Developing a tool for detecting problematic RCTs in health systematic reviews: the INSPECT-SR project

Jack Wilkinson, Centre for Biostatistics, University of Manchester. 🐦 @jd_wilko

Steering Group: Calvin Heal, Georgios Antoniou, Stephanie Boughton, Lisa Bero, Jamie Kirkham.

Some of the research discussed in this presentation 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.
Disclosures

• Currently hold or have recently held grants from NIHR, Health Research Council of New Zealand, UK Government, British Skin Foundation, Wellcome Trust.

• Stats Editor roles at Cochrane Gynaecology and Fertility, Fertility and Sterility, BJOG, Reproduction and Fertility, Journal of Hypertension
For the lawyers

• I’m not accusing anyone of fraud, data fabrication/falsification, or any other form of research misconduct here.

• I will say that some trials are unlikely to be authentic or are not trustworthy. The data or results do not appear to be compatible with a genuine RCT.

• I make no claims that this is due to deliberate action on behalf of investigators/ authors (vs catastrophic errors in data management, for example).
Outline

1. Detecting problematic studies in the context of health systematic reviews: the INSPECT-SR project.

2. Some principles for investigating potentially problematic RCTs.
1. Detecting problematic studies in the context of health systematic reviews: the INSPECT-SR project.
Ivermectin for COVID-19

**Ivermectin for Prevention and Treatment of COVID-19 Infection: A Systematic Review, Meta-analysis, and Trial Sequential Analysis to Inform Clinical Guidelines**

Bryant et al., 2021
Risk ratio for death: 0.38 (95% CI 0.19 to 0.73)
15 trials

**Meta-analysis of Randomized Trials of Ivermectin to Treat SARS-CoV-2 Infection**

Hill et al., 2021
Risk ratio for death: 0.49 (95% CI 0.28 to 0.86)
12 trials
Ivermectin for COVID-19

- SRs widely covered in media and social media.
- Used by antivax groups

A just-published, peer-reviewed study already clearly shows that ivermectin prevents and treats Covid-19 and has the potential to save and improve countless lives in the UK and worldwide right now.

The strength of evidence for ivermectin has this week been supercharged by publication of a gold standard review of 24 randomised trials conducted in 15 countries among more than 3400 people worldwide proving infections fall and deaths are dramatically reduced when ivermectin is administered. Published in the American Journal of Therapeutics the most rigorous statistical standards were applied by world-leading researchers biostatistician Mr Andrew Bryant and medical doctor and researcher Dr. Tess Lawrie.
The catch...

- It now appears that several of the trials were not authentic

Analysis by Nick Brown (@sTeamTraen) at steamtraen.blogspot.com
Meta-analyses restricted to ‘credible’ trials

Hill et al., retracted their systematic review (👍):

- “The significant effect of ivermectin on survival was dependent on the inclusion of studies with a high risk of bias or potential medical fraud.”
- Risk ratio for death 0.96 (95% CI 0.56 to 1.66, 4 studies)

Popp et al., 2022 (Cochrane) excluded seven trials overall

- Asymptomatic or mild disease: Risk ratio for death 0.77 (95% CI 0.47 to 1.25, 6 trials)
- Moderate to severe disease: Risk ratio for death 0.60 (95% CI 0.14 to 2.51, 3 trials, 1 with no events)
Systematic reviews: Fake data to patient care pipeline

1. Attempt to identify all RCTs on the review topic
   - Problematic trials will be included

2. Critically appraise study methodology, include in meta-analysis
   - Assess risk of bias
   - But do not consider authenticity
   - Many (not all) fake trials report sound methods

3. Make conclusions, recommendations, on basis of evidence
   - SRs seen as gold standard
   - Included in guidelines
   - Influence patient care
3 out of 5 trials subsequently identified as fake.

26 trials. 8 had identical or similar text, 2 no ethical approval.

