Indicators of questionable research practices were identified in 163,129 randomized controlled trials

Johanna A. Damen, Pauline Heus, Herm J. Lamberink, Joeri K. Tijdink, Lex Bouter, Paul Glasziou, David Moher, Willem M. Otte, Christiaan H. Vinkers, Lotty Hooft

PII: S0895-4356(22)00307-9
DOI: https://doi.org/10.1016/j.jclinepi.2022.11.020
Reference: JCE 10966

To appear in: Journal of Clinical Epidemiology

Received Date: 13 July 2022
Revised Date: 17 November 2022
Accepted Date: 29 November 2022

Please cite this article as: Damen JA, Heus P, Lamberink HJ, Tijdink JK, Bouter L, Glasziou P, Moher D, Otte WM, Vinkers CH, Hooft L, Indicators of questionable research practices were identified in 163,129 randomized controlled trials, Journal of Clinical Epidemiology (2023), doi: https://doi.org/10.1016/j.jclinepi.2022.11.020.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier Inc.
Indicators of questionable research practices were identified in 163,129 randomized controlled trials.

Johanna A Damen¹, Pauline Heus¹, Herm J Lamberink²,³, Joeri K Tijdink⁴, Lex Bouter⁵, Paul Glasziou⁶, David Moher⁷, Willem M Otte²,⁸, Christiaan H Vinkers⁹, Lotty Hooft¹

¹ Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands.
² Department of Child Neurology, UMC Utrecht Brain Center, University Medical Center Utrecht and Utrecht University, Utrecht, The Netherlands.
³ Department of Neurology, Haaglanden Medical Center, Den Haag, The Netherlands.
⁴ Department of Ethics, Law and Humanities, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands and Department of Philosophy, Vrije Universiteit Amsterdam, the Netherlands.
⁵ Department of Epidemiology and Data Science, Amsterdam UMC, and Department of Philosophy, Vrije Universiteit, Amsterdam, The Netherlands.
⁶ Institute for Evidence-Based Healthcare, Bond University, Gold Coast, Australia.
⁷ Centre for Journalology, Clinical Epidemiology Program, The Ottawa Hospital Research Institute, Ottawa, Canada; School of Epidemiology and Public Health, University of Ottawa, Canada.
⁸ Biomedical MR Imaging and Spectroscopy group, Center for Image Sciences, University Medical Center Utrecht and Utrecht University, Utrecht, The Netherlands.
⁹ Department of Psychiatry and Department of Anatomy and Neurosciences, Amsterdam University Medical Center (Amsterdam UMC, location VUmc), Amsterdam, The Netherlands.

https://orcid.org/0000-0001-7401-4593 (JAD)
https://orcid.org/0000-0002-6886-4652 (PH)
27 https://orcid.org/0000-0003-1379-3487 (HJL)
28 https://orcid.org/0000-0002-1826-2274 (JKT)
29 https://orcid.org/0000-0002-2659-5482 (LB)
30 https://orcid.org/0000-0001-7564-073X (PG)
31 https://orcid.org/0000-0003-2434-4206 (DM)
32 https://orcid.org/0000-0003-1511-6834 (WMO)
33 https://orcid.org/0000-0003-3698-0744 (CHV)
34 https://orcid.org/0000-0002-7950-2980 (LH)
35
36 Corresponding author
37 Johanna A Damen
38 Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht
39 P.O. Box 85500
40 Str. 6.131
41 3508 GA Utrecht
42 The Netherlands
43 J.A.A.Damen@umcutrecht.nl
44 +31 6 25777481
Abstract

Objective: To explore indicators of the following questionable research practices (QRPs) in randomized controlled trials (RCTs): (1) risk of bias in four domains (random sequence generation, allocation concealment, blinding of participants and personnel, and blinding of outcome assessment); (2) modifications in primary outcomes that were registered in trial registration records (proxy for selective reporting bias); (3) ratio of the achieved to planned sample sizes; and (4) statistical discrepancy.

Study design and setting: Full-texts of all human RCTs published in PubMed in 1996-2017 were automatically identified, and information was collected automatically. Potential indicators of QRPs included author-specific, publication-specific, and journal-specific characteristics. Beta, logistic, and linear regression models were used to identify associations between these potential indicators and QRPs.

