Improving peer review via controlled experiments

Isabelle Boutron

METHODS team
Research Centre of Epidemiology Biostatistics Sorbonne Paris Cité
Paris Descartes University
Background

- Clinical epidemiologist

- Focus on clinical research (randomized controlled trials)
  - Develop interventions to improve research and use high level evidence study designs or modelling to evaluate these interventions

METHODS Team
Methods of therapeutic evaluation of chronic diseases

The focus of the French EQUATOR centre is on increasing/improving/enhancing the value of biomedical research.
MiRoR project
Joint doctoral training program, to train 15 PhD students in Methods in Research on Research in the field of clinical research ([http://miror-ejd.eu](http://miror-ejd.eu)) funded by Marie Skłodowska-Curie Actions.

Projects
-Planning
-Conduct
-Reporting
-Peer review

7 European Universities and 10 International Partners
The peer review system

• Central to the scientific community

• 2 major goals
  – Gatekeeper of the scientific publications
  – To improve the quality of manuscripts

• A system relying mainly on work performed voluntarily by academic researchers
Is the peer-review system achieving its goal?
Two versions of a well-designed randomized controlled trial that differed only in the direction of the finding of the principal study end point

- were more likely to recommend the positive version of the test manuscript for publication than the no-difference version (97% vs 80%, P<0.001)
- detected more errors in the no-difference version than in the positive version (mean 0.85 vs 0.41, P<0.001)
- awarded higher methods scores to the positive manuscript than to the (identical) no-difference manuscript (8.24 vs 7.53, P=0.005)
119 reviewers were randomized to assess a fabricated manuscript with the prestigious authors’ names and institutions masked or visible

- Reviewers were more likely to recommend acceptance when the prestigious authors’ names and institutions were visible than when they were redacted
  - 87% vs 68%;
  - RR, 1.28 [95% CI, 1.06-1.39], P = .02

- They gave higher ratings for the methods.
Impact of peer review on reports of randomised trials published in open peer review journals: retrospective before and after study

- Comparison of the submitted and accepted versions of manuscripts of RCTs published in open access journals (n=93)

- Changes were limited
  - A median of 11% (range 1-60%) words deleted and 20% added (range 2-88%).

Hopewell S, BMJ, 2014
Impact of the peer review process on Transparency and completeness of reporting

• 93 RCTs published
• First submission / Peer-reviewers’ comments/ Published article

<table>
<thead>
<tr>
<th>Adequate reporting</th>
<th>Submitted</th>
<th>Published</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence generation</td>
<td>47%</td>
<td>59%</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>34%</td>
<td>44%</td>
</tr>
<tr>
<td>Blinding</td>
<td>33%</td>
<td>45%</td>
</tr>
<tr>
<td>Primary outcome</td>
<td>51%</td>
<td>51%</td>
</tr>
<tr>
<td>Results for the primary outcome</td>
<td>35%</td>
<td>35%</td>
</tr>
</tbody>
</table>

Hopewell S, BMJ, 2014
Transparency of published reports (peer-reviewed) vs posting in trial registries (no peer-review)

Timing and Completeness of Trial Results Posted at ClinicalTrials.gov and Published in Journals

Sample of randomized controlled trials of drugs with results both posted and published

<table>
<thead>
<tr>
<th>Completeness of reporting</th>
<th>ClinicalTrials.gov N=202</th>
<th>Published article N=202</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow of participants</td>
<td>64%</td>
<td>48%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Efficacy results</td>
<td>79%</td>
<td>69%</td>
<td>0.02</td>
</tr>
<tr>
<td>Adverse events</td>
<td>73%</td>
<td>45%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Serious adverse events</td>
<td>99%</td>
<td>63%</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Riveros Plos Med, 2013
Detection of selective reporting of outcomes

Survey of 676 authors and reviewers who had reviewed at least 1 article reporting a clinical trial in the past 2 years

34% examined information registered on a trial registry.
Detection of selective reporting of outcomes

- 226 manuscripts of drug trials submitted to 8 journals
- (29%) had changes in the primary outcome
- The pattern of editorial decisions was not statistically significantly different for manuscripts with or without changed primary outcomes ($P = 0.418$).
Can peer reviewers reduce spin?

