Journal of Clinical Epidemiology 136 (2021) 189–202

Journal of Clinical Epidemiology

REVIEW

Methods to assess research misconduct in health-related research: A scoping review

Esmee M Bordewijk[a,b], Wentao Li[b,*], Rik van Eekelen[a], Rui Wang[b], Marian Showell[c], Ben W Mol[b], Madelon van Wely[a]

[a] *Centre for Reproductive Medicine, Amsterdam UMC, Amsterdam, The Netherlands*
[b] *Department of Obstetrics and Gynecology, Monash University, Clayton, Australia*
[c] *Department of Obstetrics and Gynaecology, University of Auckland, Auckland, New Zealand*

Accepted 12 May 2021; Available online 24 May 2021

Abstract

Objective: To give an overview of the available methods to investigate research misconduct in health-related research.

Study Design and Setting: In this scoping review, we conducted a literature search in MEDLINE, Embase, The Cochrane CENTRAL Register of Studies Online (CRSO), and The Virtual Health Library portal up to July 2020. We included papers that mentioned and/or described methods for screening or assessing research misconduct in health-related research. We categorized identified methods into the following four groups according to their scopes: overall concern, textual concern, image concern, and data concern.

Results: We included 57 papers reporting on 27 methods: two on overall concern, four on textual concern, three on image concern, and 18 on data concern. Apart from the methods to locate textual plagiarism and image manipulation, all other methods, be it theoretical or empirical, are based on examples, are not standardized, and lack formal validation.

Conclusion: Existing methods cover a wide range of issues regarding research misconduct. Although measures to counteract textual plagiarism are well implemented, tools to investigate other forms of research misconduct are rudimentary and labour-intensive. To cope with the rising challenge of research misconduct, further development of automatic tools and routine validation of these methods is needed.

Trial registration number: Center for Open Science (OSF) (https://osf.io/mq89w).

Keywords: Data integrity; Research misconduct; Scientific misconduct; Randomization; Methods; Scoping review

Funding: This work was supported by Elsevier. Elsevier does not have a role in study design; in the collection, analysis and interpretation of data; in the writing of the manuscript; or in the decision to submit for publication.

Conflict of interest: MS is the information specialist of the Cochrane Gynaecology and Fertiliy Group. BWM is supported by a National Health Medical Research Council (NHMRC) Practitioner Fellowship (GNT1082548). BWM reports consultancy for ObsEva, Merck and Guerbet. MvW is coordinating editor of the Netherlands Satellite of the Cochrane Gynaecology and Fertiliy Group. All other authors have no conflicts of interest.

[*] Corresponding author. Tel.:+61 (0)3 8572 3948
E-mail address: wentao.li@monash.edu (W. Li).

https://doi.org/10.1016/j.jclinepi.2021.05.012
0895-4356/© 2021 Elsevier Inc. All rights reserved.

What is new?

Key findings

- We found 27 methods in the literature that assess research misconduct in health-related research. While the methods to detect textual plagiarism have been widely implemented, most other methods have not been adequately validated nor have they been structurally implemented.

What this adds to what is known?

- This scoping review systematically summarized reported methods that detect research misconduct in health-related research and analysed their applicability.

What is the implication, what should change now

- There exist methods that can be tested by the scientific community to proactively defend the integrity of research before publication. More efforts are needed to further develop, validate, or automate these methods to promote their routine use in academic publishing.

1. Introduction

Science relies on the integrity of findings that are reported. It was found that about 2% of scientists admitted to having fabricated, falsified, or modified data or results at least once and on average over 14% of scientists observed these behaviours among their colleagues [1]. Research misconduct may result in a waste of financial and human resources and, more importantly, it might pose an immediate risk to human health [1].

The US Office of Research Integrity (ORI) defines research misconduct as fabrication [1], falsification [2], or plagiarism [3] in proposing, performing, or reviewing research, or in reporting research results [2]. Fabrication is making up data, results, or recordings, and reporting them. Falsification is manipulating research materials, equipment, or processes, or changing or omitting data or results such that the research is not accurately represented in the research record. Plagiarism is the appropriation of another person's ideas, processes, results, or words without giving appropriate credit. Research misconduct does not include honest error or differences of opinion [2].

In case of suspected misconduct, according to the Committee on Publication Ethics (COPE) code of conduct [3], editors have the duty to take action [4]. However, only a third of top-ranking peer-reviewed journals have publicly available definitions of misconduct and less than half describe editorial procedures for handling allegations of misconduct [5]. Admittedly, investigating research misconduct is usually not straightforward, and therefore dealing with possible misconduct is not an easy task. Failure to adequately investigate possible misconduct may perpetuate unreliable research findings in the literature. When researchers who commit fraud go unchecked, they may continue to practice misconduct [4].

Methods that investigate research misconduct accumulate and evolve. However, to date, there is no complete overview of these methods and their applicability. Here, we reviewed the literature for articles that mention, describe, validate, or apply methods for screening or assessing research misconduct in health-related research.

2. Methods

The protocol of this scoping review is registered in the Center for Open Science (OSF) on July 14, 2020 (https://osf.io/mq89w). We followed the reporting guidelines for meta-analyses and systematic reviews extension for scoping reviews, as outlined by the PRISMA statement [6].

2.1. Literature search

A comprehensive and systematic literature search was undertaken in MEDLINE, Embase, The Cochrane CENTRAL Register of Studies Online (CRSO), and The Virtual Health Library for reports up to the July 14, 2020 by an information specialist (MS, Appendix 1). To identify any additional studies, we scanned reference lists of appropriate reports and communicated with experts in this field. All references were imported in Covidence (covidence.org). There was no language restriction or date restriction, but we excluded conference abstracts.

2.2. In- and exclusion criteria and study selection

Studies that refer to methods to investigate research misconduct, i.e., fabrication, falsification, and/or plagiarism in health-related research, were eligible for this scoping review. We excluded editorials, education plagiarism tools, and studies on data audits, meta-data, peer-review, and p-hacking as these methods are not directed at detecting research misconduct.

Two review authors (EB and MvW) independently screened all records on basis of titles and abstracts. After the eligibility screening, we critically reviewed the full text of the selected studies to assess eligibility. Any discrepancies between the reviewers were solved by consensus.

2.3. Data extraction and categorization

We used a data charting XLS sheet developed by EB with the help of MvW. Data were extracted by EB and checked by MvW for the non-statistical papers, and WL or RvE for the statistical papers. We extracted any method provided concerning research misconduct.

From each included study we extracted information regarding author, year of publication, journal, and the method. For each method, if applicable, we recorded the link to the method, description on how to use the method, information needed, whether qualitative or quantitative, validation method, automatic application, and performance if available. We categorised identified methods into the following four groups according to their scopes: overall concern, textual concern, image concern, and data concern. We did not perform a critical assessment because there was insufficient information to support a fair critical appraisal of the identified methods.

