. Use of paper methods rather than electronic data capture methods for patient reported outcome measures."
No additional thoughts.
Based on some findings which created a stir in my field (detailed here: http://deevybee.blogspot.com/2015/02/editors-behaving-badly.html), we have begun looking at submission to acceptance latency by extracting meta-data from PubMed. Most journals in our sample do not publish both, so we've then moved to emailing journal editors directly. Several have provided dates, but some (notably, The Lancet) have refused. One can then examine the distribution of latency to identify outliers, and use meta-regression or subgroup analysis to examine whether effects for these studies are significantly larger.
"I have additional suggestions for some of the sections, please see below. Some of these may be already encompassed by existing questions: *Inspecting research team* Given the nature of the study, does author list make sense - i.e. if you have a very simple study (esp.

theoretical/in silico) with dozen authors from different institutions and very diverse expertise. -> captured more generally in criteria for authorship, but this is more specific case Do any of the authors have inflated number of publications the topic of which is not universally aligned with this author's expertise. Do emails provided for the authors - esp. if institutional - align with stated affiliations. *Inspecting publication details* Does the speed of consideration appears unusually short (i.e. suspiciously quick time from submission to acceptance). *Inspecting results* Are statistical tests internally consistent (example: paper reports both p-value and t statistic, but these are not consistent with each other) Where original data for graphs provided (usually on request for integrity investigations), can graphs be reproduced. In particular, recapitulation of reported error bars. *Inspecting conduct/governance* Where ethical approvals provided (usually on request for integrity investigations): do dates on approval align with the description of when the study happened Is the procedure of the study aligned with local legislations - this is more for animal research, when legislation can differ significantly between countries, but there is still expectation that what is done, is at least legal in the location where the study happened. Can be important esp. if authors in various countries and laws more permissive not where actual work took place."
A very rough measure (although many other measures are the same) - publications from certain areas of the globe? Or originally published in other languages than english? The size of the study (e.g. number of enrollera participants)?

"'- are important features missing from the paper (eg test statistic or degrees of freedom expected but missing; figure expected but missing - it happens) - are there surprisingly few authors - eg a senior professor single-authoring a large experimental study - are data in the publication consistent with the preregistration - if the ethics review number is given, is it consistent with the enrolment start date (for example not ""2021-PQRST"" if enrolment started in in 2020"
"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. Perhaps a mention of the use of REDCap or similar clinical trial management software. I suspect people using these tools are less likely to be fraudulent. "
I would like to suggest if any concern regarding the amount of data collected (the robustness of the trial) and the number of authors and funding reported. Also, to consider the main country from the report of the trial come from.
Check 'information' of file to determine when it was authored and by whom. Reorder rows by different column values: sometimes patterns become apparent, which the authors obscure by 'reshuffling' on another column value after fabricating data. Nominal variables sometimes have patterns - for instance, I analysed a submission in which the first syllable of names contained repeated sequences. The title of the paper should be searched via Google, ResearchGate etc as well as Pubmed.
Are all methods in Weibel et al.,""Identifying and managing problematic trials: a research integrity assessment tool for randomized controlled trials in evidence synthesis"" included?
"Are the results reproducible at all? (if not, this is often used to hide a lot of issues in ambiguity) Multi-center trial comparisons (meta-analysis often results in detecting issues as they are so discrepant from everything else) Could compare hash files of the data provided as real (i.e., forensic investigation of the data origins) - this requires a bit more but there is value in storing these to detect changes over time in a secure manenr."
Check that when the dataset is ordered by participant ID or randomisation timestamp, the N+1th 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".
1- A frequency distribution table or graph for baseline characteristics can give a good impression of the quality of randomisation at first glance. 2- The plausibility of the number of duplicated values (cases) across numeric variables within a data set. 3- Test whether a variable is a subset of a second variable within a data set. 4- Test binary data for consistency if their mean and SD are given. 5- 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). 6- Check whether withdrawal and loss to follow-up in multiple trials by the same author are consistent with the expected (random) binomial distribution (check Carlisle 2012 and Bolland 2020). 7- For trials that did not report mean but reported median, the Box-Cox method of McGrath et al. (2020) and the MLN method of Cai et al. (2021) can be used to convert the median (IQR or range) to mean (SD). The converted mean (SD) can be used for regular checks of summary stats. Whether the Box-Cox method of McGrath et al. (2020) and the MLN method of Cai et al. (2021) are compelling needs to be evaluated.
Consider the legitimacy of the journal where the study has been published (i.e a predatory journal or publisher) Other papers which may be covered already in the inspecting results but just in case: https://pubmed.ncbi.nlm.nih.gov/28412468/ https://journals.sagepub.com/doi/full/10.1177/1948550616673876 https://www.semanticscholar.org/paper/Using-Statistics-from-Binary-Variables-to-Detect-Schumm-Crawford/52008d10f8f942672cb4d1334978e5299bd213b4
Not sure if I missed it. But what about data fields missing from the IPD i.e. the paper reports data subgrouped by sex but sex is not available in the IPD

If you have any comments about this survey, or about this topic more generally, please add them here.

It seems it would be very difficult to get some of the information to answer these questions.
Signalling questions or prompts for some of these questions will be needed to help guide users of the tool to look at the right bits of information.
It would be good to list simple first line tests for integrity to start with, not necessarily involving much statistical knowledge, to use before moving to more complex methods.

