A checklist to assess Trustworthiness in RAnomalised Controlled Trials (TRACT checklist)

Ben W Mol (ben.mol@monash.edu)
Monash University

Shimona Lai
Monash University

Ayesha Rahim
Monash University

Esmée M Bordewijk
Amsterdam UMC, location AMC

Rui Wang
Monash University

Rik van Eekelen
Amsterdam UMC, location AMC

Lyle C Gurrin
The University of Melbourne

Jim G Thornton
University of Nottingham

Madelon van Wely
Amsterdam UMC, location AMC

Wentao Li
Monash University

Research Article

Keywords: research integrity, trustworthiness, randomised controlled trials, checklist

Posted Date: January 11th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2119486/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Objectives: To develop a checklist to screen, for trustworthiness, papers reporting the results of randomised controlled trials (RCTs).

Design: A screening tool was developed using the four-stage approach proposed by Moher et al. This included defining the scope, reviewing the evidence base, suggesting a list of items from piloting, and holding a consensus meeting as part of a Delphi method.

The initial checklist was set-up by a core group who had been involved in the assessment of dubious RCTs for several years. We piloted this in a Delphi panel of several stakeholders, including health professionals, reviewers, journal editors, policymakers, researchers and evidence-synthesis specialists. Each member was asked to score three articles with the checklist and the results were then discussed in two Delphi sessions.

Results: The Trustworthiness in RAndomised Clinical Trials (TRACT) checklist includes seven domains that are applicable to every RCT: governance, author group, plausibility of intervention usage, timeframe, drop-out rates, baseline characteristics and outcomes. Each domain contains two or three signalling questions that can be answered as either no concerns, some concerns/no information, or major concerns. If a study is assessed and found to have significant concerns, then editors or reviewers should consider a more thorough investigation, including assessment of original individual participant data.

Conclusions: The TRACT checklist is the first checklist developed in a formal process to detect trustworthiness issues in RCTs. It might help editors, publishers and researchers to screen for such issues in submitted or published RCTs in a transparent and replicable manner.

Introduction

The US Office of Research Integrity (ORI) defines research misconduct as “fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results. Research misconduct does not include unintended error or differences of opinion”.¹

The prevalence of scientists who have been involved in scientific misconduct is estimated to be around 2%² and, for randomized clinical trials (RCTs), this percentage may be higher. Carlisle et al. reported recently that, based on assessment of review of individual participant data, 44% of submitted randomised controlled trials to the journal “Anaesthesia” were based on false data, with 26% of studies being entirely fabricated³. Although these rates might be over-estimated, the corresponding estimates of false and fabricated trials run into the hundreds of thousands, so even a prevalence one hundred times lower would still point to thousands of “zombie trials”⁴.

RCTs are recognized as the most reliable type of scientific investigation when assessing the effects of interventions. The results of RCTs are usually summarised in systematic reviews and meta-analyses, and
subsequently used to support clinical recommendations in guidelines. This process is intended to improve the overall effectiveness and efficiency of medical interventions by providing trustworthy evidence. As RCTs are used to inform clinical guidelines, they directly influence patient care. If, however, some of these trials report a treatment effect that is based on false or fabricated data then it has the potential to adversely affect the health of patients. Due to this influence, the trustworthiness of RCTs is potentially more important than that of other types of medical research.

RCTs have specific governance characteristics (including prospective trial registration requirements), detailed protocols (including documentation of study medication), and checklists that have to be completed at submission to journals. RCTs also have characteristics, absent in other forms of studies, that lend themselves to direct assessment of trustworthiness problems. For example, there should be comparable arms of trial participants and predictable patterns of differences in baseline variables between arms which, in some notorious example of scientific misconduct, have been very difficult to emulate. Deviation from these unique characteristics may provide strong evidence of data manipulation in RCTs.

In view of the importance of RCTs for clinical practice, and in light of their specific governance, methodological and submission requirements, there is both a need and an opportunity to develop tools that can be used to reliably assess, test and measure research integrity within and across submitted and already published RCTs whose trustworthiness has been questioned. During our efforts to assess the trustworthiness of RCTs we observed patterns, incorporated existing methods, and developed a pathway to assess potential problems. Based on our experience and a subsequent systematic review that assessed current methods of assessing research misconduct, we developed a prototype checklist to screen for trustworthiness issues due to possible scientific misconduct in RCTs. This screening tool (Trustworthiness in RAndomised Clinical Trials [TRACT]) needs to be optimised and formalised so as to be widely used by reviewers and editors, as well as by those who perform systematic reviews.