3. Results

3.1. Literature search

We identified 6,112 articles (Fig. 1). After removing duplications, we screened 4,956 articles, of which 4,750 irrelevant articles were excluded (proportionate agreement between reviewers was 0.91). After assessing the full text of 206 articles, we excluded another 149 (proportionate agreement between reviewers was 0.83). Therefore, 57 papers were included in this review.

3.2. Included studies

The included papers reported on 27 different methods, two on overall concern [7–18], four on textual concern [14–16,19–29], three on image concern [30–33], and 18 on data concern [7,10-13,15–18,34–63]. The characteristics of the articles are in Table 1. The following sections briefly explain the methods and their rationale. Table 2 expresses the applicability of the available methods. Available software links and programs can be found in Table 3, with further details on how to use the statistical methods described in Appendix 2.

4. Overall concern

4.1. Screening

The “REAPRAISSED checklist” for evaluation of publication integrity is a screening tool to assess whether a paper has characteristics that question its trustworthiness [8]. The checklist facilitates systematic evaluation through 11 categories. It covers ethical oversight and funding, research productivity and investigator workload, validity of randomization, plausibility of results, and duplicate data reporting.

4.2. Detection of patterns of misconduct in all publications of one author/group

When a fraudulent research paper is discovered, it is reasonable to assume that there may be similar problems with previous works of the authors involved [9]. Some patterns of research misconduct that are unique to the leading author/group can only be identified when all relevant works are compared, such as copying data of the group’s previous works and overlapping publications. Also, comparing conference posters or abstracts, research grants, and protocols of one author or author group can be useful in the detection of research misconduct [17].

5. Textual concern

Methods that detect textual concern are summarised in Appendix 3. Methods for anti-textual plagiarism have been widely implemented.

6. Image concern

ORI offers Forensic Image Analysis Tools to detect data image manipulation in the field of biomedicine, especially Western Blots [30].

Koppers, Wormer [31] created a tool that uses mathematical methods to detect suspicious images in large image archives, the R package called “FraudDetTools”. The tool can detect manipulation by deleting unwanted data information, duplication by reusing images in different papers or contexts, and manipulation by adding information/data points.

Acuna, Brookes [32] created a tool that analyses potential inappropriate reuse of images. This algorithm detects figure region reuse and is robust to rotation, cropping, resizing, and contrast changes, and estimates which of the reuses have biological meaning.

For all these algorithm-based tools, the final decision should always be made by human experts to avoid false positives.

7. Data concern

Methods to check the authenticity of the data are directed at the given statistical results and the original raw data. Some of the methods described in the following sections are sufficiently complicated that to apply them, we refer the readers to the original papers.

7.1. Statistics check

Reported statistical results can be reproduced with summary statistics in publications. Inconsistencies may be explained by data fabrication or falsification as well as other possible reasons such as honest error. We found four software packages: Statcheck, the GRIMMER test, SPRITE, and the R package rpsychi.

The free *Statcheck* [57–59] software extracts statistical values reported in the text. For each extracted statistical test result, the reported statistical values are used to recalculate the *P*-value for the reported statistical result. Recalculated *P*-values are checked against the reported *P*-values

Table 1. Characteristics of the included studies

No. in reference list	Study	Journal/source	Title	Method
8	Grey et al. 2020	Nature	Check for publication integrity before misconduct	Overall concern: Screening (REAPRAISSED checklist)
9	Smith 2005	BMJ	Investigating the previous studies of a fraudulent author.	Overall concern: Investigating all publications of one author
24	Errami et al. 2007	Nucleic Acids Research	eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications.	Textual concern: Textual plagiarism (Helioblast)
26	Errami et al. 2008	Bioinformatics	Déjà vu—A study of duplicate citations in Medline.	Textual concern: Textual plagiarism (Helioblast)
25	Errami et al. 2010	Bioinformatics	Identifying duplicate content using statistically improbable phrases.	Textual concern: Textual plagiarism (Helioblast)
27	Garner 2012	Nature	How to stop plagiarism.	Textual concern: Textual plagiarism (Helioblast)
28	Higgings et al. 2016	Research integrity and peer review	Plagiarism in submitted manuscripts: incidence, characteristics and optimization of screening-case study in a major specialty medical journal.	Textual concern: Textual plagiarism (iThenticate)
29	Taylor 2017	American Roentgen Ray Society	Plagiarism in Manuscripts Submitted to the AJR: Development of an Optimal Screening Algorithm and Management Pathways	Textual concern: Textual plagiarism (iThenticate)
16	Bordewijk et al. 2020a	European Journal of Obstetrics & Gynecology and Reproductive Biology	Data integrity of 35 randomized controlled trials in women' health.	Overall concern: Investigating all publications of one author Textual concern: Compare baseline characteristics and outcome tables Data concern: Baseline P value distribution for RCTs & Digit preference checks
15	Bordewijk et al. 2020b	Fertility and Sterility Dialog	Data integrity of 10 other randomized controlled trials of an author with a retracted paper.	Overall concern: Investigating all publications of one author Textual concern: Compare baseline characteristics and outcome tables Data concern: Baseline P value distribution for RCTs
22	Baydik et al. 2016	Journal of Korean medical science	How to Act When Research Misconduct Is Not Detected by Software but Revealed by the Author of the Plagiarized Article.	Textual concern Translated plagiarism
23	Wiwanitkit 2016	Macedonian Journal of Medical Sciences	How to verify and manage the translational plagiarism?	Textual concern Translated plagiarism
14	Spiroski 2016	Open Access Macedonian Journal of Medical Sciences	How to verify plagiarism of the paper written in Macedonian and translated in foreign language?	Overall concern: Investigating all publications of one author Textual concern Translated plagiarism
19	Bohannon 2015	Science	Scientific publishing. Hoax-detecting software spots fake papers.	Textual concern: Automatically generated fake papers (Scidetect)
20	Nguyen et al. 2016	BIR 2016 Workshop	Engineering a Tool to Detect Automatically Generated Papers.	Textual concern: Automatically generated fake papers (Scidetect)
21	Springer et al. 2015	Springer press release	Springer and Université Joseph Fourier release SciDetect to discover fake scientific papers	Textual concern: Automatically generated fake papers (Scidetect)
30	ORI	ORI	https://ori.hhs.gov/forensic-tools	Image concern
33	Parrish et al. 2009	Science and Engineering Ethics	Image manipulation as research misconduct.	Image concern

(continued on next page)

Table 1 (*continued*)