Results: We included 163,129 RCT publications. The median probability of bias assessed using RobotReviewer software ranged between 43% and 63% for the four risk of bias domains. A more recent publication year, trial registration, mentioning of CONSORT-checklist, and a higher journal impact factor were consistently associated with a lower risk of QRPs.

Conclusion: This comprehensive analysis provides insight into indicators of QRPs. Researchers should be aware that certain characteristics of the author team and publication are associated with a higher risk of QRPs.

Keywords: responsible research, bias, meta-research, RCT, questionable research, selective reporting

Running title: Indicators of questionable research practices
What is new?

- We included 163,129 randomized controlled trial publications and automatically extracted information about questionable research practices and potential indicators.
- Previously identified associations between indicators and questionable research practices were validated and new indicators were explored.
- A more recent publication year, trial registration, mentioning of CONSORT-checklist, and a higher journal impact factor were consistently associated with a lower risk of questionable research practices.
- Editors, peer reviewers and readers should be aware that certain characteristics of the author team, the journal and the publication might be associated with questionable research practices.
1. Introduction

Systematic reviews synthesize the results of randomized clinical trials (RCTs) and constitute the backbone of evidence-based medicine. Healthcare professionals rely on these reviews and guidelines to determine which treatments to use in clinical practice. Knowledge gained from RCTs is increasing but methods to minimize bias are not always used, leading to methodological flaws, statistical problems, and interpretation bias (spin), often making the methods and results difficult to reproduce.\(^5\)\(^6\)

Concerns about the quality of research are certainly not new. Questionable research practices (QRPs) were mentioned in the 1958 Code of Professional Ethics and Practices of Public Opinion.\(^8\) Banks et al. defined QRPs as design, analytic, or reporting practices that have been questioned because of the potential for the practice to be employed to present biased evidence in favor of an assertion.\(^10\) Examples of QRPs include selective reporting, p-hacking, HARKing (i.e., hypothesizing after results are known) etc.\(^10\)\(^-\)\(^12\)

To promote responsible research practices (RRPs), codes of conduct have been published, including the European Code of Conduct for Research Integrity\(^13\) and a report on Fostering Integrity of Research by the US National Academies of Science.\(^14\)\(^15\) Evidence exists for some indicators of QRP. For example, associations have been reported between journal impact factor and risk of bias,\(^16\) author experience and effect sizes,\(^17\) and study quality and the continent of origin of authors.\(^18\)\(^19\)

Previous studies focused on one specific QRP and explored a limited set of indicators in small datasets. Furthermore, time trends in quality indicators of RCTs have been described before in large datasets, including the dataset used in the present paper.\(^20\)\(^21\) To obtain more insight into possible factors associated with QRPs, a large study, including more QRPs and a broader set of indicators is
necessary. We therefore aimed to validate existing and identify new indicators of QRPs in RCTs. We investigated QRPs concerning risk of bias, modifications in primary outcomes, the ratio of achieved sample size compared to planned sample size, and statistical discrepancy. The rationale for these QRPs is that they all relate to quality of the study and quality of reporting, which is seen as an essential element of responsible research.\textsuperscript{13,22} We focused on demographic and bibliometric indicators, including characteristics of the author team, trial / publication and journal, available during different phases of a project: during trial registration, when a study is submitted for publication, and after a study is published.
2. Methods

A protocol for this study has been made publicly available on the Open Science Framework on December 19th, 2018, before start of data collections. Deviations from the protocol are described in Appendix 1.

2.1 Identification of RCTs

We searched PubMed using the Entrez API (see https://www.ncbi.nlm.nih.gov/home/develop/api/) via R Statistical Software on November 17th, 2017 to identify studies with publication type randomized controlled trial and automatically excluded non-randomized, animal, pilot, and feasibility studies (see Appendix 1). Additionally, articles were excluded when the language was other than English. Articles published before 1996 were excluded because in that year the CONSORT Statement was published aiming to enhance the completeness of reporting of RCTs.

We developed web scrapers to automatically download the PDF of each identified RCT via the website of the respective publisher. Downloaded PDFs were transformed to text data in Extensible Markup Format (XML), using GROBID software.

2.2 Data collection of QRPs

We assessed the following four QRPs (Box 1):

1. Risk of bias, the probability of bias as determined using RobotReviewer for the domains random sequence generation, allocation concealment, blinding of participants and personnel, and blinding of outcome assessment.