**Methods**

Development of our screening tool was adapted from the four-stage approach proposed by Moher and colleagues: define the scope, review the evidence base, suggest a list of items from piloting and hold a consensus process. For the consensus process, we utilised the Delphi technique, where stakeholders were asked to evaluate versions of the TRACT checklist and used questionnaires to help determine the final items to be included within.

**Defining the Scope**

We established a steering group of experts in evidence synthesis, with a background as clinical researchers, reviewers, editors, biostatisticians or evidence-synthesis specialists. The group was formed from colleagues who had been collaborating in the assessment of trials whose research integrity had...
been questioned, and also included journal editors and editors of specialist groups in Cochrane. Together this group agreed on the scope of the tools.

The primary decision of this steering group was to define the use of the screening tool. The screening tool would be incorporated in as a quick checklist that aims to review articles and triage them according to the risk of the results being based on fabricated data.

The steering group agreed to limit the screening tool to RCTs. The justification for this was that, while research misconduct is harmful for all types of research, RCTs are a crucial step in the assessment of medical interventions, prior to these interventions being applied in clinical practice and trustworthiness issues at this level stand to create direct harm for patients if their conclusions are not legitimate.

### Reviewing the Evidence Base

We conducted a scoping review of the literature for studies that mentioned a method for screening for or assessing and quantifying the extent of data integrity issues in health-related papers. We used this review of all available methods to complete our TRACT checklist\(^{14}\). Moreover, we conducted multiple systematic analyses (of which four have been published) to inform the development of the screening checklist\(^{10-13}\). The findings from these systematic reviews and our experience led to a first version of the checklist.

### Piloting

Over the last few years, we have assessed the integrity of > 300 RCTs in women's health for a variety of reasons. In these assessments, done by junior academics and students from several countries, the checklist identified trustworthiness issues in several RCTs that were later confirmed in formal investigations led by journals or publishers. The feedback from users helped us to further refine the checklist. EMB, WL, and BWM adjusted the draft of the checklists that was used for multiple rounds of piloting before presenting it to the Delphi panel.

### Delphi Consensus Process

We held a Delphi consensus meeting via videoconference with a panel with a range of stake-holders, including health professionals, reviewers, journal editors, policymakers, researchers and evidence-synthesis specialists. These stakeholders were invited to use the checklist draft to screen up to three articles and were then asked to complete a short questionnaire regarding their experience. For each checklist item, stakeholders were asked to rate both how useful it was in detecting possible breaches of data integrity and how easy it was to assess on a five-point Likert scale from 'Not Useful/Easy' to 'Extremely Useful/Easy'. These selections were then allocated points (from 1 to 5) based on their rating (e.g. Not Useful = 1 and Extremely Useful = 5; Not Easy = 1 and Very Easy = 5). Therefore, a higher rating would mean more usefulness or easier for each item. The questionnaire also provided a free text section where stakeholders were also able to provide detailed feedback regarding their use of each domain as well as advise on any additional items they thought should be included in the checklist.
Based on their use of the checklist and the completion of the questionnaire, we compiled a feedback summary and calculated median aggregate scores for usefulness and ease. We regarded items with a median rating of 3 or more in both usefulness and ease as relevant and included these in our final checklist version. We discussed stakeholder feedback and adjusted the content, applicability and design of the screening checklist based on group consensus. On the basis of the agreed outcomes of the meeting, we adapted the draft and completed the final version of the checklist.

Results

The TRACT checklist to screen for trustworthiness

The TRACT checklist is sectioned into seven domains that are applicable to every RCT; governance, author group, plausibility of intervention usage, timeframe, drop-out rates, baseline characteristics, and outcomes. These domains are discussed in further detail below.

Governance

For each RCT, it is important to check registration of the study in an official trial registry (i.e. ICRT Registry Network, clinicaltrials.gov). Registration should have occurred before start of the study, and at least not later than 2 weeks after the start of participant recruitment. The absence of even retrospective registration is a cause for concern. One should look for mismatch in the planned number of participants in the trial registration, and the number participants actually recruited in the study. If there is a mismatch, there should be an explanation in the final report. Changes in the design and conduct of the trial registration should appear in the registration portal and can be tracked using online audit tools. The description of research ethics should be clear and preferably there the ethics committee should be named, including a reference number of the ethics application. There should be a clear description of the process of informed consent as part of checking for ethical concerns with the study design.