No. in reference list	Study	Journal/source	Title	Method
31	Koppers et al. 2017	Science and engineering ethics	Toward a Systematic Screening Tool for Quality Assurance and Semiautomatic Fraud Detection for Images in the Life Sciences.	Image concern
32	Acuna et al. 2018	bioRxiv	Bioscience-scale automated detection of figure element reuse.	Image concern
58	Hartgerink 2016	Data	688,112 Statistical Results: Content Mining Psychology Articles for Statistical Test Results.	Data concern: Statistics check (Statcheck)
59	van der Zee et al. 2017	BMC Nutrition	Statistical heartburn: an attempt to digest four pizza publications from the Cornell Food and Brand Lab.	Data concern: Statistics check (Statcheck and Test statistics)
57	Epskamp et al. 2015	R-project	Statcheck: Extract statistics from articles and recompute p values. R package version 1.0.1.	Data concern: Statistics check (Statcheck)
61	Anaya 2016	PeerJ Preprints	The GRIMMER test: A method for testing the validity of reported measures of variability	Data concern: Statistics check (Grimmer test)
60	Brown et al. 2017	Social Psychological and Personality Science	The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology	Data concern: Statistics check (Grimmer test)
62	Heathers et al. 2018	PeerJ Preprints	Recovering data from summary statistics: Sample Parameter Reconstruction via Iterative TEchniques (SPRITE)	Data concern: Statistics check (SPRITE)
63	Li et al. 2020	Fertility and sterility	Integrity of randomized controlled trials: challenges and solutions.	Data concern: Statistics check (Test statistics)
17	Dahlberg 2010	Sci Eng Ethics	Scientific Forensics: How the Office of Research Integrity can Assist Institutional Investigations of Research Misconduct During Oversight Review	Overall concern: Investigating all publication of one author
Data concern: Statistics check (Test statistics), Benford's law, Digit preference checks & Inliers				
52	Al-Marzouki et al. 2005	BMJ	Are these data real? Statistical methods for the detection of data fabrication in clinical trials.	Data concern: Statistics check (Test statistics) & Digit preference checks
18	Carlisle 2020	Anesthesia	False individual patient data and zombie randomized controlled trials submitted to Anesthesia	Overall concern: Investigating all publication of one author
Data concern: Statistics check (Test statistics), Digit preference checks, Repeated measurements & Outliers				

(continued on next page)

Table 1 (*continued*)

No. in reference list	Study	Journal/source	Title	Method
56	Hüllemann et al. 2017	Anaesthesist	Application of Benford's law: a valuable tool for detecting scientific papers with fabricated data?	Data concern: Benford's law
53	Orita et al. 2012	Expert opinion on drug discovery	Agreement of drug discovery data with Benford's law.	Data concern: Benford's law
54	Hein et al. 2012	Anaesthesist	Scientific fraud in 20 falsified anesthesia papers Detection using financial auditing methods	Data concern: Benford's law
55	Pollach et al. 2016	Medical Hypotheses	The ''first digit law'' – A hypothesis on its possible impact on medicine and development aid	Data concern: Benford's law
10	Carlisle 2012	Anesthesia	The analysis of 168 randomized controlled trials to test data integrity	Overall concern: Investigating all publications of one author Data concern: Baseline P value distribution for RCTs
44	Carlisle et al. 2015	Anesthesia	Calculating the probability of random sampling for continuous variables in submitted or published randomized controlled trials.	Data concern: Baseline P value distribution for RCTs
11	Bolland 2016	Neurology	Systematic review and statistical analysis of the integrity of 33 randomized controlled trials.	Overall concern: Investigating all publications of one author Data concern: Baseline P value distribution for RCTs
45	Carlisle et al. 2017	Anesthesia	Evidence for non-random sampling in randomized, controlled trials by Yuhji Saitoh.	Data concern: Baseline P value distribution for RCTs
47	Mascha et al. 2017	Anesthesia and analgesia	An Appraisal of the Carlisle-Stouffer-Fisher Method for Assessing Study Data Integrity and Fraud.	Data concern: Baseline P value distribution for RCTs
48	Kharasch et al. 2017	Anesthesia	Seeking and reporting apparent research misconduct: errors and integrity.	Data concern: Baseline P value distribution for RCTs
49	Bolland et al. 2019a	Journal of clinical epidemiology	Rounding, but not randomization method, non-normality, or correlation, affected baseline *P*-value distributions in randomized trials.	Data concern: Baseline P value distribution for RCTs
50	Bolland et al. 2019b	Journal of clinical epidemiology	Baseline P value distributions in randomized trials were uniform for continuous but not categorical variables.	Data concern: Baseline P value distribution for RCTs
46	Myles 2019	Anesthesia	Evidence for compromised data integrity in studies of liberal peri-operative inspired oxygen.	Data concern: Baseline P value distribution for RCTs
51	Bolland et al. 2020	Anesthesia	Empirically generated reference proportions for baseline p values from rounded summary statistics.	Data concern: Baseline P value distribution for RCTs
34	Buyse et al. 1999	Statistics in medicine	The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials.	Data concern: Benford's law, Digit preference checks, Plausibility of Correlations between variables, Date checking, Recruitment over time, Repeated measurements, Inliers, Outliers & Centre with possible data fabrication

(*continued on next page*)

Table 1 (*continued*)

No. in reference list	Study	Journal/source	Title	Method
35	Kirkwood et al. 2013	Clinical Trials	Application of methods for central statistical monitoring in clinical trials.	Data concern: Benford's law, Digit preference checks, Plausibility of Correlations between variables, Date checking, Repeated measurements, Inliers, Outliers & Centre with possible data fabrication
37	van den Bor et al. 2017	Journal of clinical epidemiology	A computationally simple central monitoring procedure, effectively applied to empirical trial data with known fraud.	Data concern: Benford's law, Digit preference checks, Plausibility of Correlations between variables, Date checking, Recruitment over time, Missing data, Outliers & Centre with possible data fabrication:
43	Hartgerink et al. 2016	PsyArXiv	Detection of Data Fabrication Using Statistical Tools	Data concern: Benford's law, Digit preference checks, Plausibility of Correlations between variables, Standard deviations & Centre with possible data fabrication
36	Taylor et al. 2002	Drug Information Journal	Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data.	Data concern: Digit preference checks, Date checking, Inliers & Centre with possible data fabrication
38	O'Kelly 2004	Pharmaceutical Statistics	Using statistical techniques to detect fraud: A test case.	Data concern: Digit preference checks, Inliers, Outliers & Centre with possible data fabrication
41	Pogue et al. 2013	Clinical trials	Central statistical monitoring: detecting fraud in clinical trials.	Data concern: Digit preference checks, Repeated measurements & Centre with possible data fabrication
42	Knepper et al. 2016	Therapeutic Innovation and Regulatory Science	Statistical Monitoring in Clinical Trials: Best Practices for Detecting Data Anomalies Suggestive of Fabrication or Misconduct.	Data concern: Digit preference checks, Plausibility of correlations between variables, Date checking, Missing data & Centre with possible data fabrication
7	Bailey 1991	Controlled clinical trials	Detecting fabrication of data in a multicenter collaborative animal study.	Overall concern: Investigating all publications of one author Data concern: Plausibility of Correlations between, Inliers, Outliers & Centre with possible data fabrication
40	Wu et al. 2012	Pharmaceutical statistics	Detecting data fabrication in clinical trials from cluster analysis perspective.	Data concern: Plausibility of Correlations between variables
13	Hudes et al. 2017	FASEB journal	Unusual clustering of coefficients of variation in published articles from a medical biochemistry department in India.	Overall concern: Investigating all publications of one author Data concern: Plausibility of Coefficients of variation
12	Simonsohn 2013	Psychological science	Just Post It: The Lesson from Two Cases of Fabricated Data Detected by Statistics Alone.	Overall concern: Investigating all publications of one author Data concern: Plausibility of Standard deviations
39	Venet et al. 2012	Clinical Trials	A statistical approach to central monitoring of data quality in clinical trials	Data concern: Repeated measurements & Centre with possible data fabrication