2. Modifications in primary outcome measures based on comparing first and final versions of the public trial registration records from ClinicalTrials.gov.

3. The ratio of achieved sample size compared to what was planned.
4. Statistical discrepancy, for which we compared the reported p-value and actual p-value of the intervention effect estimate calculated from other reported information such as the confidence interval.

Trial registry numbers were collected by searching abstracts and full texts using regular expressions (i.e. sequences of characters that specify a search pattern) and we subsequently obtained public trial registration records from ClinicalTrials.gov.

2.3 Data collection of indicators

Potential indicators of QRPs were selected based on previous evidence, discussions with experts, availability, and feasibility. They are listed in Box 2 and included characteristics of the 1) author team (e.g., gender, number of authors), 2) publication (e.g., reporting of trial registration), and 3) journal (e.g., impact factor). Data were automatically extracted from information indexed in PubMed (e.g., authors, affiliations etc.) and from the full-text article as XML. Using the PubMed ID, RCTs were linked to Scopus and additional information on characteristics of author teams (e.g. Hirsch-index) was obtained. Detailed descriptions of definitions and methods for outcomes and indicators are described in Appendix 1.

Box 1: Methods for collecting information on questionable research practices

<table>
<thead>
<tr>
<th>Risk of bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk of Bias domains were extracted via open source software provided by RobotReviewer. RobotReviewer is developed to score bias for four domains of the Cochrane Risk of Bias tool: random sequence generation, allocation concealment, blinding of participants and personnel, and blinding of outcome assessment. RobotReviewer assesses the probability that a study has bias, rather than dichotomizing it into high or low risk of bias. Level of agreement between RobotReviewer and human raters was similar for most domains (average human–human</td>
</tr>
</tbody>
</table>
agreement 79% (range 71% to 85%), human–RobotReviewer agreement 65% (range 39% to 91%).

Modifications in primary outcomes
Changes made to the primary outcome after the trial had started, as reported in the trial registration on ClinicalTrials.gov. Changes were first automatically extracted by comparing the first and final version of the primary outcome as registered in the study protocols on clinicaltrials.gov. Additions and deletions of complete outcome measures were extracted. The algorithm was too sensitive for changes in the content: if any textual change was present, the primary outcome was flagged as ‘changed’. These flagged studies were subsequently manually checked to distinguish between significant and insignificant (e.g. typo’s) changes.

Ratio of achieved sample size compared to what was planned
We calculated the ratio of actual sample size and planned sample size based on the power calculation provided in the public trial registration records from ClinicalTrials.gov. This information could be extracted directly from the trial registration record. A manual check was performed for all publications and protocols where the ratio of the number of enrolled and estimated participants was >100, i.e. 100 times more enrolled than was estimated.

Statistical discrepancy
Comparison of reported p-value and actual p-value of the intervention effect estimate calculated from other reported information. Based on the reported relative risk, odds ratio or hazard ratio in combination with its 95% confidence intervals, the p-value was recomputed. This value was compared with the reported p-value using a script by Georgescu & Wren. For t-tests, Chi-square values, F-values z-statistics and correlations, the R-package StatCheck was used to check the correct reporting of the p-value. Inconsistent p-values (defined as a difference ≥0.01) were marked. Every inconsistency where the adjusted p-value crosses the level of 0.05 compared to the original p-value was labelled as statistical discrepancy.

Box 2: Collected demographic and bibliometric indicators (more details can be found in Appendix 1)

Author team
- Gender of first and last author (https://genderize.io/)
- Proportion of female authors in the author team
- Total number of authors\textsuperscript{17,30}
- Continent of first and last author\textsuperscript{18,19,31}
- Number of countries to which the author team is affiliated
- Hirsch-index of first and last author in the year before publication\textsuperscript{10,17,30}
- Academic age of first and last author (i.e., number of years between the trial publication and first publication by this author)\textsuperscript{17,19,32}
- Uninterrupted presence of first and last author (i.e., the number of years the author has published at least one article in sequentially without interruption)\textsuperscript{19}
- Number of collaborations of the first and last author (i.e., total number of coauthorships until year of publication)
- Number of institutions represented in the author team\textsuperscript{18}
- Ranking of institution of first and last author in the Academic Ranking of World Universities (www.shanghairanking.com)