Author Group

Aspects of author group may reflect the integrity of the study. RCTs with three or fewer authors, have a higher risk of trustworthiness issues, especially if they are multicentre RCTs. Normally, larger multicentre studies would usually be conducted in collaboration with a clinical trials unit, and have (at least) statisticians, epidemiologists and medical scientists as co-authors. Retractions of other studies by one or more members of the author group may also arouse suspicion, especially if this retraction is not requested by the author(s). A search of the authors in Retraction Watch Database or PubMed can identify earlier retracted studies from an author or their institution.

A large amount of RCTs published in a short timeframe by one trialist as first or last author or in one institute (e.g. >3 per year as first author) should also be cause for concern. A search of the trial registers and databases of published papers and reports according to author names can provide an idea of the total number of studies that is performed during a timeframe. This is especially suspicious if an author is
noted to be recruiting participants for their studies from a single institution; and on the other hand, concern may be somewhat mitigated if an author is performing trials at multiple different institutions or countries.

**Plausibility of Intervention Usage**

The plausibility of a study can be assessed via its design. An example of implausible study design is the use of one placebo when there are two active interventions administered via different routes in a three-arm study. Studies that also describe their allocation concealment process either poorly (or not at all) should also raise concern as this detail provides validity and reliability to the study's findings. Consider if the methods allow replication: whether the interventions and control/placebo are explained sufficiently well enough to allow the process to be replicated in another experiment. In a similar vein, the description of the study design, and research methodology and the subsequent statistical analyses should be relevant and appropriate to the project being undertaken. In this domain, an understanding of clinical research methodology and biostatistics can aid in identifying trustworthiness issues.

**Timeframe**

The study timeline is another important aspect to consider when appraising RCTs. The recruitment rate and the time between the end of recruitment to the submission of the manuscript have been identified by our study group as critical timeframes that require consideration, with timeframes that appear implausibly short being a point for concern.

In considering the speed of recruitment it is important to factor in the prevalence and incidence of the disease and the capacity of any recruiting centres, which usually can be obtained via an internet search. Two or more RCTs on the same topic executed simultaneously in one centre by the same author would, for example, be reason for concern. Additionally, the time needed to complete follow up is an important consideration for chronic diseases and interventions that require an extended time for assessing outcomes, such as assisted reproductive technology.

**Drop-out Rates**

Studies in which no participants were lost to follow up or for which no reasons are given for loss to follow up should be assessed with care, while taking into account the size of the study and the likely difficulty of follow up. There is often an expected level or rate of drop-out from RCT participants, especially when you consider long-term studies requiring prolonged and active engagement from participants that may prompt greater non-compliance (e.g. a questionnaire administered to participants at several points in time with large number of items); therefore lack of missing data points should be regarded as suspicious. On the other hand, a RCT study that has one point of data collection and a relatively simple technique to record of outcomes with no adequate explanation of drop-out, is similarly suspicious.

A similar drop-out rate for different treatments arms is often seen as a reassurance against bias; however, Bell et al. determined that equal number of drop-outs rates do not guarantee unbiased results in RCTs\(^\text{18}\). Still, some RCTs may fabricate their drop-out rates to appear more similar and less suspicious overall.
Therefore, perfectly balanced drop-out rates in round numbers in particular (i.e. two groups of 50, two groups of 100, etc.) with no explanation for such balance should raise concern, especially if this implies that a larger group of participants were ultimately enrolled than were initially recruited.

**Baseline Characteristics**

Baseline characteristics provide important demographic and medical information on the target population of the RCT, and whether or not balance of baseline characteristics was achieved to the extent that would be expected from the randomised allocation of interventions. A lack of detail or the absence of baseline characteristics in a paper prevents the ability of others to assess whether groups apparently generated by randomisation have been appropriately balanced. Additionally, this domain may also reveal implausible patient characteristics when considering common sense, previously reported results, and local data (e.g. standard deviations are similar for completely different characteristics with different means or distributions). We also note that baseline characteristics in near-perfect balance is unrealistic if allocation was determined via randomised allocation of a sample of participants from a real population. Studies relating to specific research fields or medical conditions should highlight important and pertinent prognostic factors at baseline, and the absence of summary reporting of these should be flagged. Differences in baseline characteristics from different studies on the same topic, or differences in uniform characteristics such as body length are also reason for concern.

