Journal of Clinical Epidemiology 180 (2025) 111670

Journal of Clinical Epidemiology

METHODOLOGICAL ASPECTS OF RESEARCH INTEGRITY AND CULTURE

How should we assess trustworthiness of randomized controlled trials?

Jack Wilkinson[a,*], David Tovey[b]

[a]*Centre for Biostatistics, Manchester Academic Health Science Centre, Division of Population Health, Health Services Research and Primary Care, University of Manchester, Manchester, UK*

[b]*Journal of Clinical Epidemiology, London, UK*

Accepted 6 January 2025; Published online 13 January 2025

Keywords: Trustworthiness; Fraud; Misconduct; Data fabrication; Falsification; Research integrity

Readers of this journal will be well aware that many randomized controlled trials (RCTs) have serious methodological flaws, which undermine the credibility of their results. In addition, it is increasingly recognized that some trials are afflicted by problems of a different nature; they contain false data or results, and some have been entirely fabricated. These *problematic studies* may describe sound methods [1], which means that they are not flagged by common critical appraisal frameworks, such as risk of bias (RoB) tools [2,3]. Alternative frameworks are, therefore, required to detect these studies so that they can be removed from the literature and prevented from influencing healthcare decisions. These concerns are not hypothetical, as illustrated by the reversal of the National Institute for Health and Care Excellence (NICE) recommendation of Fetal Pillow, which had been made on the basis of potentially problematic studies [4]. In the current issue, JCE presents two contributions to this effort [5,6]. The publication of these articles and the appearance of a spate of tools designed to assess the trustworthiness of RCTs [7−12] are welcome signs that the problem is finally being taken seriously. While there is considerable overlap between many of the proposed approaches, there are important differences in content, form, and implementation. This prompts questions of how we can tell which methods for identifying problematic studies are the most reliable and how they should be implemented. In this commentary, we present some considerations around the development and evaluation of these tools.

1. Trustworthiness assessment as a diagnostic test for fraud

Some researchers have suggested that trustworthiness assessment should be seen as akin to diagnosis of fraud, and that method development should follow principles used for the development of a diagnostic test [13]. Such an approach would necessitate consideration of concepts such as the prevalence of RCTs "with confirmed misconduct" [13], as well as measures of predictive accuracy assessed using a suitable reference standard. Contemplation of the first of these, prevalence, might then cause us to abandon the endeavour altogether. We lack robust estimates of prevalence of fraudulent trials; but if we assume the proportion to be low, then any imperfect diagnostic test risks producing an unacceptably high false-positive rate. Falsely classifying genuine studies as fraudulent harms both the accused researchers and the evidence base, as valuable data are discarded.

Nonetheless, there have been attempts to develop trustworthiness tools according to a diagnostic testing paradigm, for example, by comparing potential trustworthiness assessment criteria between retracted and nonretracted trials and reporting measures of diagnostic accuracy according to the number of criteria passed [14]. There are challenges with this approach. A sound diagnostic test accuracy study requires us to apply the index test to a representative sample

Funding: The INSPECT-SR project is funded by the National Institute for Health and Care Research (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.

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

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

https://doi.org/10.1016/j.jclinepi.2025.111670

from the target population, avoiding case-control designs, and to compare the test result to a reliable reference standard [15]. This might not be achievable in the context of trustworthiness assessment. First, there are limitations of adopting retraction as a reference standard. For example, it can take years for a problematic study to be detected and retracted [16]. Moreover, it is unclear how many of these studies go undetected, and when they are identified, there is no guarantee of retraction. A second problem is that the reference standard must have been applied without reference to the results of the index test. This will almost certainly be violated when considering retractions, since at least some of the trustworthiness criteria being assessed are likely to have been considered in the process of identifying problems with the retracted study and determining its fate.