**Trial / publication**
- Trial registration
- Financial support (industrial, other, none)\textsuperscript{17,18,31,33}
- Year of publication
- Conflict of interest
- Mentioning of the CONSORT Statement
- Positive and negative word frequencies in abstract\textsuperscript{34}
- Number of words and number of names mentioned in acknowledgments

**Journal**
- Medical field\textsuperscript{18,33}
- Journal impact factor in the year before publication\textsuperscript{33,35,36}
- Impact factor change compared to previous year
- Number of publications of the journal per year
- Journal publisher
- Continent of journal

2.4 Statistical analyses
Detailed analyses are described in Appendix 1. In short, associations between indicators and outcomes were assessed using univariable and multivariable regression models. Three multivariable regression models were fitted per outcome: 1) a full model including all indicators (Box 2); 2) a reduced model including indicators available upon journal submission of an article but before publication; and 3) a reduced model including indicators available upon trial design and registration but before the trial is completed. Indicators in the model were selected based on a priori group discussions on the relevance of the indicators. We used beta regression models (R package ‘betareg’), logistic regression (R package ‘rms’) for modifications in primary outcomes and statistical discrepancy, and linear regression (R package base) for the log-transformed ratio of achieved compared to planned sample size. For all multivariable models, one indicator from an indicator pair was excluded based on discussions between authors if there was multicollinearity (i.e., Spearman correlation >0.8). Indicators with more than 40% missing values were excluded from analyses. For the other indicators and QRPs, missing values were imputed twenty times using Multiple Imputation by Chained Equations. Data transformations were applied if required and a Bonferroni correction for multiple testing was applied. Goodness of fit was assessed in terms of explained variance (i.e., $R^2$).
3. Results

3.1 Study flow

The search identified 445,159 records of which 138,422 were excluded automatically because they were likely not describing an RCT (Figure 1). After excluding references for which we could not obtain the full text (n=122,810) and that were published before 1996 or with an unclear publication year (n=20,798), we included 163,129 in the analyses for each of the four probability of bias outcomes.

For ratio of achieved compared to planned sample size and modifications in primary outcomes, we excluded additional references due to the unavailability of registration on ClinicalTrials.gov. For statistical discrepancy references were excluded because no combination of p-value and test statistic could be identified, leaving 21,230 references.

3.2 Components of Questionable Research Practices (Table 1)

The median probabilities of bias ranged between 43% (IQR 18%-59%) for randomization and 63% (IQR 40%-75%) for blinding of patients and personnel. 22% (95% CI 21%-23%) of studies had modified their primary outcome and we found a median ratio of achieved compared to planned sample size of 1 (IQR 0.98-1.04). In 370 out of 21,230 publications (1.7% [95% CI 1.6%-1.9%]), we identified statistical discrepancy.

Table 1: Descriptive statistics of Questionable Research Practices

<table>
<thead>
<tr>
<th>Questionable Research Practice</th>
<th>Value*</th>
<th>Number of references for which this outcome was available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of bias (as assessed by Robot Reviewer**)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probability of bias in randomization</td>
<td>0.43 (0.18-0.59)</td>
<td>163,129</td>
</tr>
<tr>
<td>Probability of bias in allocation concealment</td>
<td>0.59 (0.40-0.71)</td>
<td>163,129</td>
</tr>
<tr>
<td>Questionable Research Practice</td>
<td>Value*</td>
<td>Number of references for which this outcome was available</td>
</tr>
<tr>
<td>--------------------------------------------------------------------</td>
<td>-------------------</td>
<td>-----------------------------------------------------------</td>
</tr>
<tr>
<td>Probability of bias in blinding of patients and personnel</td>
<td>0.63 (0.40-0.75)</td>
<td>163,129</td>
</tr>
<tr>
<td>Probability of bias in blinding of outcome assessment</td>
<td>0.55 (0.44-0.64)</td>
<td>163,129</td>
</tr>
<tr>
<td>Modifications in primary outcome in public registration</td>
<td>3615/16349 (22.1% [95% CI 21.5-22.8])</td>
<td>16,349</td>
</tr>
<tr>
<td>Ratio of achieved compared to planned sample size</td>
<td>1 (0.98-1.04)</td>
<td>24,385</td>
</tr>
<tr>
<td>Statistical discrepancy</td>
<td>370/21230 (1.7% [95% CI 1.6-1.9])</td>
<td>21,230</td>
</tr>
</tbody>
</table>

*Values are N (% [95% CI]) or median (25th – 75th percentile). **Robot Reviewer assesses the probability that a study has bias, rather than dichotomizing it into high or low risk of bias. We here present the median probabilities. See methods section for definitions of Questionable Research Practices.