for inconsistencies. This tool is not able to search tables and can miss tests that are not in APA format. It checks only if the *P*-value is consistent with the test statistic and degrees of freedom. It cannot check if the test statistic or degrees of freedom are correct [59].

The GRIMMER Test [60,61] (Granularity-Related Inconsistency of Means Mapped to Error Repeats) is built

Table 2. Applicability of the methods to assess research misconduct in health-related research

Method/technique	Application				
	Minimum information required[a]	Type	Scope	Automated	Validated
Overall concern					
Screening: REAPRAISSED checklist	Manuscript + tables + figures	Mixed	Fabrication / falsification / plagiarism	-	-
Detection of patterns of misconduct in all publications of one author/group	Manuscript + tables + figures	Qualitative	Fabrication / falsification / self-plagiarism	-	-
Textual concern					
Textual plagiarism: Helioblast/ iThenticate	Manuscript	Quantitative	Plagiarism	√	√
Compare baseline and outcome tables	Tables	Mixed	Fabrication / falsification	-	-
Translated plagiarism	Manuscript	Qualitative	Plagiarism	-	-
Automatically generated fake papers: SciDetect	Manuscript	Quantitative	Fabrication	√	-
Image concern					
Image manipulation detection tools	Figures	Quantitative	Fabrication / falsification	√	√
Data concern					
Statcheck	Manuscript	Quantitative	Fabrication / falsification	√	√
Grimmer test	Manuscript + tables	Quantitative	Fabrication / falsification	√	-
SPRITE	Manuscript + tables	Quantitative	Fabrication / falsification	√	-
Recalculate test statistics	Manuscript + tables	Quantitative	Fabrication / falsification	-[b]	-
Benford's law and digit preference checks	Manuscript + tables + figures	Quantitative	Fabrication / falsification	-[b]	-[c]
Baseline P value distribution for RCTs	Manuscript + baseline table	Quantitative	Fabrication / falsification	-[b]	-[c]
Plausibility of Correlations between variables	Manuscript + tables	Quantitative	Fabrication / falsification	-[b]	-
Plausibility of Coefficients of variation	Manuscript + tables	Quantitative	Fabrication / falsification	-	-
Plausibility of Standard deviations	Manuscript + tables	Quantitative	Fabrication / falsification	-[b]	-
Date checking	Manuscript + tables + raw data	Quantitative	Fabrication / falsification	-[b]	-
Recruitment over time	Manuscript + tables	Quantitative	Fabrication / falsification	-[b]	-
Repeated measurements	Manuscript + tables + raw data	Quantitative	Fabrication / falsification	-[b]	-
Missing data	Manuscript + tables + raw data	Quantitative	Fabrication / falsification	-[b]	-
Inliers and Outliers	Manuscript + tables + raw data	Quantitative	Fabrication / falsification	-[b]	-
Centre with data manipulation (multicenter study)	Manuscript + tables + raw data	Quantitative			
	Fabrication / falsification	-	-		

[a] May be complemented by supplementary materials

[b] R-program available although automated software is absent at the moment

[c] Preliminary attempts for validation exist

Table 3. Methods and the links to find the software or R-programs

Method/technique	**Link**
Overall integrity	
Screening: REAPRAISSED checklist (M)	http://resource-cms.springernature.com/springer-cms/rest/v1/content/17589730/data/v1 (This checklist is licensed under a Creative Commons licence: CC BY-NC-SA).
Investigate all publications of one author / author group (M)	Manual
Textual integrity	
Textual plagiarism: Helioblast/ iThenticate (S)	Helioblast: https://helioblast.heliotext.com Turnitin: https://www.crossref.org/services/similarity-check/, http://www.ithenticate.com/ & https://www.turnitin.com/
Compare baseline characteristics and outcome tables (M)	Manual
Translated plagiarism (S/M)	Manual or after translation use of textual plagiarism software
Scidetect (S)	https://gricad-gitlab.univ-grenoble-alpes.fr/labbecy/scidetect
Image Integrity	
Image manipulation detection tools (S)	https://ori.hhs.gov/forensic-tools, https://github.com/lkoppers/FraudDetTool
Data Integrity	
Statcheck (S)	https://cran.r-project.org/web/packages/statcheck/index.html.
Grimmer test (S)	http://www.prepubmed.org/grimmer/
Sprite algorithm (S)	http://www.prepubmed.org/sprite/
Recalculate test statistics (R) (M)	https://cran.r-project.org/web/packages/rpsychi/index.html, https://github.com/OmnesRes/pizzapizza
Benford's law (leading digit) and Digit preference (last digit) (R) (M)	http://www.ctc.ucl.ac.uk/Training.aspx, R-program available of the Web Appendix of van den Bor, et al. 20, https://github.com/chartgerink/ddfab
Baseline P value distribution for RCTs (R) (M)	Manual, https://cran.r-project.org/web/packages/simdistr/index.html
Plausibility of Correlations between variables (R) (M)	http://www.ctc.ucl.ac.uk/Training.aspx, R-program available of the Web Appendix of van den Bor, et al. 20
Plausibility of Coefficients of variation (M)	Manual
Plausibility of Standard deviations (M)	Manual, https://github.com/chartgerink/ddfab
Date checking (R) (M)	http://www.ctc.ucl.ac.uk/Training.aspx, R-program available of the Web Appendix of van den Bor, et al. 20
Recruitment over time (R) (M)	R-program available of the Web Appendix of van den Bor, et al. 20
Repeated measurements (R) (M)	http://www.ctc.ucl.ac.uk/Training.aspx
Missing data (R) (M)	R-program available of the Web Appendix of van den Bor, et al. 20
Inliers and Outliers (R) (M)	http://www.ctc.ucl.ac.uk/Training.aspx, R-program available of the Web Appendix of van den Bor, et al. 20
Centre with possible data fabrication	
Mean at each center to overall mean of other centers (R) (M)	http://www.ctc.ucl.ac.uk/Training.aspx
Substantial differences in outcomes/treatment effects (M)	Manual
(S) Software package, (R) R-program available, (M) manual	

upon the GRIM test [60]. This freely available software allows for testing whether reported measures of variability are mathematically possible. GRIMMER relies upon the statistical phenomenon that variances display a simple repetitive pattern when the data is discrete, i.e., granular. The algorithm created by Anaya [61] can identify whether a reported statistic is consistent with the sample size and granularity. The ability of the test relies upon: (1) the sample size; (2) the granularity of the data; (3) the precision (number of decimals) of the reported statistic; and (4) the size of the standard deviation or standard error (but not the variance). A limitation of the test is that it is at present restricted to a sample size of 99.