Use of retraction status may introduce other design issues. For example, in one study, the authors had to use a sample of retracted studies from 1995 onwards to satisfy sample size requirements [14]. This was then compared to a sample of nonretracted studies published between 2018 and 2020, using criteria such as prospective registration, presence of a data sharing statement, and adherence to the CONsolidated Standards Of Reporting Trials (CONSORT) guidelines. We might expect to see considerable differences in these characteristics between trials published in different time periods, regardless of the retraction status. Such a design clearly violates the requirement to have a representative sample.

In light of these considerations, it seems doubtful that a useful diagnostic test for fraudulent RCTs represents a realistic objective. We must either abandon the pursuit of trustworthiness assessment altogether, or jettison the diagnostic testing paradigm.

2. Moving from researcher to research integrity

If reliable diagnosis of fraud is not feasible, how should we proceed with the development of a trustworthiness tool? A distinction between ‘‘researcher integrity’’ and ‘‘research integrity,’’ proposed by MacLeod, et al [17] and discussed by O’Connell et al [18], suggests one way forward. A failure of ‘‘researcher integrity’’ implies deliberate misconduct, including plagiarism, fabrication, and falsification. By contrast, concerns with ‘‘research integrity’’ may stem from researcher misconduct but may alternatively be explained by critical errors in the conduct of various aspects of the trial. We might be able to see that there are numerical contradictions in the results of a published study, but the matter of whether this is attributable to data fabrication or errors in data management or analysis will generally remain unknowable. Noticing an anomaly will usually be much easier than explaining how it occurred. If our goal is to distinguish trustworthy from untrustworthy evidence to prevent the influence of the latter, then identifying the anomaly might suffice. If we can see (or have strong reasons to suspect) that data in a study are incorrect, the question of whether or not this is due to fraud is not the primary concern [9]. Whether due to error or fraud, we should avoid using the results to guide patient care.

By focusing on the trustworthiness of RCTs rather than the question of whether or not the investigators committed fraud, we avoid concerns relating to the unknown prevalence of fraud, while also avoiding legal consequences and harms caused to wrongly accused researchers [18]. Developing a trustworthiness tool can then be viewed as an endeavour to identify and operationalize criteria that would lead us to doubt a study’s veracity. There are parallels to RoB tools, which do not seek to *detect* bias, but rather help the reviewer to judge whether aspects of the study are likely to cause bias. Similarly, a trustworthiness tool should guide the user through an assessment to help them determine whether they have serious concerns about the trustworthiness of the study. The validity of a trustworthiness tool should not depend on the prevalence of untrustworthy studies any more than the validity of RoB tools rests upon the prevalence of biased trials.

3. Selecting suitable criteria for assessing trustworthiness

If measures of diagnostic accuracy are not suitable for the selection of trustworthiness criteria, how can we determine which are worthwhile? Further comparison with the development of RoB tools is instructive. RoB tools have been informed by empirical meta-epidemiological evidence regarding methodological features of RCTs and their association with treatment effect estimates. For example, a comparison of RCTs with and without adequate allocation concealment suggested that the latter group tends to exaggerate treatment effects [19]. It is unclear whether similar approaches would aid in the evaluation of candidate trustworthiness criteria, as it is not known whether bogus results systematically differ from authentic findings. Upon initial consideration, we might expect results in problematic studies to be inflated. However, this would not be the case if a fabricator opted to emulate existing RCTs or report a negative result to avoid detection. Nor would a lack of systematic differences reassure us that problematic studies do not cause harm, as this would tell us little about the impact of these trials on particular instances. Nonetheless, other forms of empirical evidence remain important. For example, many useful studies by Bolland et al [20,21] have evaluated data patterns in fraudulent clinical trials, which advance our understanding of statistical approaches to fraud detection [22].