Something else I've noticed in potentially problematic studies are samples that appear to be connected across studies even though articles fail to indicate they are connected or even imply they are separate samples. In these cases, the investigator has published what appear to be multiple separate RCTs of a method, but in fact the sample is the same and the data has simply been re-analyzed with a few participants removed. I can't quite say what tips off when this is happening- sample sizes that are quite close, sample means that are very similar?
No comments. This looks like a great resource - best of luck with the project!
"Something that systematic reviewers would appreciate is a tool that is easy to use and possibly not too time intensive. A thought is to have both a short version of a research integrity tool as well as a longer version. Or implementing it at different parts of the SR process (e.g. basic first screen and then a more intense assessment further to the finished process, or as a part of a sensitivitet analysis)" It would be good to work with ScreenIT on this. They are a great group and very collaborative in my experience.
Very complete. Thanks for having me here.
We could do with a searchable record of authors who have submitted unpublished false data (similar to the RetractionWatch database of published papers), populated by verified journal editors. Most of the false trials I identified have been published elsewhere (often with major changes); there is no mechanism to reduce this happening. I built a sandcastle against the tide.
Is it worth considering conflict of interest (declared or not) - often a source of considerable bias which is certainly problematic, clearly uncontrolled. Depends on what one considers a problematic trial - is it one that cripples reproducibility? Barbara
"Lots falls and stands based on the details of implementing. Lots of these methods are theoretically feasible, but their in-practice validity is hard with low prevalence of fabrication. If applied often and without prior suspicion, this can lead to many false positives and reduces trust in the institutions providing these assessments. It would be best to take a two-staged approach - if you suspect an author/group based on a paper, move to other papers to see whether those too have problems. What also works particularly well is comparing studies of a similar nature - genuine and ungenuine data will be often easier to sort out. If it's harder to sort out, the effect of the ungenuine data (if present) will also be smaller and the risk is smaller)."
Very comprehensive - well done to the team!
This is fairy comprehensive. Clustering methods may be useful but I haven't used them. In the only retraction I've been involved in, the individual patient data was so poor that it didn't merit any formal statistical methods: the rows in the dataset did not correspond to individual patients (it seemed that

each column of data was constructed individually: pairwise scatterplots were sufficient to detect the problem). Also, the randomisation sheet was just a set of random numbers - but no treatments listed beside them. I don't think the full gambit of methods described above are necessary for every case. I'll e-mail you my checklist for new submissions - it's very similar to many lists out there - but it doesn't go into methods for individual patient data assessment.
Some of the questions in this survey are very technical or difficult to understand. Have the people who suggested them run the suggested tests against IPD from their own trials?
This is really important work!
"V useful list. I've thought for ages that we need a Uni to set up a MSc coures to train people in methods of fraud detection. Maybe Manchester could be persuaded to do it. I'm now retired and doing this kind of thing on an amateur basis. For instance, I have reservations about a study by XXXX that I've flagged on PubPeer - I haven't tried to get the data, but the study seemed particularly problematic in terms of plausibility of time scale etc. https://pubpeer.com/publications/888BF5CC8DBCD2B080DC4189AA604D. I've tried to get confirmation from the Ethics organisation in iran but not had any success. One other thing: this task would be made MUCH easier if all ethics approvals were made open. There seems huge reluctance to do that. XXXXXX"
This seems a very exhaustive and comprehensive list.
"As scientific misconduct is a continuum, it may not be easy to distinguish between scientific misconduct, poor methodology, poor reporting and major protocol deviations, with possible overlap between these notions. For instance, I found an article where the methods section specified that some costy and hard-to-retreive data was collected (SARS-Cov-2 RT-PCR every week in 10000 asymptomatic indian healthy volounteers), but there were no result related to these data in the Results section, suggesting that it was actually never done. Strangely, there extensive data with poor clinical relevance was reported (serological responses) while the most important results (COVID-19 symptomatic and asymptomatic infections) had very little reporting in the Results section.

I suspect that there were major deviations from the original protocol and that most of the clinically relevant data could not be used, so that authors reported as few data as possible on that. However, they reported nothing on these problems and even claimed to have very few lost to follow-up. To what degree is there an intentional hiding of these problems? Where does start the poor reporting? Where does start the scientific misconduct? Is poor reporting a scientific misconduct? "
"Wow there are some methods there I never heard of. Great work!"
Are you working with XXXXXX on this?
"A technical note, the space to provide comments is very limited. On the topic, most of listed activities are sensible but from the practical point of view not sure how one would be able to perform all the checks within a reasonable time when we already know that syst. reviews take too much time to complete. As for the IPD checks, haven't heard about 60% of listed methods "
"Just adding my response here per our meeting to note my software in development, that will permit semi-automatic assessments of some of the items in this list (starting with the distribution of baseline characteristics from RCTs). I hope to have it online in the next month. - XXXXXX"
I would have like to be able to score the items as to how important they were.
Thanks for the opportunity to comment. My first reaction is that this is a great list. My second reaction is that there are a lot of items to check. Would the average reviewer have the expertise and time to go through this list? Could the list be shortened or prioritised (e.g., the retraction question might be the first one, in which case the reviewer wouldn't need to proceed)? Are there opportunities to automate some of these steps (e.g., plagiarism, attaching a link of retractions connected to the paper or author team)? All the best
I think methods should be divided in those that can be used more for screening (so high sensitivity, even with the risk of lots of false positives) and those that are more diagnostic.