Spin: A way of reporting to convince the reader that the beneficial effect of the experimental treatment (efficacy, safety) is higher than shown by the results

Spin is frequent in published reports
- 50% of abstract conclusions of RCTs have spin

Boutron JAMA 2009
Can peer reviewers reduce spin?

- Assessment of the manuscript submitted, peer reviewers comments, and final manuscript of non-randomized studies assessing a therapeutic intervention published in open access journals (n=128).

- 55% of submitted manuscripts, peer reviewers identified at least one example of spin

- Of the spin identified by peer reviewers
  - 67% were completely deleted,
  - 16% partially deleted
  - 17% not removed in the final published article.

- For 15%, peer reviewers requested adding some spin

# Level of spin before and after peer review in the abstract conclusion

<table>
<thead>
<tr>
<th>Level</th>
<th>Before peer-review</th>
<th>N=66</th>
<th>After peer-review</th>
<th>N=61</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td></td>
<td>24</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td>10</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>No</td>
<td></td>
<td>28</td>
<td></td>
<td>26</td>
</tr>
</tbody>
</table>

76% Peer reviewers failed to identify spin in abstract conclusions

Controlled experiments to improve the peer-review process
Interventions to improve transparency

**CONSORT Statement, recommendations for reporting RCTs**

- **CONSORT:** 25-item checklist + E&E paper
- Requested in **instructions to authors** of several journals
- Authors are requested to submit the relevant checklist
- **Adherence of authors** to these guidelines remains **low**
Impact of a web-based tool (WebCONSORT) to improve the reporting of randomised trials: results of a randomised controlled trial

Sally Hopewell1,2,3*, Isabelle Boutron3,4, Douglas G. Altman2, Ginny Barbour2, David Moher6, Victor Montori7, David Schriger8, Jonathan Cook2, Stephen Gerry2, Omar Omar2, Peter Dutton2, Corran Roberts2, Eleni Frangou2, Lei Clifton2, Virginia Chiocchia2, Ines Rombach2, Karolina Wartolowska2, and Philippe Ravaud3,4

Combined checklist
19 manuscripts selected for RCT by the editorial staff were not randomised

- 46 journals actively recruited into the trial
- 324 manuscripts were randomised
Are CONSORT checklists submitted by authors adequately reflecting what information is actually reported in published papers?

David Blanco\textsuperscript{1,3}, Alice M. Biggane\textsuperscript{2,3}, Erik Cobo\textsuperscript{1} and MiRoR network

![Figure 1](image_url) Reporting inconsistencies found for the six papers that used the appropriate CONSORT checklist.
Format
- Checklist of items: not understood by most authors
- Elaboration & Explanation manuscript: >30 pages (never read by authors)

Checklist

Item 6 a
"Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed"

Format
- Template with the checklist tailored and expanded with details on what should be reported for each item

Writing Aid Tool
Expanded, combined and tailored checklist

Please report
- Which outcome(s) is the primary outcome(s) (i.e. pre-specified outcome considered to be of greatest importance).

For each primary and secondary outcome, report
- The variable of interest (eg, pain)
- How the outcome was assessed (eg, VAS)
- The analysis metric (eg, change from baseline)
- Time point of interest for analysis (eg, 3 months)
- ...
Completeness of reporting (0-10) writing aid tool vs none
Mean Difference = 2.1 [1.5;2.7]; p <0.001

Completeness of reporting (0- 10) writing aid tool vs NEJM/JCO
Mean Difference = 1.7 [1.1; 2.4]; p <0.001

• Design: “Split-manuscript randomised controlled trial” with blinded outcome assessment
• 41 students (masters and PhD)
• Session: 4 hours

Barnes C, Boutron I, ... Ravaud P. BMC Med. 2015
Cobweb

**Methods**

**Trial design**

Participants

Interventions

Outcomes

Sample size

Randomisation

Blinding

Statistical methods

**Results**

*Save changes*

**Description of trial design (such as parallel, factorial) including allocation ratio**

Please report:

- The type of trial design (parallel group)
- The conceptual framework (superiority, non-inferiority, equivalence, other)
- The allocation ratio
- Any other pertinent information (for drug development (phase 1,2,3), other)