Of course, items are not selected for inclusion in RoB tools solely on the basis of empirical evidence. For example, the MetaBLIND study failed to demonstrate any association between blinding and trial results [23]. It does

not follow that the assessment of blinding in individual RCTs should be discontinued. There are clear theoretical principles explaining how and why lack of blinding might impact study results. Rather, the MetaBLIND result might discourage the sort of algorithmic thinking that declares any study without blinding to be at high RoB, instead of encouraging careful consideration about the potential for lack of blinding to cause bias in any particular trial. This is exactly the approach encouraged by the RoB 2 tool [3]. Similarly, it seems likely that theoretical considerations have a role to play in the selection, elaboration, and application of trustworthiness criteria.

Corroborating this idea, researchers, to date, have relied upon theory and expert opinion for the selection and evaluation of trustworthiness criteria, with the majority selecting criteria on the basis of the experience of and discussions in the research team. For example, the developers of Trustworthiness in RAndomized Controlled Trials (TRACT) devised a preliminary list of checks based on their own experience before asking a Delphi panel to apply the checks to several RCTs and score these in terms of their usefulness and feasibility [7].

If we are to select trustworthiness criteria on the basis of theoretical considerations, they should, at the very least, display face validity. In this regard, some of the checks included in existing tools may be challenged. For example, in the current issue, Au et al use a discrepancy of 15% or more between the intended and achieved sample size as a marker of untrustworthiness [6]. This may reflect the professional experiences of the trustworthiness in randomized controlled trials development team, who largely represent the field of Obstetrics and Gynaecology. Nonetheless, it may be viewed with skepticism by trialists in many fields where recruitment challenges are common. Indeed, the problem of underrecruitment to RCTs is so well recognized that strategies to improve recruitment are a focus of trial methodology research [24]. In addition, the intent behind the TRACT domain *Plausibility of Intervention Usage* is not entirely clear, as it appears to conflate allocation concealment with blinding, prompting the user with the example '*use of sealed envelopes in a placebo-controlled trial*'. It is not clear why the use of sealed envelopes to prospectively conceal the allocation sequence should preclude the use of placebo to ensure blinding throughout the trial. The study by Au et al introduces further confusion, as the authors prompted ChatGPT with the text ''Specifically, evaluate if the explanation of the interventions and control/placebo is detailed enough to allow for replication in another experiment.'' This suggests that the primary concern is to ensure the methods are clearly reported. This critique is not intended to denigrate the pioneering work of the team so much as to illustrate that potential trustworthiness criteria can be criticized using the methodological principles. The fact that we can discuss individual checks in a principled way suggests that the selection of criteria on the basis of theory is not an arbitrary endeavour.

4. Implementation

Having selected suitable criteria to use for the assessment of RCTs, we must consider how to operationalize each of these in the form of a check that can be performed by reviewers and how to assemble and arrange these checks to produce a useful and practicable trustworthiness assessment tool. In doing so, we must be mindful that the idea of routinely examining trustworthiness of RCTs is relatively new; so, the majority of potential users of a tool will have limited experience and expertise in the assessment of trial integrity. There are considerable risks here. The prospect of taking on the role of data detective might be alluring to many principled researchers, and while the desire to eliminate bogus research from the literature is admirable, the combination of enthusiasm and inexperience may produce a torrent of spurious complaints, needlessly burdening accused researchers and research integrity professionals and bringing the whole enterprise of trustworthiness assessment into disrepute. To minimize these risks, the wording of these trustworthiness checks requires careful consideration, and clear guidance is needed for each, detailing the ways in which they might malfunction.

A cautionary example can be found in the article by Nielsen et al [5]. The authors recommend recalculating *P* values from reported summary statistics. The authors correctly note that *P* values may not be exactly reproducible from rounded summary data and suggest that ''only large differences between the reported and recalculated *P* values should be concerning''. But as guidance to users, this might be highly misleading; recalculated *P* values might be very different from reported *P* values but consistent with the reported summary data. We are familiar with an example in which identical mean values ('no difference') in the study groups would be compatible with highly significant *P* values, once rounding has been taken into account. It is appropriate to recommend checking the consistency of statistical results with reported data, but this needs to be accompanied by clear guidance, describing how to check whether reported *P* values are consistent with reported summary data, taking rounding into account. Where t-tests have been used, it is not difficult to calculate the largest and smallest *P* values that would be compatible with the reported summary data [25]. Indeed, it might be more suspicious if we can reproduce the authors' *P* values from the reported summary data in every line of a table of results, as it might suggest that no underlying dataset was analyzed to obtain the results. We expect that these points are well-understood by Neilsen, et al, but suitable training and guidance are required to make sure that this knowledge is appropriately transferred to users of these methods.