Many aspects relevant for the research integrity assessment (e.g. assessing proper randomization, author contributorship, funding details, ethics...) can only be checked if sufficient details are reported in a publication, protocol or registry record. Problematic studies often do not report in sufficient detail on aspects relevant for assessment of RI. Thus, I believe that insufficient reporting or non-reporting itself on several relevant RI aspects is a matter of ignoring research integrity and should be a red flag.
It seems like it would be useful to further organise and categorise the list as presented here. e.g. Inspecting results and inspecting data should juxtaposed; different approaches to testing for fabricated data (digit analysis vs multivariate distributions) should be grouped; etc.
It might be useful to review the reported cases noted above and other more, e.g. retracted papers, famous investigation reports (https://www.nature.com/articles/d41586-021-00733-5) and propose other methods to inspect the problems?
This is really useful and I am so eager to see the final product, and learn more about the various methods you have mentioned. If I can be helpful in any way I'd love to remain involved. My stake is as a consumer of the medical literature and critical appraisal enthusiast responsible for teaching clinicians how to read the literature, which includes knowing how to detect fraud. In case there is not space for this later, my name is XXXXXXX. Thanks for doing this important work!
"Do not make the list too long. Aks yourself the question 'are the data true' and in doubt, ask for original data? PS: I really like the input of Jack and his team in the field of research integrity; you are making a difference!!"
This is a very comprehensive list of identifiers, much more than I have thought of previously.
"This is an excellent list and I have nothing much to add (I have limited statistical expertise) except under 'inspecting the research team: Are there authors on the paper from the country from which the data was collected (i.e to counter safari/parachute research). "
Diagnostic yield versus time taken to test. Major impediment is probably getting access to the raw data, and some simple tests (such as time to get the data) may end up performing as well as more complex ones. I think this is a structural problem with the medical literature, and is solvable to a large extent by deconstructing study information into freely accessible bits which are commented on by content experts for that piece (much would still be with statisticians) and then the linked red flag appears in the public domain.

Section B. Updated list of checks for problematic studies

Checks are arranged in five domains. Numbers in brackets/ parentheses are citation in main text.

Domain 1: Inspecting the results in the paper (28 checks)

Check
Is the number of participant withdrawals compatible with the disease, age and timeline?(7)
Are subgroup means incompatible with those for the whole cohort?(7)
Are the reported summary data compatible with the reported range?(7)
Are the summary outcome data identical across study groups?(7)
Are there any discrepancies between data reported in figures, tables and text?(7)
Are statistical test results compatible with reported data?(7, 15)
Are any baseline data implausible with respect to magnitude or variance? (7)
Are any outcome data, including estimated treatment effects, implausible?(16)
Are any of the baseline data excessively similar between randomized groups?(17, 18)
Are any of the baseline data excessively different between randomised groups? (17, 18)
Are there any discrepancies between the values for percentage and absolute change?(7)
Are there any discrepancies between reported data and participant inclusion criteria?(7)
Are the variances in biological variables surprisingly consistent over time?(7)
Are correct units reported?(7)
Are numbers of participants correct and consistent throughout the publication?(7)
Are calculations of proportions and percentages correct?(7)
Are results internally consistent?(7)
Are non-first digits compatible with a genuine measurement process?(19)
Are the variances of integer data possible?(20)
Are the means of integer data possible?(21)
Are data simulated from reported summary statistics plausible?(22)
Are differences in variances in baseline variables between randomised groups plausible(using

summary data)?(23, 24)
Are coefficients of variation plausible?
Is the amount of missing data plausible?(25)
Are the results substantially divergent from the results of multiple other studies in meta-analysis(26)
Is there heterogeneity across studies in degree of imbalance in baseline characteristics (in meta-analysis)(27)
Are statistical tests internally consistent? (example: paper reports both p-value and t statistic, but these are not consistent with each other) [origin: from current survey]
Are important features missing from the paper? [origin: from current survey]

Domain 2: Inspecting the research team and their work (19 checks)

Have the data been published elsewhere by the research team in an illegitimate fashion?(7)
Is any duplicate reporting acknowledged or explained?(7)
Are duplicate-reported data consistent between publications?(7)
Are relevant methods consistent between publications?(7)
Is there evidence of duplication of figures?(7)
Does consideration of other studies from members of the research team highlight causes for concern?(6)
Is the distribution of non-first digits in manuscripts from one author compatible with a genuine measurement process?(28)
Is the standard deviation of summary statistics in multiple studies by same authors plausible (when compared to simulated or bootstrapped data?)(29)
Do all authors meet criteria for authorship?(7)
Are contributorship statements present?(7)
Are contributorship statements complete?(7)
Is authorship of related papers consistent?(7)
Can co-authors attest to the reliability of the paper?(7)
Have other studies from the author team been retracted, or do they have expressions of concern, relevant post-publication amendment, or critical Retraction Watch or PubPeer

comment?(6)
Are the authors on staff of institutions they list?(6)
Do any authors have a professorial title but no other publications on PubMed?(6)
Does the statistics methods section use generic language, suggesting lack of expert statistical input?(6)
Are withdrawal and loss to follow-up in multiple trials by the same author consistent with the expected (random) binomial distribution? [origin: from current survey]
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. [origin: from current survey]