SPRITE (Sample Parameter Reconstruction via Interative Techniques) is a technique for reconstructing potential discrete data sets using only basic summary information about a sample, namely the mean, the standard deviation, the sample size, and the lower and upper bounds of the range of item values. SPRITE complements the GRIM and GRIMMER tests [62]. SPRITE does not have

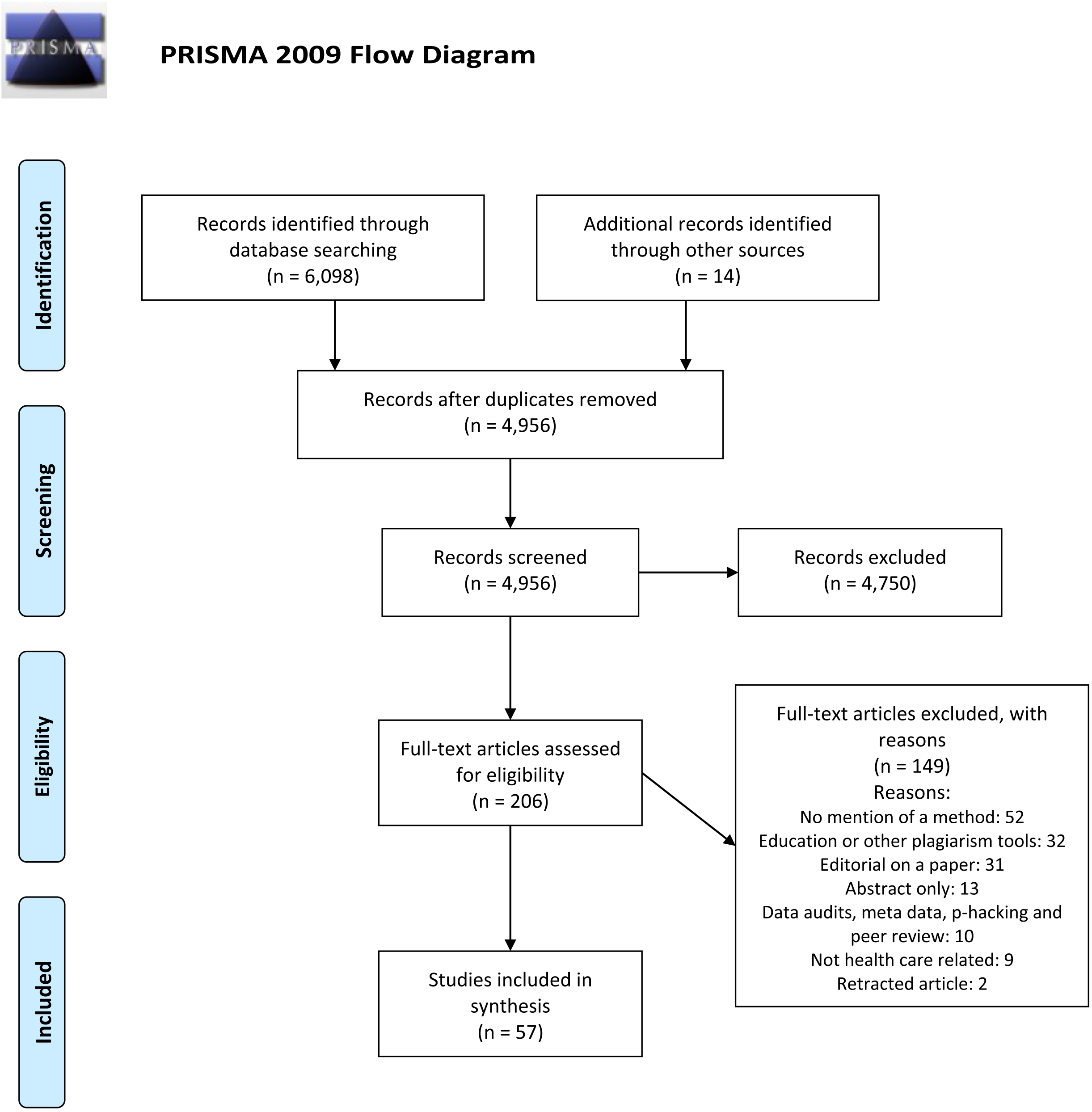
PRISMA
PRISMA 2009 Flow Diagram
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 6,098)
Additional records identified through other sources (n = 14)
Records after duplicates removed (n = 4,956)
Records screened (n = 4,956)
Records excluded (n = 4,750)
Full-text articles assessed for eligibility (n = 206)
Full-text articles excluded, with reasons (n = 149)
Reasons:
No mention of a method: 52
Education or other plagiarism tools: 32
Editorial on a paper: 31
Abstract only: 13
Data audits, meta data, p-hacking and peer review: 10
Not health care related: 9
Retracted article: 2
Studies included in synthesis (n = 57)

Figure 1. PRISMA 2009 flow diagram.

a sample size limitation. SPRITE also takes into consideration the range of possible values of the raw data. It can identify cases in which the summary statistics are theoretically possible, but imply a highly skewed or otherwise unusual distribution of individual responses. SPRITE is not a complete mathematical solution, and some degree of interpretation of its output will always be required [62].

Statistics checks could be performed for trials that performed univariable analyses. Independent *t*-test, one- and two-way ANOVAs can be checked using the means, standard deviations, and the sample size reported in articles [59]. Chi-square tests, Fisher's exact tests, unadjusted odds ratios, and risk ratios can be reproduced using absolute numbers given in crosstabs [63].

If an original raw dataset is available the statistics can be recalculated and a comparison can be made between the results of these recalculations and the resulting claims in the paper [17]. ORI developed a method that focuses on the insignificant data or numbers of a paper whenever possible. This principle is based on their repeated observation, that when falsifying or fabricating data, an individual will focus on the desired outcome and pay less attention to the other data to make it appear authentic [17].