The matter of how to arrange the checks in the form of a useful and reliable tool also requires care. In particular, there is the question of how responses to a barrage of checks should be used to arrive at an overall rating about the trustworthiness of a study. In the current issue, Au et al. suggest that a study should be flagged for investigation if most items are rated as ''major concerns'' [6]. Anderson et al treated the number of quality criteria achieved by an RCT as a score, reporting discriminative performance according to each cut-off, and suggested that a ''predictive scoring system'' could be developed [14]. Again, there might be important lessons to learn from the RoB literature. Quality scores, such as the Newcastle-Ottawa Scale, may be criticized for implying that a critical flaw in a study can be compensated by other strengths. For example, using a quality score, a study with serious uncontrolled confounding might be rated as good-quality evidence, provided other aspects were done well [26]. This phenomenon should be avoided by a reliable trustworthiness tool. If a study's reported summary data are incompatible with the reported statistical results, for example, we shouldn't trust those results even if no other red flags are present. One way to ensure that a trustworthiness tool reflects that critical flaws would be to adopt the approach used in RoB 2, where the overall study-level judgement is at least as severe as the judgement for the lowest rated domain. Problematic studies could display any of a variety of warning signs, but there is no particular reason to expect any particular problematic study to possess a large number of them.

5. Automation of trustworthiness assessment

The introduction of trustworthiness tools as a supplement to existing methodological appraisals is likely to increase the burden of assessment. There is a clear rationale for deploying trustworthiness tools prior to RoB assessment. After all, why should we care about the apparent validity of an inauthentic study? This would preclude the RoB assessment for problematic trials. However, if we assume that a majority of RCTs will not be judged as being problematic and will therefore be subjected not only to trustworthiness but also RoB assessment, the overall time required to assess trials is likely to increase. Both in the current issue [5] and elsewhere [22], it has been highlighted that concurrent examination of an author's full body of work is an effective way to identify integrity issues, but assessing a large number of trials at once is particularly onerous. Solutions to increase feasibility are therefore needed.

To this end, Au et al have demonstrated reasonable performance of a large language model (LLM) in the implementation of TRACT in a case study [6]. Reliable, automated trustworthiness assessment would be gold dust to journal editors, research integrity specialists, and systematic reviewers. Some checks are likely to be easier to automate using AI than others. For example, checks of unambiguous, clearly defined study features, such as checking whether the sample size discrepancy exceeds 15%, might be easier to assess using AI than those requiring subject matter knowledge and a degree of expert judgement. Unfortunately, these 'objective' checks are also those subjects to concerns relating to arbitrariness and generalisability. We have already expressed doubts about the 15% sample size threshold. Determining whether or not such a discrepancy is really a cause for concern and would require understanding of the particular context in which the study was taking place. Similarly, as correctly noted in the wording of TRACT, low numbers of participants who lost to follow-up would be more concerning in some trials than in others [7]. Specifically, low attrition rates might be more surprising in studies with longer duration of follow-up or more demanding treatment and outcome measurement protocols. Again, assessment of this item cannot be reduced to a check against an objective criterion. Expert understanding is required. Guidance for RoB 2 emphasizes the importance of subject matter and methods expertise in RoB assessment, and we anticipate that expert judgement will also be an essential feature of many trustworthiness checks. On the face of things, objectivity might appear to be a desirable quality of a trustworthiness check, but if this entails the application of arbitrary standards, it may be a double-edged sword. A further problem with arbitrary, objective criteria is that they are easily gameable by fraudsters, who can manipulate features of the manuscript in accordance with these rules. On the other hand, if we introduce checks that require careful consideration in their application, we may raise the level of difficulty for both users of the checks and people trying to circumvent them.