3 of 27 trials from one investigator suggested to be implausible (huge effects, no attrition).
EDITORIAL

When beauty is but skin deep: dealing with problematic studies in systematic reviews

Stephanie L Boughton, Jack Wilkinson, Lisa Bero

Managing potentially problematic studies  https://bit.ly/3SsJO9F
EDITORIAL

When beauty is but skin deep: dealing with problematic studies in systematic reviews

Stephanie L Boughton, Jack Wilkinson, Lisa Bero

Managing potentially problematic studies  
https://bit.ly/3SsJO9F

• Do not include studies until serious concerns about trustworthiness have been resolved.
When beauty is but skin deep: dealing with problematic studies in systematic reviews

Stephanie L Boughton, Jack Wilkinson, Lisa Bero

Managing potentially problematic studies

- Do not include studies until serious concerns about trustworthiness have been resolved.
- How do we define a ‘problematic study’?
EDITORIAL

When beauty is but skin deep: dealing with problematic studies in systematic reviews

Stephanie L Boughton, Jack Wilkinson, Lisa Bero

Managing potentially problematic studies https://bit.ly/3SsJO9F

• Do not include studies until serious concerns about trustworthiness have been resolved.

• How do we define a ‘problematic study’?

• How can we detect them?
Aim: To develop a tool for identifying problematic randomised controlled trials in the context of health systematic reviews.

1. Convene a panel of people with expertise and experience of investigating problematic studies.

2. Create an extensive list of methods for detecting problematic studies.

3. Apply the list to a sample of systematic reviews (feasibility, impact on review conclusions)

4. Enter the items into a Delphi process

5. Prospective testing in production and update of systematic reviews.
Aim: To develop a tool for identifying problematic randomised controlled trials in the context of health systematic reviews.

1. Convene a panel of people with expertise and experience of investigating problematic studies.

2. Create an extensive list of methods for detecting problematic studies.

3. Apply the list to a sample of systematic reviews (feasibility, impact on review conclusions)

4. Enter the items into a Delphi process

5. Prospective testing in production and update of systematic reviews.
Making an extensive list of methods

102 checks or tests identified

- Implemented as online survey of experts
- “Are we missing anything?”
## Preliminary classifications and examples

<table>
<thead>
<tr>
<th>Inspecting results in the paper</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Are the results substantially divergent from others in the meta-analysis?</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inspecting conduct, governance and transparency</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Is the recruitment of participants plausible within the stated time frame for the research?</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inspecting the research team</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Have other studies by the research team been retracted, or do they have expressions of concern?</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inspecting text and publication details</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Is there evidence of copied work, such as duplicated or partially duplicated tables?</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inspecting individual participant data</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Does the dataset contain repeated sequences of baseline values?</em></td>
</tr>
<tr>
<td>Study or Subgroup</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>Alda 2011</td>
</tr>
<tr>
<td>Carson 2006</td>
</tr>
<tr>
<td>Ersek 2008</td>
</tr>
<tr>
<td>Greco 2004</td>
</tr>
<tr>
<td>Kaapa 2006</td>
</tr>
<tr>
<td>Keefe 1990</td>
</tr>
<tr>
<td>Keefe 1996</td>
</tr>
<tr>
<td>Kraaimaat 1995</td>
</tr>
<tr>
<td>Litt 2009</td>
</tr>
<tr>
<td>Lumley 2014</td>
</tr>
<tr>
<td>Lumley 2017</td>
</tr>
<tr>
<td>Mangels 2009</td>
</tr>
<tr>
<td>Monticone 2013</td>
</tr>
<tr>
<td>Monticone 2016</td>
</tr>
<tr>
<td>Monticone 2017</td>
</tr>
<tr>
<td>Nicholas 2013</td>
</tr>
<tr>
<td>Smee 2006</td>
</tr>
<tr>
<td>Tava 2011</td>
</tr>
<tr>
<td>Thieme 2006</td>
</tr>
<tr>
<td>Thorn 2011</td>
</tr>
<tr>
<td>Thorn 2018</td>
</tr>
<tr>
<td>Thorsell 2011</td>
</tr>
<tr>
<td>Turner 2006</td>
</tr>
<tr>
<td>van Eijk 2013</td>
</tr>
<tr>
<td>Vitiello 2013</td>
</tr>
<tr>
<td>Vlaeyen 1994</td>
</tr>
<tr>
<td>Zautra 2008</td>
</tr>
</tbody>
</table>