Domain 3: Inspecting conduct, governance and transparency (21 checks)

Is the volume of work reported by research group plausible, including that indicated by concurrent studies from the same group?(7)
Is the reported staffing adequate for the study conduct as reported?(7)
Is the recruitment of participants plausible within the stated time frame for the research?(7)
Is the recruitment of participants plausible considering the epidemiology of the disease in the area of the study location?(7)
Is the interval between study completion and manuscript submission plausible?(7)
Could the study plausibly be completed as described?(7)
Are the study methods plausible, at the location specified?(7)
Are the locations where the research took place specified, and is this information plausible?(7)
Is a funding source reported?(7)
Has the study been prospectively registered?(7)
Are details such as dates and study methods in the publication consistent with those in the registration documents?(7)
Is there evidence that the work has been approved by a specific, recognized committee? (ethics)(7)
Are there any concerns about unethical practice?(7)

Is the grant funding number identical to the number in unrelated studies?(6)
Are the data publically available?(6)
Do the authors agree to share individual participant data?(6)
Are additional patient data recorded in patient case records beyond what is reported in the paper?(6)
Do authors cooperate with requests for information? [author team suggestion]
Do authors provide satisfactory responses to requests? [author team suggestion]
In which country was the study conducted? [origin: from the current survey]
Is the procedure of the study aligned with local legislations? [origin: from the current survey]

Domain 4: Inspecting text and publication details (8 checks)

Is there evidence of copied work, such as duplicated or partially duplicated tables?(7)
Is there evidence of text reuse (cutting and pasting text between papers), including text that is inconsistent with the study?(7, 30-34)
Are there typographical errors?(7)
Is there evidence of automatically-generated text?(35)
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?(6)
Was the study published in a journal from a list of predatory/ low quality journals?(6)
Is there evidence of manipulation or duplication of images?(7)
Was the time between submission to acceptance reasonable? [origin: from current survey]

Domain 5: Inspecting individual participant data (40 checks)

Compare leading digits in individual participant data to Benford's Law(28)
Compare non-first digits in individual participant data to expected (36)
Compare distribution of leading digits in individual participant data between randomised arms [author team suggestion]
Compare distribution of non-first digits in individual participant data between randomised

arms(36)
Comparing multiple versions of a spreadsheet containing study data for consistency(9)
Examining spreadsheet for formulae used to fabricate data(9)
Change global format to 'general' in Excel - calculated values display long strings of numbers to right of decimal place, fabricated values may not(9)
Can the results in the paper be reproduced from the underlying dataset?(9)
Statistical test to compare variances in baseline variables between groups using IPD (Levene, Brown-Forsythe)(36-38)
Identifying inliers using singular value decomposition(39)
Identifying inliers using Mahalanobis distance(10)
Colour code values in Excel spreadsheet to highlight outlying values, patterns and repetition(40)
Plot column values in order provided by author and by group (check for repetition, patterns, differences in patterns between groups)(40)
Plot differences between consecutive column values in order provided by author and by group (check for repetition, patterns, differences in patterns between groups)(40)
Probability of column sequences via simulation and resampling(40)
Test of runs of the same value (e.g. resampling, Wald-Wolfowitz)(41)
Checks of sequences in decimal places (after deleting integer)(40)
Examine relationships between variables for biological plausibility (e.g. by plotting against each other)(10)
Statistical test to compare multivariate correlations between variables between treatment groups (10)
Plot correlation coefficients for each pair of variables by group using greyscale/ heatmap(14)
Compare kurtosis of baseline variables between groups(10)
Consider whether distribution of variable follows simple but implausible model (such as Normal)(10)
Check repeated measures for interpolation and duplication(10)
Inspect recruitment over calendar time (compare between groups)(10)
Inspect time between participant visits(10)

Check visit dates (plausibility of visits on Sundays)(10)
Plot Chernoff faces for each group(10, 14)
Make Star Plots for each group(10, 14)
Apply neighbourhood clustering method(42)
Are data internally consistent? [author team suggestion]
Are only a small number of baseline characteristics collected in IPD(6)
Calculate autocorrelation between column values, overall and by group [author team suggestion]
Check whether the randomisation sequence consistent with the description in the paper [author team suggestion]
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. [origin: from current survey]
Reorder rows by different column values: sometimes patterns become apparent, which the authors obscure by 'reshuffling' on another column value after fabricating data. [origin: from current survey]
Check that when the dataset is ordered by participant ID or randomisation timestamp, the N+1th 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. [origin: from current survey]
Data fields missing from the IPD i.e. the paper reports data sub-grouped by sex but sex is not available in the IPD. [origin: from current survey]
Test whether a variable is a subset of a second variable within a data set. [origin: from current survey]
The plausibility of the number of duplicated values (cases) across numeric variables within a data set. [origin: from current survey]
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). [origin: from current survey]

Section C. Items excluded from the survey. Numbers I brackets/ parentheses are citation in main document.