The study by Au et al highlights some other ways in which AI might facilitate trustworthiness assessment. For example, statistical forensic methods typically require summary data from the paper to be extracted and formatted for further analysis, which is time-consuming if done manually. The ability to use LLMs to automate this task would represent a considerable saving of time. If statistical checks of these data could also be reliably automated, this would lower the barrier to entry for many users of trustworthiness tools, while potentially reducing errors in the application and interpretation of these methods.

One potential risk of using LLMs in the peer review process is that it could, in principle, lead to breaches of confidentiality. These concerns could potentially be addressed through the development of bespoke LLMs deployed in secure environments for use in editorial assessment.

6. Concluding remarks

We have described some methodological considerations for the development of trustworthiness tools. In doing so, it could be argued that we are making a mountain out of a molehill. Perhaps, the exact content and form of these tools

are not important. Rather, it could be argued, the primary objective is to ensure that some forms of trustworthiness assessment of RCTs are routinely performed, as this will deter fraudsters who might otherwise have proceeded to submit fake research to journals in the belief that no one would ever think to question its authenticity. This is indeed an important anticipated consequence of routine trustworthiness assessment and hopefully offers some reassurance against concerns that these tools will serve as training manuals for LLMs and individual fraudsters on how to produce convincing forgeries. However, we believe that it is nonetheless important to construct these tools with due regard to their validity. Efforts to develop a trustworthiness tool, INSPECT-SR, according to the principles described here, are almost completed at the time of writing [27].

The preceding discussion has been concerned with trustworthiness assessment without recourse to the underlying trial dataset. When the individual participant data (IPD) can be accessed, a more rigorous assessment of study integrity is possible. A tool for use when IPD are available has recently been published [11], and a large consensus process to develop an IPD extension to INSPECT-SR (working name, INSPECT-IPD) has been funded. The problem of problematic studies is clear, and various tools for thwarting their influence have been proposed. Stakeholders must now decide which of these tools are viable. Burying one's head in the sand is no longer a defensible option.

CRediT authorship contribution statement

Jack Wilkinson: Conceptualization, Writing − original draft, Writing − review & editing. **David Tovey:** Conceptualization, Writing − review & editing.

Declaration of competing interest

J.W. is primary investigator on the INSPECT-SR project (NIHR203568) and primary supervisor of the INSPECT-IPD project (NIHR303046). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. J.W. additionally declares Statistics or Methodological Editor roles for BJOG, Fertility and Sterility, Reproduction and Fertility, Journal of Hypertension, and for Cochrane Gynecology and Fertility. He also undertakes confidential integrity investigations for various journals and publishers.

Data availability

No data were used for the research described in the article.