3.3 Demographic and bibliometric indicators (Appendix 2)

The majority of the included publications had a male first author (61.8% [95% CI 61.6%-62.1%]) and a male last author (73.6% [95% CI 73.4%-73.8%]), with a median of 33% (IQR 17%-50%) female authors. Author teams included a median of 6 (IQR 5-9) authors.

The most frequent medical discipline was general medicine (10.0% [95% CI 9.9%-10.2%]), 12.8% (95% CI 12.6%-13.0%) of publications mentioned the word CONSORT, and for 28.8% (95% CI 28.6%-29.1%) we identified a trial registration number. Articles were published in journals with a median impact factor of 2.93 (IQR 1.99-4.41).

The indicators ranking of the institution of the first author, ranking of the institution of the last author, financial support, number of words in the acknowledgment, and number of names in the acknowledgment were excluded from further analyses because of the large amount of missing data.
3.4 Univariable analyses

Results of univariable analyses are presented in Appendix 2. None of the indicators showed a statistically significant consistently positive or negative association for every type of QRP.

3.5 Multivariable models with data available from the trial publication

The indicators continent of first author, academic age of first and last author, academic presence of first author, and number of collaborations of first and last authors were excluded from multivariable models due to high correlations with other indicators in the model.

3.5.1 Risk of bias

In the multivariable models (Appendix 2), the following indicators were found to be associated with a lower probability of bias for at least three out of four domains: a higher proportion of female coauthors, publications with the last author from Oceania, a more recent publication year, reporting a trial registration number, mentioning of CONSORT, higher journal impact factor, and publications from a large publisher. Publications with the last author from North America were associated with a higher probability of bias than publications with the last author from Europe. Compared to the category of general medicine, many medical disciplines were associated with either consistently higher (e.g., hematology) or lower (e.g., anesthesiology) probability of bias.

3.5.2 Modifications in the primary outcome

Publications with the last author from North America or Oceania had a higher risk of modifications in the outcome than publications with the last author from Europe. Also, a higher h-index of the first and last authors and having more institutions involved were associated with a higher risk.

3.5.3 Ratio of achieved compared to sample size

A higher number of countries involved was associated with a higher ratio of achieved compared to planned sample size (i.e., higher achieved sample size). Having more institutions involved was associated with a lower ratio.

3.5.4 Statistical discrepancy
Publications reporting a trial registration number were associated with a lower risk of statistical discrepancy. We found conflicting associations or found no associations with consistent directions over multiple QRP\'s for the research experience of the last author (i.e., active research years), use of positive or negative words in abstracts, or changes in journal impact factors.

3.6 Multivariable models restricted to data available upon submission to a journal (i.e. before trial publication)
Models that contain indicators available upon submission of an article to a journal but before publication (model 2) showed similar trends to the models with post-publication indicators (Table 2 and Appendix 2). Again, a higher proportion of female authors was associated with a lower probability of bias (except for the domain blinding of participants and personnel where the reverse was found). The h-index of first and last authors was associated with a higher risk of primary outcome modifications in the public registration. Studies that mentioned CONSORT and reported a trial registration number were consistently associated with a lower probability of bias.

Differences with the full models were that a higher number of authors was associated with a lower probability of bias. The h-index of the first author was associated with a lower probability of bias in all four domains in the reduced models but not in the full models.

3.7 Multivariable models with data available upon trial registration
The models that contained indicators available upon trial registration (but before trial completion) only included the indicators gender of last author, the continent of last author, h-index of last author, academic age of last author, and medical discipline. For the four domains, almost all of these indicators were associated with probability of bias (Appendix 2).
### Table 2: Results from multivariable reduced models that include indicators of QRPs available upon submission of an article but before publication (model 2)