**Total (95% CI):**

2017: 1718, 100.0%, -0.33 [-0.56, -0.10]

**Heterogeneity:** Tau² = 0.32, Chi² = 293.71, df = 26 (P < 0.00001); I² = 91%

**Test for overall effect:** Z = 2.82 (P = 0.005)

**Test for subgroup differences:** Not applicable

---

**Psychological therapies for chronic pain**

Williams, et al. 2020
## Preliminary classifications and examples

<table>
<thead>
<tr>
<th>Inspection Category</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inspecting results in the paper</strong></td>
<td>Are the results substantially divergent from others in the meta-analysis?</td>
</tr>
<tr>
<td><strong>Inspecting conduct, governance and transparency</strong></td>
<td>Is the recruitment of participants plausible within the stated time frame for the research?</td>
</tr>
<tr>
<td><strong>Inspecting the research team</strong></td>
<td>Have other studies by the research team been retracted, or do they have expressions of concern?</td>
</tr>
<tr>
<td><strong>Inspecting text and publication details</strong></td>
<td>Is there evidence of copied work, such as duplicated or partially duplicated tables?</td>
</tr>
<tr>
<td><strong>Inspecting individual participant data</strong></td>
<td>Does the dataset contain repeated sequences of baseline values?</td>
</tr>
</tbody>
</table>
Results

- 71 participants – 5 continents, but mostly Europe (55%), Australia/ NZ (21%), N America (14%).
- 25 pages of comments: 16 new checks proposed, many suggestions to modify existing checks (e.g. merging, splitting or rewording).

<table>
<thead>
<tr>
<th>Domain</th>
<th>Number of checks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Inspecting results in the paper</td>
<td>28</td>
</tr>
<tr>
<td>2. Inspecting the research team and their work</td>
<td>19</td>
</tr>
<tr>
<td>3. Inspecting conduct, governance and transparency</td>
<td>22</td>
</tr>
<tr>
<td>4. Inspecting text and publication details</td>
<td>7</td>
</tr>
<tr>
<td>5. Inspecting individual participant data</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>117</td>
</tr>
</tbody>
</table>
Aim: To develop a tool for identifying problematic randomised controlled trials in the context of health systematic reviews.

1. Convene a panel of people with expertise and experience of investigating problematic studies.

2. Create an extensive list of methods for detecting problematic studies.

3. Apply the list to a sample of systematic reviews (feasibility, impact on review conclusions)

4. Enter the items into a Delphi process

5. Prospective testing in production and update of systematic reviews.
Approximately 50 researchers...

<table>
<thead>
<tr>
<th>Domain</th>
<th>Number of checks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Inspecting results in the paper</td>
<td>28</td>
</tr>
<tr>
<td>2. Inspecting the research team and their work</td>
<td>19</td>
</tr>
<tr>
<td>3. Inspecting conduct, governance and transparency</td>
<td>22</td>
</tr>
<tr>
<td>4. Inspecting text and publication details</td>
<td>7</td>
</tr>
<tr>
<td>5. Inspecting individual participant data</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>117</td>
</tr>
</tbody>
</table>

Applying the long list of checks...

To RCTs in 50 Cochrane Reviews

- How often is each check failed?
- How feasible are the checks?
- What is the impact of removing flagged trials?
Interested?

- Need input and collaboration at all stages – methodologists, trialists, systematic reviewers, editors, publishers, patients, research integrity professionals, or researchers with experience.

- Credible tool needs to be feasible, backed by broad consensus.

- INSPECT-SR Workshop at Colloquium: come and try an early draft version, feedback – join us!