7.2. Benford's law and digit preference checks

Benford's law is a description of the probability of the digits in naturally occurred numbers. The first digits tend to follow a non-uniform distribution in the natural occurrence which means that the first digits one, two, and three account for more than 60% of the total probability distribution. The last digits, however, are expected to approach uniform distribution.

It is possible to collect all the numbers from published articles and assess the frequencies of the digits by applying Benford's law. There may be some legitimate reasons for unequal distributions of digits, for example, biomarker measurements might be rounded to the last digit of either zero or five because the technology is insufficiently accurate. However, preference for one digit over another might be an indication of digit preference of a fraudulent researcher. Limited validation studies on cases of proven fraud and non-verified controls found that this approach is highly sensitive, but the specificity is unclear [56]. Only limited exploratory studies have revealed that some natural biomedical data obey Benford's law [53].

7.3. Baseline P*-value distribution for randomized trials*

This method only applies to randomised trials. In a trial, the *P*-values in the baseline characteristics table (often "Table 1") are expected to be "uniform" with equal distributions between zero and one. This balance of baseline characteristics due to random allocation is difficult to be perfectly fabricated. Using summary statistics in publications, baseline *P*-values can be obtained through parametric tests or Monte Carlo simulations [11,44].

The effectiveness of this method in locating trials with concerns has been demonstrated empirically in several notorious cases of research misconduct [11,44–46]. However, there is concern about the validity of the expected uniform distribution given several assumptions may be violated such as independence between variables, exclusive use of simple randomisation, no rounding of summary statistics, and no publication bias [47,48]. Reassuringly, recent simulation studies suggest that although correlation, randomisation method, and non-normality do not have important effects on baseline *P*-value distribution, those calculated from rounded summary statistics are not uniformly distributed [49]. Also, it was found that baseline *P*-value distributions were uniform for continuous but not categorical characteristics [50]. Based on these findings, the true expected (i.e., reference) distributions for baseline *P*-values from rounded summary statistics were established empirically [51].

Positive findings using this method may be due to one or a combination of the inaccuracy of the method, honest errors regarding data analysis and reporting, chance, or fraud [47].

7.4. Plausibility of correlations between variables, coefficients of variation and standard deviations

Researchers may create false data and use sensible values for a single variable. However, it is difficult to fabricate several variables that together are consistent with real data [35]. By eyeballing the baseline and results sections, unlikely values may come to light. Some variables should be correlated based on knowledge or common sense, the correlation after manipulations of the data may end up too strong or too weak to be plausible [7,34,35,37,40,42,43].

Similar to correlations, it is difficult to fabricate multiple means and standard deviations for separate variables or groups in a way that they differ enough to be realistic but not so much that it attracts attention. Coefficients of variation indicate variable variation regardless of its unit, defined as dividing the sample standard deviation by the sample mean. Researchers who commit fraud could unconsciously pick coefficients of variation that are too similar for unrelated variables with very different scales [13].

Fabricated data may tend to have too similar standard deviations to be plausible [12,43]. When researchers fabricate different means for two or more study arms, they might be reluctant to change the standard deviation. The standard deviation of multiple standard deviations across groups can indicate that they are unrealistically similar [12].

7.5. Date checking and recruitment over time

In presence of raw data, all dates should occur after the first participant being recruited or randomised, and before final events such as death or the end of the study [35]. Also, it could be checked whether there is a relative irregularity of subject visits taking place during weekends [37] and whether routine measurements were not taken at weekends or holidays [36], as randomisation or clinic appointments are unlikely to heavily fall on these days. Care must be taken in choosing which dates to check, because dates of death, emergency treatment, or some clinic visits may occur at any time [35,36,42].

Furthermore, the rate by which real participants are recruited might not be perfectly constant over time as studies often have a "start-up" period. Inclusions for fabricated data might be more constant over time [37]. In trials, a comparison of treatment groups by week or month of randomization can reveal periods with unrealistic inclusion [34].

7.6. Repeated measurements and missing data

Some variables are measured repeatedly on the same individuals. An insufficient variability over time may reveal propagation of previous values rather than genuine observations [34]. If data are fabricated the false values may not vary enough compared to real data [35]. Repeated sequences of values of different included individuals can also be found within a whole column by plotting the column values in order [18].

Also, fabricated data might be “too perfect” in the sense of containing relatively few missing values. Missing data rates can be checked in raw data and missing data rates can be compared between centres in case of a multicentre trial [37,42].

7.7. Outliers and Inliers: unrealistically large or low variance in data variables

Outliers are observations that appear to be inconsistent with the rest of the data, usually appearing as too large. Here, this method compares the observed value for a single participant to those from all other participants. Outliers at the participant level are more likely to result from errors rather than fraud [35]. On the other hand, researchers who create false data tend to choose values close to the mean of the other observations, as outliers might be noticed by others [35]. Thus, in their tendency to avoid creating outliers, researchers who commit fraud might create odd distributions in which individual values are unusually close to the overall study mean (inliers) [38]. Having several participants with a very small difference from the study mean at the same site, the site may require further investigation [35]. ORI uses this principle to compare suspicious datasets with control datasets of a similar topic [17].

7.8. Centre with possible data fabrication

Central statistical monitoring (CSM), using statistical methods to compare data of one centre with data of all the centres combined, focuses on multicentre studies. This principle is based on the assumption that a fraudulent staff member does not have access to any trial data of other centres. Consequently, fabricated observations might be different from true observations [37].

Most above-mentioned methods that assess data concerns can be used in CSM. Some further comparisons may be helpful. For example, if a particular site has mean values that are very different from the other sites, it might indicate that some participants have been fabricated, or those recruited are so different from other centres that this may require further investigation [34,35,39,41]. Also, fraudulent researchers might wish to demonstrate positive findings. Hartgerink, Voelkel [43] compared genuine and fabricated summary statistics and found that the fabricated effects were in general larger than the genuine ones.

8. Discussion

This scoping review describes numerous methods to assess research misconduct. The methods to detect textual plagiarism have been regularly implemented as detection tools. Most other methods have not been adequately validated nor structurally implemented. Some methods are based on eyeballing and experience. There is a need for automation to facilitate the detection of potential misconduct.