**Example**

“This was a multicenter, stratified (6 to 11 years and 12 to 17 years of age, with imbalanced randomisation [2:1]), double-blind, placebo-controlled, parallel-group study conducted in the United States (41 sites).” (Blumer et al., Pediatrics 2009)
The tasks expected from a peer-reviewer are not realistic
- More than 200 different tasks identified
- The tasks involved different level of expertise and different backgrounds
  - Statistical and methodologic expertise
  - Content expertise (novelty, relevance, external validity etc)
  - None
    - Verification (adherence to guidelines, consistency with registries)
    - Formatting
Multiple tasks asked to reviewers (important tasks for reviewers are not congruent with important tasks for editors)

- The most important task for peer reviewers (evaluating the risk of bias) was clearly requested by 5% of editors.
- The task most frequently requested by editors (provide recommendations for publication), was rated in the first tertile only by 21% of all participants.
A 2-step peer-review system to improve completeness of reporting. Theoretical background of the intervention

- Peer-reviewers consider that assessing adherence to reporting guidelines and switch in outcomes is not their remit.

- Detection of inadequate reporting does not involve high level of expertise and could be performed by early career researchers who would gain expertise in peer-review.

- Online tool and a training module could early career researchers detected misreported items and switched outcomes when evaluating a report of a RCT.
Development of the intervention - COBPeer

COBPeer A Consort-Based Peer-Review Tool

The ARTISTIC (Arthritis Intervention StAnDarized) study:
A pragmatic randomized controlled trial comparing standardized consultation to usual care for patients with knee osteoarthritis under primary care

P. Revault (MD, PhD), B-A. Filipe (MD); J. Boutron (MD, PhD), C. Ray (MD); A. Mahmoudi (MD), B. Quirous (MD), T. Plass (MD)

1 INSERM, U738, Paris, France; 2 Université Paris 7 Denis Diderot, UFR de Médecine, Paris, France; 3 AP-HP, Hôpital Bichat, Département d’Épidémiologie, Biostatistique et Recherche Clinique, Paris, France

Automatic review reports

OUTCOMES
Completely defined your primary outcome measures, including how and when they were assessed

- Was the primary outcome clearly identified? [ ] Yes [ ] No

RANDOMIZATION / SEQUENCE GENERATION
Method used to generate the random allocation sequence

Did the author report:

- The method of sequence generation (e.g., table, random number table, or computerized random number generator, or other) [ ] Yes [ ] No

ALLOCATION CONCEALMENT
Mechanism used to implement the random allocation sequence (e.g., sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned

Did the author report:

- Peer-Review report

Dear Authors,

Please note that the completeness of reports of your manuscript has been assessed and some essential information is missing. We have summarized below the list of missing information that should be reported before sending your manuscript to our external peer review.

Should you need any further information, please do not hesitate to contact us.

Sincerely,

Primary Outcome:
- Please make sure you clearly report which outcome is the primary outcome (i.e., the pre-specified outcome considered to be of greatest importance and usually used in the sample size calculation)
- For the primary outcome please clearly report the following information:
  - The variable of interest (e.g., pain, Back Depression Inventory score, all-cause mortality)
  - How the variable of interest was measured (e.g., VAS, Back Depression Inventory Score)
  - The analysis model (e.g., change from baseline, final value, time to event)
  - The summary measure for each study group (e.g., mean, proportion with score > 2)
  - The time point of interest for analysis (e.g., one hour, 3 months)
  - How the outcome was assessed (e.g., EVA)

Participant flow:
- Please report a flowchart in the flow diagram or in the text, please report:
  - Number of participants lost to follow-up with reasons in each group
  - Number of participants analyzed for the primary outcome in each group

Outcomes and estimation
For the primary outcome, please report:
- The precision for difference between groups (e.g., 95% CI)

Harms
Please report:
- For each group, the number of participant withdrawals due to harms
- Results in each group for each adverse event type (mean (SD) or number of events)

Registration
- If you registered your protocol, please reported the registration number. If you did not register your protocol, please give your reasons.
Development of the intervention – COBPeer training programme

**COBPeer training module**

**Item 6a. Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed**

--

**Outcomes**

After one-year anthropometric data, information on smoking behavior, changes in medication and medical history, fasting blood glucose, lipids and creatinine were collected. The baseline questionnaire was repeated and treatment plans were analyzed on the number and duration of visits and use of self-monitoring. For the calculation of the SCORE risk assessment both at baseline and after one year, the age at baseline was used, entailing slight underestimation of the risk after one year.