- Need people to participate in Delphi (methods experts and potential users of the tool)

- Need people who would be willing to test the tool while undertaking a systematic review.

- If you have any expertise, experience or interest, please contact me:

  - jack.wilkinson@manchester.ac.uk or @jd_wilko
Available tools or frameworks

Experts identified warning signs of fraudulent research: a qualitative study to inform a screening tool
Lisa Parker, Stephanie Boughton, Rosa Lawrence, Lisa Bero

Checklist to assess Trustworthiness in RAnAomised Controlled Trials (TRACT checklist): concept proposal and pilot
Ben W. Moli, Shimona Lai, Ayesha Rahim, Esmée M. Bordeuwijk, Rui Wang, Rik van Eekelen, Lyle C. Gurin, Jim G. Thornton, Madelon van Wely, and Wentao Li

Identifying and handling potentially untrustworthy trials in Pregnancy and Childbirth Cochrane Reviews
Alfirevic Z, Kellie FJ, Stewart F, Jones L, Hampson L, on behalf of Pregnancy and Childbirth Editorial Board

Identifying and managing problematic trials: A research integrity assessment tool for randomized controlled trials in evidence synthesis
Stephanie Weibel, Maria Popp, Stefanie Reis, Nicole Skoetz, Paul Garner, Emma Sydenham
Individual participant data integrity assessment tool

- Tool consisting of a checklist and semi-automated scripts to assess clinical trials for individual participant data meta-analyses (IPD-MA)
- Based on existing literature, mapping exercise and expert consensus
- Pilot tested and refined using two large IPD-MA in child health
- Developed by NextGen Evidence Synthesis Team at NHMRC Clinical Trials Centre, University of Sydney

For more information, or if you are interested in piloting the tool, please contact Kylie Hunter at kylie.hunter@sydney.edu.au
Semi-automated software for analytical forensics: RCT baseline tables as a proof of concept

- Upload and view PDF
- Automated baseline table identification
- Automated p-value extraction and distribution visualization
- Manually correct extraction errors
2. Some principles for investigating potentially problematic RCTs.
• For this part of the talk, drawing on my experience investigating potentially problematic trials for journals and publishers over past four years. Trained by Stephen Evans.

• **Confidentiality** – following examples are illustrative – inspired by real cases, but I have changed details.

• Investigation involves a thorough examination of the manuscript, data, and other sources (registration, correspondence with authors, potentially other papers from the authors).

• Not trying to prove misconduct. Could these data have arisen from a genuine RCT?
• Conclusions based on a **holistic assessment** – not a single statistical test.

• Usually a day’s work (at least).

• Illustrating some basic principles here – not comprehensive, not a tutorial!
Stage 1: Concerns with study

• Usually on the basis of published or submitted information (manuscript, trial registration)
Stage 1: Concerns with study

- Usually on the basis of published or submitted information (manuscript, trial registration)

Stage 2: Detailed investigation

- Request additional documentation, individual participant data (IPD)
- Analysis of IPD
Stage 1: Concerns with study

- Usually on the basis of published or submitted information (manuscript, trial registration)

Stage 2: Detailed investigation

- Request additional documentation, individual participant data (IPD)
- Analysis of IPD

How?
Stage 1: Concerns with study
- Usually on the basis of published or submitted information (manuscript, trial registration)

Stage 2: Detailed investigation
- Request additional documentation, individual participant data (IPD)
- Analysis of IPD
Stage 1: Concerns with study

- Usually on the basis of published or submitted information (manuscript, trial registration)

Stage 2: Detailed investigation

- Request additional documentation, individual participant data (IPD)
- Analysis of IPD

- Want to ensure fabricated data cannot influence patient care (e.g. through meta-analysis).
- Want to avoid unintentionally removing genuine data from the literature.
Stage 1: Concerns with study
• Usually on the basis of published or submitted information (manuscript, trial registration)

Stage 2: Detailed investigation
• Request additional documentation, individual participant data (IPD)
• Analysis of IPD

• Want to ensure fabricated data cannot influence patient care (e.g. through meta-analysis).
• Want to avoid unintentionally removing genuine data from the literature.
Recruitment and allocation to treatments in an RCT

1. Potential participants present over time, have their baseline measurements taken before randomisation.

Baseline measurements assessed
1. Potential participants present over time, have their baseline measurements taken before randomisation.