Excluded check	Origin	Reason for exclusion
Is the number of participant deaths compatible with the disease, age and timeline?	Review of papers included in scoping review (7)	Covered by checks relating to plausibility of study outcomes
Are any data impossible?	Review of papers included in scoping review (7)	Covered by several more specific checks
Are any of the outcome data unexpected outliers?	Review of papers included in scoping review (7)	Covered by *Are any outcome data, including estimated treatment effects, implausible?*
Are any data outside the expected range for sex, age or disease?	Review of papers included in scoping review (7)	Covered by several more specific checks
Are the results of statistical testing internally consistent and plausible?	Review of papers included in scoping review (7)	Covered by *Are statistical test results compatible with reported data?* [Note: was subsequently added to the list in response to the survey]
How many data are duplicate reported?	Review of papers included in scoping review (7)	Covered by checks relating to duplicate reporting.
Are other data errors present?	Review of papers included in scoping review (7)	Considered too vague, covered by other more specific items.
Investigating all publications of one author	Scoping review	Covered by several more specific items (what and how to check the work of the author)

Inspect variances over time	Review of papers included in the scoping review (10)	Covered by *Are the variances in biological variables surprisingly consistent over time?*
Was a reporting checklist used?	Qualitative study	Outside of scope (reporting quality)
Do the numbers of animals purchased and housed align with numbers in the publication?	Review of papers included in scoping review (7)	Could not apply to human studies
Have the correct analyses been undertaken and reported?	Review of papers included in scoping review (7)	Outside of scope (study quality)
Is there evidence of poor methodology, including: missing data, inappropriate data handling, 'P-hacking': biased or selective analyses that promote fragile results, other unacknowledged multiple testing.	Review of papers included in scoping review (7)	Outside of scope (risk of bias)
Is there outcome switching — that is, do the analysis and discussion focus on measures other than those specified in registered analysis plans?	Review of papers included in scoping review (7)	Outside of scope (risk of bias)
Have not included specific methods for statistical monitoring/ fraud detection of multicentre trials using IPD (trying to detect fabrication at one site by comparing to others)	Scoping review (e.g. (8))	Not applicable outside of central trial monitoring context
Checking grant applications for inconsistencies in reported preliminary results	Review of papers included in scoping review (9)	Proposed in context of institutional integrity investigations – relies on

		access to grant applications.

References

1. 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.
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(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(7789):167-9.
8. Kirkwood AA, Cox T, Hackshaw A. Application of methods for central statistical monitoring in clinical trials. Clin Trials. 2013;10(5):783-806.
9. Dahlberg JE, Davidian NM. 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(24):3435-51.
11. Qualtrics. Qualtrics XM Provo, Utah, USA. [cited 2024 January]. Available from: https://www.qualtrics.com.
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. Ther Innov Regul Sci. 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 (1985–2013). Behav Res. 2016;48:1205-26.
16. Li W, van Wely M, Gurrin L, Mol BWJ. Integrity of randomized controlled trials: challenges and solutions. Fertility and Sterility. 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 [version 2; peer review: 2 approved]. F1000 Research. 2023;11:783.
19. Mosimann J, Dahlberg J, Davidian N, Krueger J. Terminal digits and the examination of questioned data. Accountability in Research. 2002;9(2):75-92.
20. Anaya J. The GRIMMER test: A method for testing the validity of reported measures of variability. Peer J Preprints. 2016;4:e2400v1.
21. Brown NJ, Heathers JA, Science P. The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. Social Psychological and Personality Science. 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. Report No.: 2167-9843.
23. Snedecor G, Cochran WG. Statistical methods, 8th ed. Wiley-Blackwell. 1989:84-6.
24. Bartlett MS. Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences. 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(2):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 U. Just post it: The lesson from two cases of fabricated data detected by statistics alone. Psychological Science. 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-W5.
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 J. Plagiarism in submitted manuscripts: incidence, characteristics and optimization of screening—case study in a major specialty medical journal. Res Integr Peer Rev. 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. American Journal of Roentgenology. 2017;208(4):712-20.
35. Bohannon J. Hoax-detecting software spots fake papers. Science. 2015; 348(6230)
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(7511):267-70.
37. Schultz BB. Levene Test for Relative Variation. Systematic Zoology. 1985;34(4):449-56.
38. Brown MB, Forsythe AB. Robust tests for the equality of variances. Journal of the American Statistical Association. 1974;69(346):364-7.
39. Greenacre M, Ayhan HÖ. 2014. BSE Working Paper: 763.
40. Carlisle JB. False individual patient data and zombie randomised controlled trials submitted to Anaesthesia. Anaesthesia. 2021;76(4):472-9.
41. Barton DE, David FN. Multiple runs. Biometrika. 1957;44(1/2):168-78.
42. Wu X, Carlsson M. Detecting data fabrication in clinical trials from cluster analysis perspective. Pharm Stat. 2011;10(3):257-64.
43. 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.
44. 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-British Medical Journal. 2010;340.
45. Woodhead M. 80% of China's clinical trial data are fraudulent, investigation finds. BMJ. 2016;355:i5396.
46. Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One. 2009;4(5):e5738.