The strength of this scoping review is that it brings together all literature-reported methods that detect research misconduct in health-related research. We sorted the collection of methods and summarised their applicability to build a quick reference guide for readers. However, it is possible that some unpublished methods were missed in this literature-based effort, especially in-house methods that belong to publishers. These methods are usually commercial products for certain aims and may not have good generalizability. We could not obtain enough information to assess them via the public domain. We are also aware of some methods that pass from mouth-to-mouth, such as implausible productivity of researchers, implausibly high recruitment rates given the stringent eligibility criteria and the capacity of the recruiting centres, and inability to identify the claimed Institutional Review Board. Another limitation is that it was not possible to make a comparison of the available methods because they focus on different dimensions and behaviours of research misconduct. Only a few tests were validated and there was almost no information on how reliable the results are, let alone systematic critical appraisal of the identified methods. Limitations of few methods have been preliminarily discussed, for example, limitations of the increasingly used baseline *P*-value distribution in randomized trials have been touched upon [64,65]. But for most methods identified in this review, there is no reference to their strengths and weaknesses. Even the validated tests have limitations, as there are still discussions on setting thresholds for plagiarism. These underpin the necessity to use multiple methods for any investigation.

We advise using multiple methods to detect potential research misconduct because a single method is usually insufficient. At this moment there is no one particular method that we recommend using alone. The main research gap is that we need to know what minimal set of tests are required to optimize detection of misconduct; this includes the necessity of validation of available methods and determining their diagnostic capacity. Second, it always helps to ask for the raw datasets and apply statistical checks. These attempts are usually hampered by the poor accessibility and stewardship of research data. As an obligation of publication, a unified requirement to submit research data to appropriate data repositories along with meta-data like data dictionaries may be part of the solution. Third, it is important to check research governance including protocols, ethics approval, and documentation of study medication as this will contribute to either trust or distrust of the research. Last, we advise automating these methods as much as possible. Automation of “ready” methods would promote wide use. Automation of methods in development would encourage validation and testing. We also encourage new methods to be automated in advance to expedite the process of validation and application.

A thorough investigation of suspected research misconduct is currently a difficult, time-consuming, and labour-intensive process. The scientific community needs to develop better detection tools that are validated. Subsequently, these tools can be automated for routine assessments and tested by the community to proactively defend the integrity of research before publication.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' roles

EB designed the study, managed the literature search, extracted data, and drafted the manuscript; MvW designed the study, managed the literature search, checked data, and critically revised the manuscript; WL designed the study, helped to interpret the statistical methods and critically revised the manuscript; RW and RvE helped to interpret the statistical methods and critically revised the manuscript; MS performed the literature search and critically revised the manuscript; BwM designed the study, and critically revised the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2021.05.012.

References

[1] Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One 2009;4 e5738.

[2] Dancet EAF, D'Hooghe TM, Dreischor F, van Wely M, Laan ETM, Lambalk CB, et al. The 'Pleasure&Pregnancy' web-based interactive educational programme versus expectant management in the treatment of unexplained subfertility: protocol for a randomised controlled trial. BMJ open 2019;9:e025845.

[3] Practice Committee of the American Society for Reproductive MedicineElectronic address aao, practice committee of the american society for reproductive m. evidence-based treatments for couples with unexplained infertility: a guideline. Fertil Steril 2020;113:305–22.

[4] Wager E. Coping with scientific misconduct. Bmj 2011;343 d6586.

[5] Bosch X, Hernández C, Pericas JM, Doti P, Marušić A. Misconduct policies in high-impact biomedical journals. PloS one 2012;7 e51928-e.

[6] Tricco A, Lillie E, Zarin W, O'Brien K, Colquhoun H, Levac D. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med 2018;169(7):467–73.

[7] Bailey KR. Detecting fabrication of data in a multicenter collaborative animal study. Control Clin Trials 1991;12:741–52.

[8] Grey A, Bolland MJ, Avenell A, Klein AA, Gunsalus C. Check for publication integrity before misconduct. Nature Publishing Group; 2020.

[9] Smith R. Investigating the previous studies of a fraudulent author. Br Med J. 2005;331:288–91.

[10] Carlisle JB. The analysis of 168 randomised controlled trials to test data integrity. Anaesthesia 2012;67:521–37.

[11] Bolland MJ, Avenell A, Gamble GD, Grey A. Systematic review and statistical analysis of the integrity of 33 randomized controlled trials. Neurology 2016;87:2391–402.

[12] Simonsohn U. Just post it: the lesson from two cases of fabricated data detected by statistics alone. Psychol Sci 2013;24.

[13] Hudes ML, McCann J, Ames B. Unusual clustering of coefficients of variation in published articles from a medical biochemistry department in India. FASEB J 2009;23(3):689–703.

[14] Spiroski M. How to verify plagiarism of the paper written in Macedonian and translated in foreign language? Open Access Maced J Med Sci 2016;4:1–4.

[15] Bordewijk EM, Wang R, van Wely M, Li W, Mol BW. Data integrity of 10 other randomized controlled trials of an author with a retracted paper. Fertil Steril 2020. https://www.fertstertdialog.com/posts/data-integrity-of-10-other-randomized-controlled-trials-of-an-author-with-a-retracted-paper. [Accessed 05 January 2021].

[16] Bordewijk EM, Wang R, Askie LM, Gurrin LC, Thornton JG, van Wely M. Data integrity of 35 randomised controlled trials in women' health. Eur J Obstet Gynecol Reprod Biol 2020;249:72–83.

[17] Dahlberg JE, Davidian NM. Scientific forensics: how the office of research integrity can assist institutional investigations of research misconduct during oversight review. Sci Eng Ethics 2010;16:713–35.

[18] Carlisle JB. False individual patient data and zombie randomised controlled trials submitted to Anaesthesia. Anaesthesia 2020;76(4):472–9.

[19] Bohannon J. Scientific publishing. Hoax-detecting software spots fake papers. Science 2015;348:18–19.

[20] Nguyen M, Labbé C. Engineering a tool to detect automatically generated papers. BIR@ECIR; 2016.

[21] Springer and Université Joseph, Fourier release SciDetect to discover fake scientific papers. 2020, https://www.springer.com/gp/about-springer/media/press-releases/corporate/scidetect/541662015. [Accessed 05 January 2021].

[22] Baydik OD, Gasparyan AY. How to act when research misconduct is not detected by software but revealed by the author of the plagiarized article. J Korean Med Sci 2016;31:1508–10.

[23] Wiwanitkit V. How to verify and manage the translational plagiarism? Maced J Med Sci 2016;4:533.

[24] Errami M, Wren JD, Hicks JM, Garner HR. eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. Nucleic Acids Res 2007;35:W12–WW5.

[25] Errami M, Sun Z, George AC, Long TC, Skinner MA, Wren JD. Identifying duplicate content using statistically improbable phrases. Bioinformatics 2010;26:1453–7.

[26] Errami M, Hicks JM, Fisher W, Trusty D, Wren JD, Long TC. Déjà vu—A study of duplicate citations in Medline. Bioinformatics 2007;24:243–9.

[27] How to stop plagiarism. Nature 2012;481:21–3.

[28] Higgins JR, Lin F-C, Evans JP. Plagiarism in submitted manuscripts: incidence, characteristics and optimization of screening-case study in a major specialty medical journal. Res Integr Peer Rev 2016;1:13.