<table>
<thead>
<tr>
<th>Continent of last author - Africa</th>
<th>Bias in randomization</th>
<th>Bias in allocation concealment</th>
<th>Bias in binding of patients and personnel</th>
<th>Bias in binding of outcome assessment</th>
<th>Probability of bias</th>
<th>Probability of bias outcome</th>
<th>Ratio of achieved compared to target sample size</th>
<th>Statistical discrepancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continent of last author - Asia</td>
<td>-0.144 (-0.200; -0.088)</td>
<td>-0.053 (-0.103; -0.003)</td>
<td>-0.071 (0.019; 0.123)</td>
<td>-0.047 (-0.080; -0.014)</td>
<td>-0.105 (-0.602; 0.391)</td>
<td>-0.107 (-0.090; 0.055)</td>
<td>-0.847 (-0.495; 2.189)</td>
<td>-0.156 (-0.439; 0.751)</td>
</tr>
<tr>
<td>Continent of last author - Middle and South America</td>
<td>0.063 (-0.008; 0.135)</td>
<td>0.056 (-0.004; 0.115)</td>
<td>-0.073 (-0.139; -0.006)</td>
<td>-0.068 (-0.105; -0.026)</td>
<td>0.089 (-0.350; 0.529)</td>
<td>0.013 (-0.053; 0.080)</td>
<td>-0.605 (-3.021; 1.812)</td>
<td>0.064 (-0.883; 0.290)</td>
</tr>
<tr>
<td>H-index of first author</td>
<td>0.001 (0.000; 0.001)</td>
<td>0.000 (-0.000; 0.001)</td>
<td>-0.001 (0.000; 0.001)</td>
<td>-0.000 (0.000; -0.000)</td>
<td>0.005 (0.001; 0.009)</td>
<td>0.001 (-0.000; 0.001)</td>
<td>-0.002 (-0.016; 0.012)</td>
<td>0.064 (-0.111; 0.240)</td>
</tr>
<tr>
<td>Academic age of last author - sqrt</td>
<td>0.002 (-0.007; 0.010)</td>
<td>-0.000 (-0.007; 0.007)</td>
<td>0.018 (0.010; 0.025)</td>
<td>0.009 (0.004; 0.014)</td>
<td>-0.039 (-0.097; 0.018)</td>
<td>0.000 (-0.009; 0.010)</td>
<td>-0.017 (-0.206; 0.173)</td>
<td>0.064 (-0.883; 0.290)</td>
</tr>
<tr>
<td>Percentage of positive words in abstract</td>
<td>0.036 (-0.019; 0.090)</td>
<td>0.043 (-0.004; 0.089)</td>
<td>0.063 (0.012; 0.112)</td>
<td>0.017 (-0.015; 0.048)</td>
<td>-0.125 (-0.533; 0.284)</td>
<td>-0.015 (-0.075; 0.045)</td>
<td>0.355 (-4.00; 1.109)</td>
<td>0.064 (-0.883; 0.290)</td>
</tr>
<tr>
<td>Medical discipline*</td>
<td>See Appendix 2</td>
<td>See Appendix 2</td>
<td>See Appendix 2</td>
<td>See Appendix 2</td>
<td>See Appendix 2</td>
<td>See Appendix 2</td>
<td>See Appendix 2</td>
<td>See Appendix 2</td>
</tr>
<tr>
<td>Percentage of negative words in abstract</td>
<td>0.111 (-0.245; 0.290)</td>
<td>-0.058 (-0.157; -0.046)</td>
<td>0.110 (-0.233; -0.087)</td>
<td>0.110 (-0.233; -0.087)</td>
<td>-0.096 (-0.018; 0.050)</td>
<td>-0.096 (-0.018; 0.050)</td>
<td>-0.096 (-0.018; 0.050)</td>
<td>-0.096 (-0.018; 0.050)</td>
</tr>
<tr>
<td>Trial registration</td>
<td>-0.437 (-0.462; -0.412)</td>
<td>-0.417 (-0.438; -0.395)</td>
<td>-0.155 (-0.178; -0.133)</td>
<td>-0.193 (-0.207; -0.178)</td>
<td>Not applicable**</td>
<td>Not applicable**</td>
<td>Not applicable**</td>
<td>-0.685 (-1.260; -0.110)</td>
</tr>
</tbody>
</table>

All values are regression coefficients from multivariable models for 1 unit increase in the indicator. For all outcomes, except for the ratio of achieved compared to planned sample size, negative values are good, i.e. less questionable and more responsible (e.g. lower risk of bias). Statistically significant values are marked in blue (lower risk of QRP) and red (higher risk of QRP). For all categorical variables regarding the continent of authors or journal, Europe is taken as the reference category. Sqrt: square root.