<table>
<thead>
<tr>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Was the primary outcome clearly identified?*

**Outcomes**

After one-year anthropometric data, information on smoking behavior, changes in medication and medical history, fasting blood glucose, lipids and creatinine were collected. The baseline questionnaire was repeated and treatment plans were analyzed on the number and duration of visits and use of self-monitoring. For the calculation of the SCORE risk assessment both at baseline and after one year, the age at baseline was used, entailing slight underestimation of the risk after one year.

**Explanation**

Authors reported numbers data collected, but they did not indicate which was the primary outcome. So the primary outcome reported by authors is unclear.

1/ The variable of interest: NOT REPORTED BY AUTHORS
2/ How the outcome(s) were assessed: NOT REPORTED BY AUTHORS
3/ The format of the primary outcome data: NOT REPORTED BY AUTHORS
4/ The summary measure for each study group: NOT REPORTED BY AUTHORS
5/ The time frame to measure primary outcome: NOT REPORTED BY AUTHORS
6/ Who assessed the outcome: NOT REPORTED BY AUTHORS

→ Congratulations! Your answers are correct.
Assessment of the intervention

• Objective
  – To compare the performance of early career peer reviewers who use the peer reviewer tool (COBPeer) with usual peer reviewers in identifying incomplete reporting and switched outcomes in reports of RCTs.

• Study design
  – Cross-sectional study comparing the accuracy of early career peer reviewers using COBPeer to that of usual peer reviewers when evaluating the completeness of reporting and a switched in primary outcome(s) in completed reports of RCTs at the first submission
  – Gold standard: assessment of systematic reviewers
Assessment of the intervention

Initial submitted report
N=120

Usual peer review

Assessment by 2 researchers
(standard peer review)

Gold standard

Assessment by 2 systematic reviewers

Early career researchers

Assessment by 1 early career researchers
with the tool

1. The completeness of reporting
2. Switched outcomes
Several interventions are implemented or proposed to improve the system.
What is the evidence?
A Systematic Review and Meta-analysis of Interventions to Improve the Quality of Peer Review

- Only 22 reports of RCTs
- Only 7 were published over the past 10 years
- Interventions assessed
  - Blinding
  - Open peer review
  - Training
  - Use of checklist
  - Adding experts
- Most are performed in one single journal
- Low methodologic quality

Bruce, BMC Med, 2016
What should be the study design?
What should be the study design?

**Manuscripts**
- Intervention
- Control
  - Outcome assessment

**Peer reviewers**
- Intervention
- Control
  - Outcome assessment
  - Type of manuscript (real, fabricated)

**Cluster RCTs**
- Intervention
- Control
  - Outcome assessment

**Stepped-wedge RCT**

**Time series analysis**
- Active implementation
  - (Lancet, Ann Intern Med)
  - Level change: $p = 0.0035$
  - Trend change: $p = 0.0154$

**Paired study design**
- Intervention
- Usual process

- **Type of manuscript (real, fabricated)**
- **Setting single journal, several journals, several publisher**
What should be the study design?

Which design would you choose?

- Series of vignette based studies
- 6 interventions
- Pairs of vignettes evaluated by 204 experts
What should be the outcome?

• Quality of the final manuscript (how is high quality defined, who is to decide?)

• Quality of the peer review report (editor’s subjective assessment, validated scales to assess the quality of the peer review report)
What should be the outcome?

Systematic review of all tools assessing the quality of a peer-review report

• 23 scales
• 1 checklist
• No tool defined the concept of quality
• Number of domains varied

Superchi C, PEERE, 2018
Make peer review scientific

Thirty years on from the first congress on peer review, Drummond Rennie reflects on the improvements brought about by research into the process — and calls for more.
Acknowledgements

• Amytis Heim (master student)
• Anthony Chauvin (PhD student)
• Philippe Ravaud (Director of the Centre of Epidemiology Biostatistics Sorbonne Paris Cité)
• Cecilia Superchi (MiRoR PhD student)
A 2-step peer-review system to improve completeness of reporting. Theoretical background of the intervention

Impossible to adequately assess / lack of transparency

Assessment of a transparent manuscript

Authors

Assessment of a transparent manuscript