2. Eligible participants are sequentially allocated to study arms according to a random sequence.
Recruitment and allocation to treatments in an RCT

1. Potential participants present over time, have their baseline measurements taken before randomisation.

2. Eligible participants are sequentially allocated to study arms according to a random sequence.

Baseline measurements assessed
Recruitment and allocation to treatments in an RCT

1. Potential participants present over time, have their baseline measurements taken before randomisation.

2. Eligible participants are sequentially allocated to study arms according to a random sequence.

Baseline measurements assessed
Recruitment and allocation to treatments in an RCT

1. Potential participants present over time, have their baseline measurements taken before randomisation.

2. Eligible participants are sequentially allocated to study arms according to a random sequence.
Recruitment and allocation to treatments in an RCT

1. Potential participants present over time, have their baseline measurements taken before randomisation.

2. Eligible participants are sequentially allocated to study arms according to a random sequence.

Baseline measurements assessed
1. Potential participants present over time, have their baseline measurements taken before randomisation.

2. Eligible participants are sequentially allocated to study arms according to a random sequence.

- No systematic differences between groups in baseline characteristics.
- Any patterns in baseline characteristics over time should appear in both groups...
Recruitment and allocation to treatments in an RCT

1. Potential participants present over time, have their baseline measurements taken before randomisation.

2. Eligible participants are sequentially allocated to study arms according to a random sequence.

Change in process
Recruitment and allocation to treatments in an RCT

1. Potential participants present over time, have their baseline measurements taken before randomisation.

2. Eligible participants are sequentially allocated to study arms according to a random sequence.

Change in process
Recruitment and allocation to treatments in an RCT

1. Potential participants present over time, have their baseline measurements taken before randomisation.

2. Eligible participants are sequentially allocated to study arms according to a random sequence.

Change in process
1. Potential participants present over time, have their baseline measurements taken before randomisation.

2. Eligible participants are sequentially allocated to study arms according to a random sequence.
Recruitment and allocation to treatments in an RCT

1. Potential participants present over time, have their baseline measurements taken before randomisation.

2. Eligible participants are sequentially allocated to study arms according to a random sequence.

3. Patterns in baseline characteristics should be apparent in both arms.
Example – genuine data

3-arm RCT. Blue lines divide into treatment groups – plotted in randomisation order.
Example – genuine data

3-arm RCT. Blue lines divide into treatment groups – plotted in randomisation order.
Example – genuine data

3-arm RCT. Blue lines divide into treatment groups – plotted in randomisation order.
Example – dubious data

2-arm RCT

Group A

Group B
Example – dubious data

2-arm RCT

Group A

Group B
Example – dubious data

2-arm RCT

Group A

Group B

?
Other problems can be revealed by these plots

Take a moment – can you spot any problems?
Other problems can be revealed by these plots
Looking at outcome variables

- Outcomes are a bit different

- They are influenced by treatment, so we do expect to see differences

- But plotting against randomisation order can still reveal improbable patterns...
Looking at outcome variables

- Outcomes are a bit different

- They are influenced by treatment, so we do expect to see differences

- But plotting against randomisation order can still reveal improbable patterns...
Looking at outcome variables

- Outcomes are a bit different
- They are influenced by treatment, so we do expect to see differences
- But plotting against randomisation order can still reveal improbable patterns...
Correlation across rows in the dataset

• We don’t expect to see substantial correlation between the baseline values of successive participants.

• E.g. each participant’s duration of infertility shouldn’t be related to the duration of infertility of the person recruited after them, or to the next person’s, or the next person’s etc.