Figure legends:

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

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

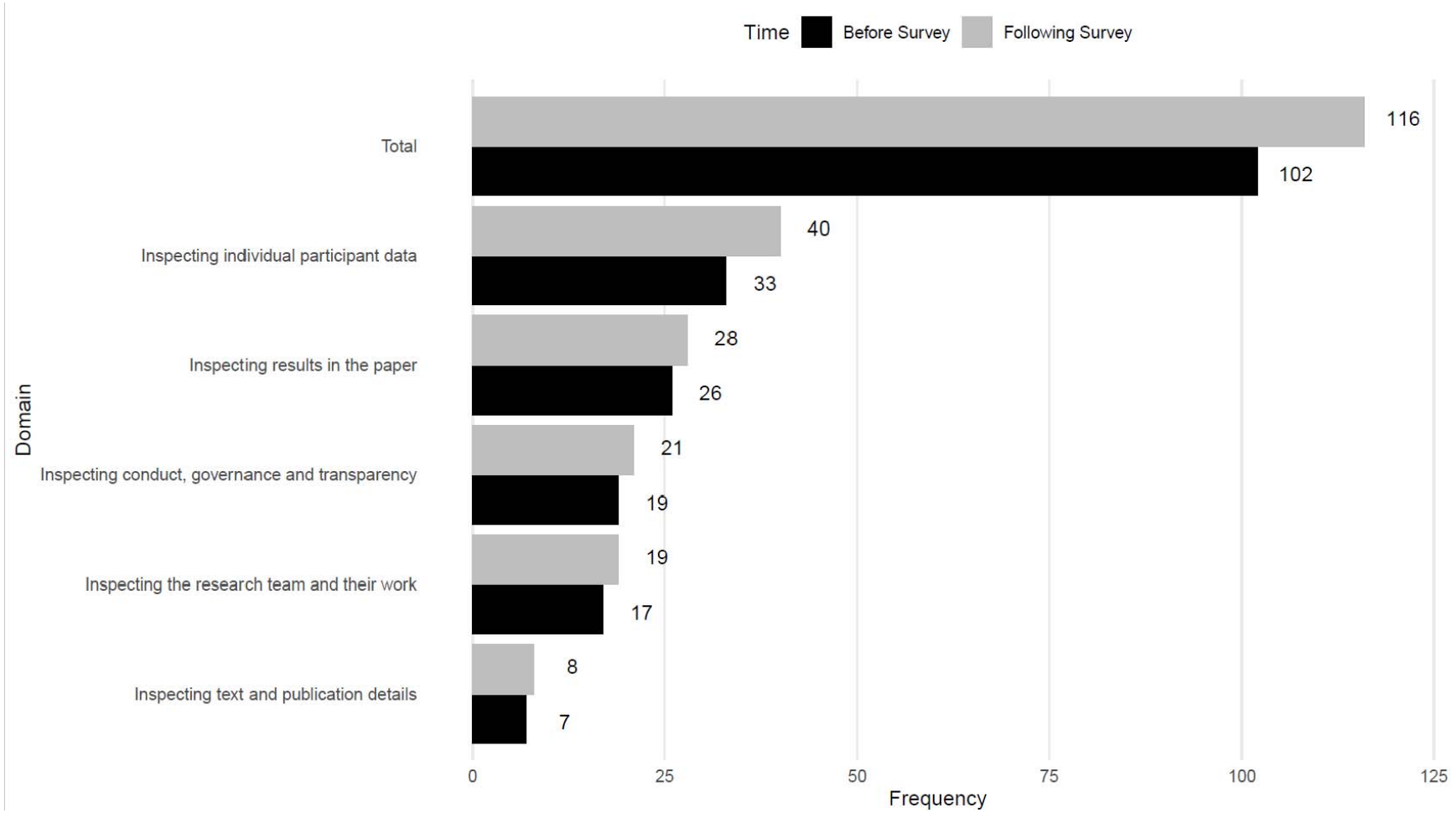


Figure 1: Number of checks in each domain before and after the survey

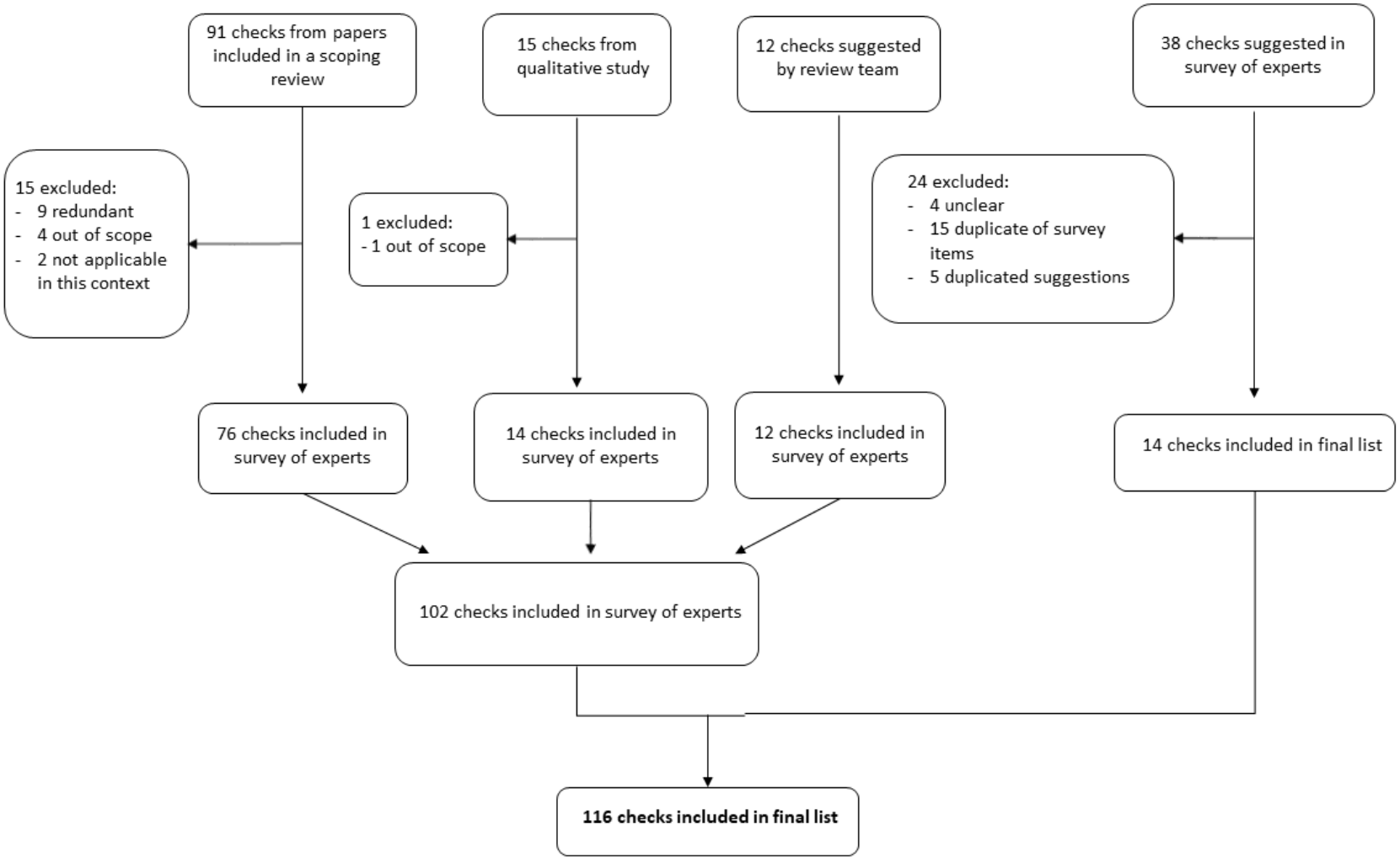
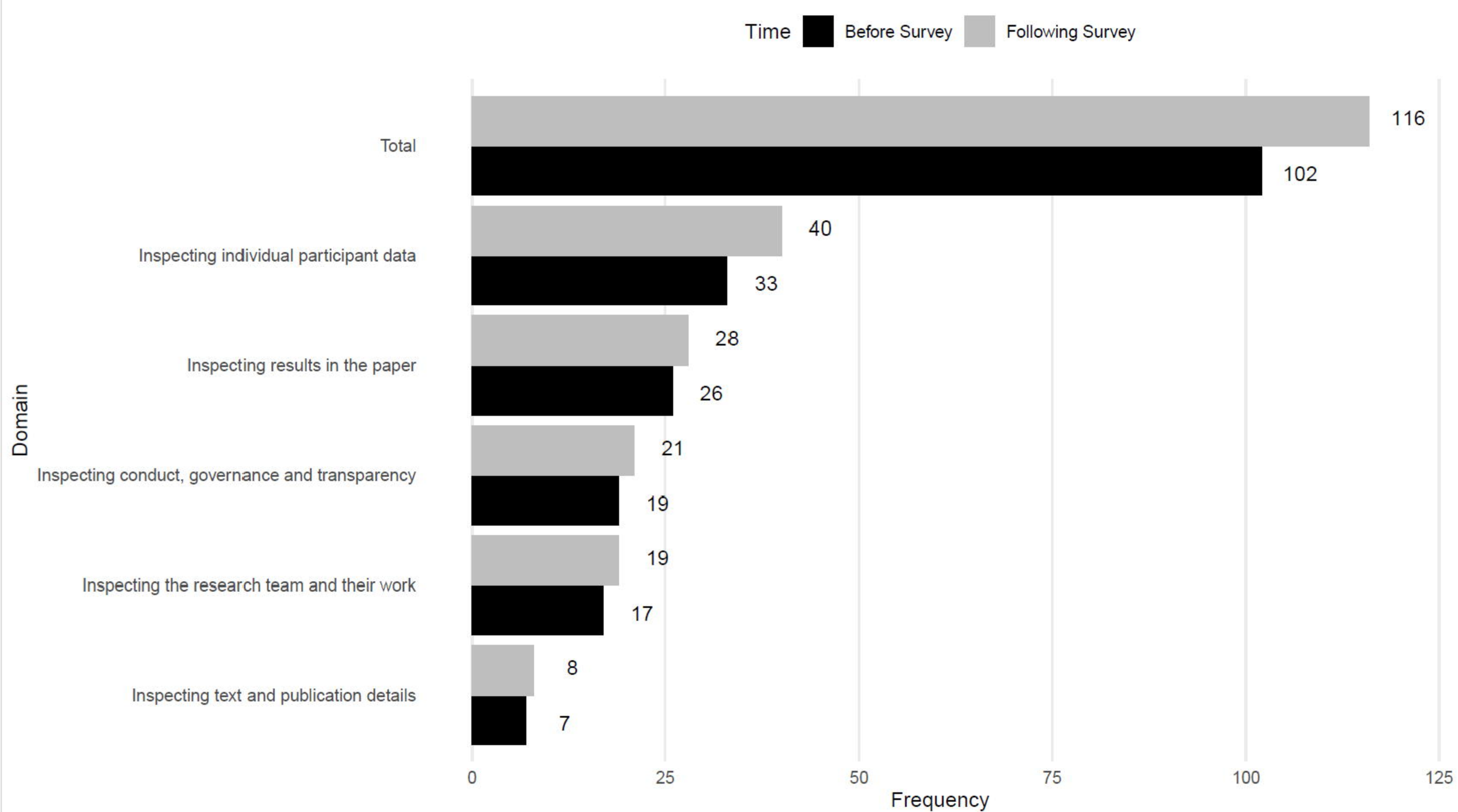
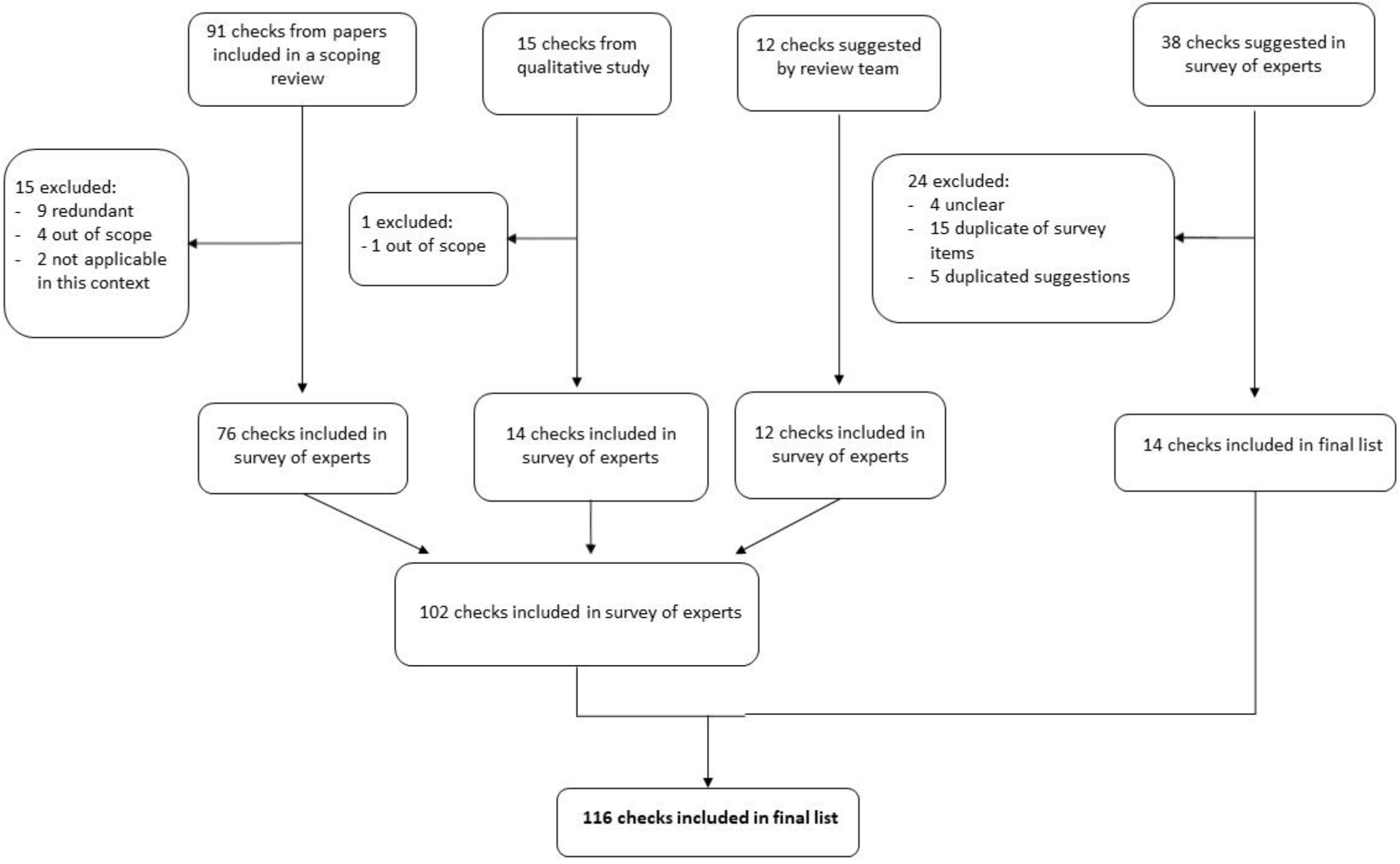


Figure 2: Flow chart showing origin of checks included in final list.


Time
Before Survey
Following Survey
Domain
Total
Inspecting individual participant data
Inspecting results in the paper
Inspecting conduct, governance and transparency
Inspecting the research team and their work
Inspecting text and publication details
116
102
40
33
28
26
21
19
19
17
8
7
0
25
50
75
100
125
Frequency


91 checks from papers included in a scoping review
15 checks from qualitative study
12 checks suggested by review team
38 checks suggested in survey of experts
15 excluded:
- 9 redundant
- 4 out of scope
- 2 not applicable in this context
1 excluded:
- 1 out of scope
24 excluded:
- 4 unclear
- 15 duplicate of survey items
- 5 duplicated suggestions
76 checks included in survey of experts
14 checks included in survey of experts
12 checks included in survey of experts
14 checks included in final list
102 checks included in survey of experts
116 checks included in final list