* The indicator ‘medical discipline’ was included in the model but removed from the table to improve readability.

** The indicator having a trial registration could not be included in the models predicting modifications in the outcome and ratio of achieved compared to planned sample size, as these outcomes were only available for trials that have a trial registration.
3.8 Explained variance

In terms of explained variance the reduced models had lower values than the full models (Appendix 2). The highest $R^2$ values were seen for full models predicting bias in allocation concealment and bias in randomization (0.138 and 0.122, respectively). The lowest $R^2$ was found for the reduced model, using data available during trial design and registration, predicting ratio of achieved compared to planned sample size ($R^2$ 0.002).
4. Discussion

We investigated the association between trial characteristics and QRPs and found associations with QRPs for many of the studied indicators (e.g., gender, publication year, h-index, mentioning of CONSORT). The most robust indicators that were consistently associated with a lower risk of several QRPs included: 1) a higher journal impact factor, 2) a journal from a large publisher (such as Elsevier or Springer), 3) having a trial registration, and 4) mentioning of the CONSORT reporting guideline. We could not identify any association between the percentage of positive or negative words in an abstract and the risk of QRP.

4.1 Comparison to previous literature

Several researchers mapped the frequency of QRPs. In our study, we observed that p-values did not correspond to the given test statistics in 1.7% of the articles. This is similar to a previous publication in which inconsistencies were observed in 1.6% of studies. It is, however, lower than a study that found statistical discrepancy in 38% of articles published in 2001 in Nature and 25% in the BMJ. A possible explanation for these differences is that p-values were manually collected and checked in that study while we made use of an automated script which might have missed large parts of the p-value test-statistic combinations while a manual check is not restricted to this specific type of format and therefore could identify more of these combinations.

We found that gender (higher proportion of female authors) was associated with lower probability of bias. Previous research has shown that female authors tend to report more conservative effect sizes, but also that female first authors are more likely to overestimate effects. A recent survey conducted amongst Dutch academics has shown that a lower academic rank and a female gender were associated with a lower RRP score (i.e. less responsible). Many studies have focused on the association between impact factor and QRPs. In agreement with our findings, higher impact factors were found to be associated with a lower probability of bias, but also with better reporting.
Surprisingly, we found that a higher h-index was associated with a higher risk of primary outcome modifications. We hypothesize this might be related to the fact that h-index is partly driven by the number of publications and therefore, more experienced researchers often have a higher h-index. In the past years, there has been a change in research culture, with more attention to responsible conduct or research.48

4.2 Recommendations for future research

Although it is not possible to draw conclusions about causal relations based on our study, our results might inform future strategies to identify those RCTs at high risk of QRPs. In this explorative study, we showed there are associations between indicators and the presence of QRPs. However, the low explained variance of our regression models suggests these cannot be used for individual risk predictions. Furthermore, this suggests there is still a lot of variation between studies that could not be explained by the indicators we studied. A future step could be to study more indicators from other QRP domains to inform a prediction model that can be applied to flagging trial protocols, manuscripts or articles with a high risk of QRPs which need to be scrutinized more closely. It should also be noted that such a prediction model should not be used on its own, but always combined with further (manual) examination of the existence of QRPs.

Two indicators that consistently showed associations with a lower probability of bias across all four studied QRPs are reporting a trial registration number and mentioning CONSORT in a manuscript, which both relate to strategies aimed at enhancing usability of study results. This confirms the importance of requiring trial registration and complying with reporting guidelines.

Surprisingly, we found that a higher number of countries was associated with a higher ratio of achieved compared to planned sample size, while a higher number of institutions was associated with a lower ratio. Furthermore, having a last author from Oceania was associated with a lower probability of bias and a higher risk of modifying outcomes. Further research can focus on finding out
whether these associations can be confirmed independently, or whether they are just chance findings.