[29] Taylor DB. Plagiarism in manuscripts submitted to the AJR: Development of an optimal screening algorithm and management pathways. AJR Am J Roentgenol 2017;208:712–20.

[30] The office of research integrity (ORI). Forensic Tools, 2020. https://ori.hhs.gov/forensic-tools. [Accessed 05 January 2021].

[31] Koppers L, Wormer H, Ickstadt K. Towards a systematic screening tool for quality assurance and semiautomatic fraud detection for images in the life sciences. Sci Eng Ethics 2017;23:1113–28.

[32] Acuna DE, Brookes PS, Kording KP. Bioscience-scale automated detection of figure element reuse. BioRxiv 2018:269415.

[33] Parrish D, Noonan B. Image manipulation as research misconduct. Sci Eng Ethics 2009;15:161–7.

[34] Buyse M, George SL, Evans S, Geller NL, Ranstam J, Scherrer B. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. Stat Med 1999;18:3435–51.

[35] Kirkwood AA, Cox T, Hackshaw A. Application of methods for central statistical monitoring in clinical trials. Clin Trials 2013;10:783–806.

[36] Taylor RN, McEntegart DJ, Stillman EC. Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. Drug Inf J 2002;36:115–25.

[37] van den Bor RM, Vaessen PWJ, Oosterman BJ, Zuithoff NPA, Grobbee DE, Roes KCB. A computationally simple central monitoring procedure, effectively applied to empirical trial data with known fraud. J Clin Epidemiol 2017;87:59–69.

[38] O'Kelly M. Using statistical techniques to detect fraud: A test case. Pharm Stat 2004;3:237–46.

[39] Venet D, Doffagne E, Burzykowski T, Beckers F, Tellier Y, Genevois-Marlin E. A statistical approach to central monitoring of data quality in clinical trials. Clin Trials 2012;9:705–13.

[40] Wu X, Carlsson M. Detecting data fabrication in clinical trials from cluster analysis perspective. Pharm Stat 2011;10:257–64.

[41] Pogue JM, Devereaux PJ, Thorlund K, Yusuf S. Central statistical monitoring: detecting fraud in clinical trials. Clin Trials 2013;10:225–35.

[42] Knepper D, Lindblad AS, Sharma G, Gensler GR, Manukyan Z, Matthews AG. Statistical monitoring in clinical trials: best practices for detecting data anomalies suggestive of fabrication or misconduct. Ther Innov Regul Sci 2016;50:144–54.

[43] Hartgerink CHJ, Voelkel JG, Wicherts JM, van Assen MALM. Detection of data fabrication using statistical tools. PsyArXiv 2019. doi:10.31234/osf.io/jkws4.

[44] Carlisle JB, Dexter F, Pandit JJ, Shafer SL, Yentis SM. Calculating the probability of random sampling for continuous variables in submitted or published randomised controlled trials. Anaesthesia 2015;70:848–58.

[45] Carlisle JB, Loadsman JA. Evidence for non-random sampling in randomised, controlled trials by Yuhji Saitoh. Anaesthesia 2017;72:17–27.

[46] Myles PS, Carlisle JB, Scarr B. Evidence for compromised data integrity in studies of liberal peri-operative inspired oxygen. Anaesthesia 2019;74:573–84.

[47] Mascha EJ, Vetter TR, Pittet J-F. An Appraisal of the Carlisle-Stouffer-Fisher Method for Assessing Study Data Integrity and Fraud. Anesth Analg 2017;125:1381–5.

[48] Kharasch ED, Houle TT. Seeking and reporting apparent research misconduct: errors and integrity. Anaesthesia 2018;73:125–6.

[49] Bolland MJ, Gamble GD, Avenell A, Grey A. Rounding, but not randomization method, non-normality, or correlation, affected baseline P-value distributions in randomized trials. J Clin Epidemiol 2019;110:50–62.

[50] Bolland MJ, Gamble GD, Avenell A, Grey A, Lumley T. Baseline P value distributions in randomized trials were uniform for continuous but not categorical variables. J Clin Epidemiol 2019;112:67–76.

[51] Bolland MJ, Gamble GD, Grey A, Avenell A. Empirically generated reference proportions for baseline p values from rounded summary statistics. Anaesthesia 2020.

[52] Al-Marzouki S, Evans S, Marshall T, Roberts I. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. BMJ. 2005;331:267-70.

[53] Orita M, Hagiwara Y, Moritomo A, Tsunoyama K, Watanabe T, Ohno K. Agreement of drug discovery data with Benford's law. Expert Opin Drug Discov 2013;8:1–5.

[54] Hein J, Zobrist R, Konrad C, Schuepfer G. Scientific fraud in 20 falsified anesthesia papers : detection using financial auditing methods. Der Anaesthesist 2012;61:543–9.

[55] Pollach G, Brunkhorst F, Mipando M, Namboya F, Mndolo S, Luiz T. The "first digit law" - A hypothesis on its possible impact on medicine and development aid. Med Hypotheses 2016;97:102–6.

[56] Hullemann S, Schupfer G, Mauch J. Application of Benford's law: a valuable tool for detecting scientific papers with fabricated data?: A case study using proven falsified articles against a comparison group. Anaesthesist 2017;66:795–802.

[57] Epskamp S, Nuijten MB. statcheck: Extract statistics from articles and recompute p values. R package version 1.0.1. http://CRAN.R-project.org/package=statcheck. 2015.

[58] Hartgerink C. 688,112 Statistical results: content mining psychology articles for statistical test results. Data 2016;1:14.

[59] van der Zee T, Anaya J, Brown NJL. Statistical heartburn: an attempt to digest four pizza publications from the Cornell Food and Brand Lab. BMC Nutr 2017;3:54.

[60] Brown NJL, Heathers JAJ. The GRIM test:a simple technique detects numerous anomalies in the reporting of results in psychology. Soc Psychol Personal Sci 2017;8:363–9.

[61] Anaya J. The GRIMMER test: A method for testing the validity of reported measures of variability. PeerJ Preprints 2016;4 e2400v1.

[62] Heathers J, Anaya J, van der Zee T, Brown N. Recovering data from summary statistics: Sample parameter reconstruction via iterative techniques. (SPRITE) Peer J Preprints 2018;6:e26968v1.

[63] Li W, van Wely M, Gurrin L, Mol BW. Integrity of randomized controlled trials: challenges and solutions. Fertil Steril 2020;113:1113–19.

[64] Betensky RA, Chiou SH. Correlation among baseline variables yields non-uniformity of p-values. PLoS One 2017;12:e0184531.

[65] Bland M. Do baseline P-values follow a uniform distribution in randomised trials? PLoS One 2013;8:e76010.