**Outcomes**

When assessing study outcomes, effect sizes should be taken into consideration. If RCTs of the same topic with similar patient demographics indicating that effect size should be similar is instead substantially different, this should raise concern. Assessing the heterogeneity of studies is commonly used in meta-analyses and provides detail on the variability of included studies. Conflicting information between outcomes should also be assessed – for example, if a study has less clinical pregnancies (pregnancy confirmed via the presence of a gestation sac on ultrasound) compared to ongoing pregnancies (often defined as the presence of a viable pregnancy with cardiac activity noted from at least 12 weeks gestation), which should be impossible. It is also important to consider the primary outcome detailed at trial registration (if applicable) and check if this has now changed at publication\(^\text{19}\). All of these domains combined provide an indication of the trustworthiness of the published RCT.

**Discussion**

In this article, we have presented a screening checklist tool to detect publication trustworthiness issues in RCTs. This TRACT checklist is designed to screen papers and triage their risk of data fabrication to allow for better detection of research misconduct.

Our checklist is straightforward, easy to apply, and enables the analysis of research misconduct without any significant prior experience or training. If a study is assessed and found to be suspicious, then
reviewers should consider continuing with a more thorough investigation into the data integrity issues identified, in which we recommend assessment of the original data.

The TRACT checklist also has its limitations. Firstly, although we developed this checklist, it is not yet validated. Secondly, some patterns or reasons for concern are somewhat crude: other patterns of misconduct may not be picked up on by using this checklist. Fabricators can use the checklist to fabricate a paper that fits all the items of the checklist. As a screening tool, it may misclassify papers that either do or do not warrant further investigation. Lastly, using our checklist can be time consuming depending on the article being screened and the capacity and experience of users.

We have not provided a formal cut-off at which our checklist scores positive. This will be addressed in subsequent validity studies. Also, trustworthiness screening is an issue of common sense. Until now, the question ‘Are the data from a real study?’ and ‘Did the study actually take place “in real life”?’ have simply not been asked in assessment of submitted articles or during meta-analysis. Awareness of data-fabrication, even without using a formal score, will already make a large difference.

This is the first checklist developed for the detection of trustworthiness issues specific to RCTs. We believe the TRACT checklist can help editors, publishers and researchers who suspect scientific misconduct to make an efficient analysis of RCTs. If using this checklist raises suspicions or even provide evidence for research misconduct, the authors should be asked for an explanation. If they cannot provide a satisfactory or reasonable explanation, the next step for published articles is to consider investigation according to the Committee on Publication Ethics (COPE) guidelines. For researchers and readers who performed this assessment, they should contact the journals in which the papers are published.

We strongly believe that (anonymised) data sharing should become the standard before a paper can be accepted for publication, and that raw data should be publicly available at the moment of publication. Science is often thought of as a self-correcting system in which hypotheses and data are constantly being tested, replicated and validated, which is only possible when data is shared. The availability of participant data and the willingness to share these data may be a good indicator of quality, methodological soundness and integrity of RCTs. The burden of proof of integrity of a paper should be with the authors and not with the editors or peer reviewers.
Ultimately, the safety of patients was our primary concern and our motivation to develop this checklist. Providing medical treatment based on fraudulent research could be harmful, even if proposed treatment does nothing at all. Research misconduct is a major problem that all fields are facing, and the only way to tackle the problem is for all stakeholders take shared responsibilities for improving the system maintaining a high standard of research integrity and for the publication and dissemination of the results of research that has been shown to meets these standards.

**Declarations**

**Authors’ roles**

BWM conceived the idea. BWM, WL, RW, EB, MvW, LCG, RvE and JT designed the study. BWM drafted the manuscript. SL and AR organised the Delphi process and the consensus meetings.

We acknowledge the members of the Delphi panel: Zarko Alfirevic, Rik van Eekelen, Vincenzo Berghella, Jack Wilkinson, Madeline Flanagan, Steve McDonald, Helen Nagels, Lyle Gurrin, Francis Githae Muriithi, Kurt Barnhart, Esmée Bordewijk, and Naomi Lee.

**Funding**

No external funding was sought to support this work.

**Conflict of interest**

MvW is co-editor of the Netherlands Satellite of the Cochrane Gynaecology and Fertility Group.

BWM is supported by a National Health Medical Research Council (NHMRC) Practitioner Fellowship (GNT1082548). BWM reports consultancy for ObsEva, Merck and Guerbet. All other authors have no conflicts of interest.

**References**


21. Bordewijk EM, et al., Wang R, van Wely M et al To share or not to share data: how valid are trials evaluating first-line ovulation induction for polycystic ovary syndrome?. Human reproduction update 2020;Accepted for publication.

Tables

Table 1 is available in the supplementary files section.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- IntegrityTRACTTable.docx