• We do expect correlation across rows if someone has typed (fabricated) values into the column – people are poor random number generators.

• Certainly don’t expect this to differ between randomised groups.
Autocorrelation plot

- Plot of correlation between duration of infertility values 1 row apart (Lag = 1), 2 rows apart (Lag = 2), 3 rows apart etc.
Autocorrelation plot

- Plot of correlation between duration of infertility values 1 row apart (Lag = 1), 2 rows apart (Lag = 2), 3 rows apart etc.
- In treatment group, there is a correlation between successive rows, which decays as we get further apart.
Autocorrelation plot

- Plot of correlation between duration of infertility values 1 row apart (Lag = 1), 2 rows apart (Lag = 2), 3 rows apart etc.
- In treatment group, there is a correlation between successive rows, which decays as we get further apart.
- Control group looks like genuine data – no serial correlation between rows.
Autocorrelation plot

- Plot of correlation between duration of infertility values 1 row apart (Lag = 1), 2 rows apart (Lag = 2), 3 rows apart etc.
- In treatment group, there is a correlation between successive rows, which decays as we get further apart.
- Control group looks like genuine data – no serial correlation between rows.
- The treatment group correlation is suspicious, and the difference between groups is more so.
Relationships between variables

• Are expected relationships between variables present?

• Hard to fake (unless you know what you are doing).

• Requires contextual knowledge (e.g. should we expect relationship between gestational age and birthweight in a particular trial).

• Don’t expect multivariate distribution to differ between randomised groups.
Closing comments

- Have shown just a few basic checks here. Different approaches may be more or less appropriate for particular cases.

- We understand a lot about characteristics of data arising from RCTs. Can use this to assess whether data are (in)compatible with a genuine RCT.

- “Could there be an explanation for this?”

- Sometimes there is clear evidence of fabrication (e.g. certain cases with repeating sequences in the data). Other times, unclear whether misconduct or very poor conduct.

- Either way, may have reservations about using the data to decide how patients are treated.
Thanks to expert panel members

<table>
<thead>
<tr>
<th>Elizabeth Loder</th>
<th>Toby Lasserson</th>
<th>Kyle Sheldrick</th>
<th>Andrew Grey</th>
<th>Susan Garfinkel</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Carlisle</td>
<td>Tianjing Li</td>
<td>Emily Lam</td>
<td>David Torgerson</td>
<td>Andreas Lundh</td>
</tr>
<tr>
<td>Karla Soares-Weiser</td>
<td>Neil O’ Connell</td>
<td>Rebecca Jones</td>
<td>Esmée Bordewijk</td>
<td>Lyle Gurrin</td>
</tr>
<tr>
<td>Rita Redberg</td>
<td>Lisa Parker</td>
<td>Darren Dahly</td>
<td>Nick Brown</td>
<td>Lene Seidler</td>
</tr>
<tr>
<td>Jo Dumville</td>
<td>Virginia Barbour</td>
<td>Alison Avenell</td>
<td>Wentao Li</td>
<td>Kylie Hunter</td>
</tr>
<tr>
<td>Mike Clarke</td>
<td>Ben Mol</td>
<td>James Heathers</td>
<td>Richard Stevens</td>
<td></td>
</tr>
<tr>
<td>Emma Sydenham</td>
<td>Barbara Redman</td>
<td>Gideon Meyerowitz-Katz</td>
<td>Rafael Perera-Salazar</td>
<td></td>
</tr>
<tr>
<td>Jane Dennis</td>
<td>Jill Hayden</td>
<td>Madelon van Wely</td>
<td>Sarah Lensen</td>
<td></td>
</tr>
</tbody>
</table>

- Need people to participate in Delphi.
- Need people who would be willing to test a tool while undertaking a systematic review.
- If you have any expertise, experience or interest, please contact me:

  - [jack.wilkinson@manchester.ac.uk](mailto:jack.wilkinson@manchester.ac.uk) or [@jd_wilko](https://twitter.com/jd_wilko)