4.3 Strengths and limitations

We evaluated QRPs in RCTs covering a large proportion of all published RCTs included in PubMed. Using automated data collection, we were able to obtain a large amount of data. Our analysis also has several limitations. Firstly, we have not manually screened all in- and excluded articles. This allowed us to include a large number of RCTs, but it is possible that we have included articles that do not report an RCT, that we have included multiple publications about the same RCT, and that we have excluded articles that were reporting an RCT. Especially, poorly written articles were more likely to be misclassified. Articles for which no PDF was available had to be excluded, which might have led to a selective set of RCTs included in our analyses. For the QRP related to selective reporting of outcomes, we restricted ourselves to RCTs registered on ClinicalTrials.gov while many European trials are only registered in European Union Drug Regulating Authorities Clinical Trials Database. This might have led to selective exclusion of European RCTs for this QRP. Secondly, the automated data collection might have led to misclassification of indicators and QRPs. We expect that this has diluted the associations. Due to the high amount of missing data, caused by problems with automatic data collection, we, had to exclude five of our predefined indicators. Furthermore, only 21,230 articles were available for evaluating statistical discrepancy because in the other articles we were not able to identify a p-value test-statistic combination in the required format. For the outcome risk of bias, we relied on RobotReviewer software. Evaluations of this tool indicated moderate to good agreement with human reviewers for the random sequence generation and allocation concealment domains, however, a varying agreement was found for the domains on blinding.\textsuperscript{3 20 49}
Thirdly, we planned to collect information on the quality of reporting, defined as adherence to the CONSORT reporting guideline as determined with software developed by StatReviewer, but this turned out not possible due to time constraints of the software developers. Lastly, although we applied a Bonferroni correction, we still tested hundreds indicator-outcome associations. Furthermore, the large size of our dataset might have resulted in statistically significant but irrelevant associations (i.e. small effect size).

5. Conclusion

Our analyses show that gender, author continent, publication year, h-index, mentioning of CONSORT, trial registration, medical discipline, and journal impact factor were all associated (in different directions) with the risk of QRPs.
Data Availability Statement: The risk of bias characterization was done with a large-batch customized Python scripts (version 3; https://github.com/wmotte/robotreviewer_prob). The data management and analyses used R (version 3.6.1). All data are available at https://github.com/wmotte/RCTQuality).

Funding: This work was supported by The Netherlands Organisation for Health Research and Development (ZonMw) grant “Fostering Responsible Research Practices” (445001002) (CV, JT, WO). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declarations of interest: none.
References


39. Harrell FE. Package ‘rms’. *Vanderbilt University* 2019

40. Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *Journal of statistical software* 2011;45(3)


Records identified through PubMed searching (n=445,159)

- Records excluded (n=138,422)
  - 21,382 no abstract
  - 67,169 not ‘random’ or ‘assign’ etc.
  - 3,147 study protocol
  - 9,084 pilot study
  - 17,979 not English
  - 19,369 published before 1988
  - 292 published after 2017

RCTs identified (n=306,737)

- Records excluded (n=122,810)
  - 82,560, no institutional license
  - 24,549 full text not successfully downloaded
  - 15,701 no conversion to XML

Full text articles obtained (n=183,927)

- Records excluded (n=20,798)
  - 13,491 published before 1996
  - 7,307 publication year unclear

Full text articles included (n=163,129)

Per outcome:
- Probability of bias (n=163,129)
- Outcome modifications (n=16,349)
- Ratio of sample size (n=24,385)
- Statistical discrepancy (n=21,230)
What is new?

- We included 163,129 randomized controlled trial publications and automatically extracted information about questionable research practices and potential indicators.
- Previously identified associations between indicators and questionable research practices were validated and new indicators were explored.
- A more recent publication year, trial registration, mentioning of CONSORT-checklist, and a higher journal impact factor were consistently associated with a lower risk of questionable research practices.
- Editors, peer reviewers and readers should be aware that certain characteristics of the author team, the journal and the publication might be associated with questionable research practices.
Declarations of interest: none.
Johanna A Damen: Conceptualization, Methodology, Formal analysis, Data curation, Writing - Original Draft. Pauline Heus: Conceptualization, Methodology, Writing - Review & Editing. Herm J Lamberink: Conceptualization, Methodology, Data curation, Writing - Review & Editing. Joeri K Tijdink: Conceptualization, Methodology, Writing - Review & Editing, Funding acquisition. Lex Bouter: Conceptualization, Methodology, Writing - Review & Editing. Paul Glasziou: Conceptualization, Writing - Review & Editing. David Moher: Conceptualization, Writing - Review & Editing. Willem M Otte: Conceptualization, Methodology, Data curation, Writing - Review & Editing. Christiaan H Vinkers: Conceptualization, Methodology, Writing - Review & Editing, Supervision, Funding acquisition. Lotty Hooft: Conceptualization, Methodology, Writing - Review & Editing, Supervision.