References

[1] 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.
[2] 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.
[3] Sterne JAC, Savovic J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898.
[4] Grey A, Avenell A, Bolland MJ, Thornton JG. The fetal Pillow deflates-lessons for all. BJOG 2024.
[5] Nielsen J, Bordewijk EM, Gurrin L, et al. Assessing the scientific integrity of the collected work of one author or author-group. J Clin Epidemiol January 2025. https://doi.org/10.1016/j.jclinepi.2024.111603.
[6] Au LS, Qu L, Nielsen J, et al. Using artificial intelligence to semi-automate trustworthiness assessment of randomized controlled trials: A case study. J Clin Epidemiol January 2025. https://doi.org/10.1016/j.jclinepi.2025.111672.
[7] Mol BW, Lai S, Rahim A, Bordewijk EM, Wang R, van Eekelen R, et al. Checklist to assess Trustworthiness in RAndomised Controlled Trials (TRACT checklist): concept proposal and pilot. Res Integr Peer Rev 2023;8(1):6.
[8] Weibel S, Popp M, Reis S, Skoetz N, Garner P, Sydenham E. Identifying and managing problematic trials: a research integrity assessment tool for randomized controlled trials in evidence synthesis. Res Synth Methods 2023;14(3):357−69.
[9] Grey A, Bolland MJ, Avenell A, Klein AA, Gunsalus CK. Check for publication integrity before misconduct. Nature 2020;577(7789):167−9.
[10] Weeks J, Cuthbert A, Alfirevic Z. Trustworthiness assessment as an inclusion criterion for systematic reviews—what is the impact on results? Cochrane Evid Synth Methods 2023;1(10):e12037.
[11] Hunter KE, Aberoumand M, Libesman S, Sotiropoulos JX, Williams JG, Aagerup J, et al. The individual participant data integrity tool for assessing the integrity of randomised trials. Res Synth Methods 2024;15(6):917−35.
[12] Abbott J, Acharya G, Aviram A, Barnhart K, Berghella V, Bradley CS, et al. Trustworthiness criteria for meta-analyses of randomized controlled studies: OBGYN journal guidelines. J Obstet Gynecol MFM 2024;6(12):101481.
[13] Khan KS, Fawzy M, Chien PFW. Integrity of randomized clinical trials: performance of integrity tests and checklists requires assessment. Int J Gynaecol Obstet 2023;163(3):733−43.
[14] Anderson KM, Doulaveris G, Bennett C, Mol BW, Berghella V. Standard quality criteria in retracted vs nonretracted obstetrical randomized controlled trials. Am J Obstet Gynecol MFM 2023;5(5):100889.
[15] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529−36.
[16] Chambers LM, Michener CM, Falcone T. Plagiarism and data falsification are the most common reasons for retracted publications in obstetrics and gynaecology. BJOG 2019;126(9):1134−40.
[17] Macleod M, University of Edinburgh Research Strategy G. Improving the reproducibility and integrity of research: what can different stakeholders contribute? BMC Res Notes 2022;15(1):146.
[18] O'Connell N, Richards GC, Soliman N, Ferraro MC, Segelcke D, Eccleston C, et al. ENTRUST-PE: An Integrated Framework for Trustworthy Pain Evidence. White Paper 2024. https://doi.org/10.31219/osf.io/e39ys.
[19] Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273(5):408−12.

[20] Bolland MJ, Gamble GD, Avenell A, Cooper DJ, Grey A. Distributions of baseline categorical variables were different from the expected distributions in randomized trials with integrity concerns. J Clin Epidemiol 2023;154:117−24.

[21] 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.

[22] Bolland MJ, Avenell A, Grey A. Statistical techniques to assess publication integrity in groups of randomized trials: a narrative review. J Clin Epidemiol 2024;170:111365.

[23] Moustgaard H, Clayton GL, Jones HE, Boutron I, Jorgensen L, Laursen DRT, et al. Impact of blinding on estimated treatment effects in randomised clinical trials: meta-epidemiological study. BMJ 2020; 368:l6802.

[24] Boxall C, Treweek S, Gillies K. Studies within a trial priorities to improve the evidence to inform recruitment and retention practice in clinical trials. Res Methods Med Health Sci 2022; 3(4):121−6.

[25] Brown NJ, Heathers J. Rounded Input Variables, Exact Test Statistics (RIVETS) 2024. https://doi.org/10.17605/OSF.IO/UBWM3.

[26] Wilkinson J, Stocking K. Some common, fatal flaws in systematic reviews of observational studies. Fertil Steril 2024;121(6): 918−20.

[27] Wilkinson J, Heal C, Antoniou GA, Flemyng E, Avenell A, Barbour V, et al. A survey of experts to identify methods to detect problematic studies: stage 1 of the INveStigating ProblEmatic Clinical Trials in Systematic Reviews project. J Clin Epidemiol 2024; 175:111512.