Journal of Clinical Epidemiology 170 (2024) 111365

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

Statistical techniques to assess publication integrity in groups of randomized trials: a narrative review

Mark J. Bolland[a,b,*], Alison Avenell[c], Andrew Grey[a]

[a]*Department of Medicine, University of Auckland, Private Bag 92 019, Auckland 1142, New Zealand*
[b]*Department of Endocrinology, ADHB, Private Bag 92 024, Auckland 1142, New Zealand*
[c]*Health Services Research Unit, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, Scotland*

Accepted 9 April 2024; Published online 15 April 2024

Abstract

Objectives: To describe statistical tools available for assessing publication integrity of groups of randomized controlled trials (RCTs).
Study Design and Setting: Narrative review.
Results: Freely available statistical tools have been developed that compare the observed distributions of baseline variables with the expected distributions that would occur if successful randomization occurred. For continuous variables, the tools assess baseline means, baseline *P* values, and the occurrence of identical means and/or standard deviation. For categorical variables, they assess baseline *P* values, frequency counts for individual or all variables, numbers of trial participants randomized or withdrawing, and compare reported with independently calculated *P* values. The tools have been used to identify publication integrity concerns in RCTs from individual groups, and performed at an acceptable level in discriminating intentionally fabricated baseline summary data from genuine RCTs. The tools can be used when concerns have been raised about RCT(s) from an individual/group and when the whole body of their work is being examined, when conducting systematic reviews, and could be adapted to aid screening of RCTs at journal submission.
Conclusion: Statistical tools are useful for the assessment of publication integrity of groups of RCTs.

Keywords: Statistical methods; Research integrity; *P* values; Randomization; Fabricated data; Data integrity; R

1. Introduction

A reliable biomedical literature is essential , but errors and/or unreliable data in publications are common [1]. These publications can be said to have compromised publication integrity. Often, the issues arise from honest mistakes, such as typographical or coding/analytical errors, but sometimes they result from questionable research practices, including research misconduct such as fabrication or falsification. Detecting compromised publication integrity, particularly arising from questionable research practices, can be difficult. However, techniques have been described for assessing individual publications [2,3]. Assessments can be conducted using summary information from the publication [4], although assessing individual patient data is more powerful [5].

Publication integrity concerns about a body of work from a research group can arise either de novo or in the context of existing retractions. Here, techniques for single publications can be used, but it is also possible to simultaneously examine data from all studies. For groups of randomized controlled trials (RCTs), which are among the highest levels of evidence and strongly influence clinical practice, potentially powerful techniques can aid assessment of publication integrity. A fundamental premise of an RCT is that as group allocation occurs randomly (by chance), any between-group differences are due to chance. This principle can be exploited to assess publication integrity. Baseline variables (and outcomes unrelated to randomization) will differ by chance in an RCT. Therefore, the observed distributions of variables can be compared with the distributions expected to arise by chance. If the

Funding: This research received no specific funding. M.B. is a recipient of an HRC Clinical Practitioners Fellowship. The authors are independent of the HRC. The HRC had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

* Corresponding author. Bone and Joint Research Group, Department of Medicine, Faculty of Medical and Health Sciences, University of Auckland, Private Bag, 92019, Auckland, New Zealand.
E-mail address: m.bolland@auckland.ac.nz (M.J. Bolland).

https://doi.org/10.1016/j.jclinepi.2024.111365

What is new?

Key findings

- Statistical tests have been developed to aid in the assessment of publication integrity of groups of randomized controlled trials and are freely available.

What this adds to what was known?

- The tools compare observed and expected distribution of continuous and categorical baseline variables, and numbers of trial participants randomized or withdrawing, matching summary statistics, and compare reported and independently calculated *P* values.
- The tools have been used to identify integrity concerns in groups of published randomized trials and performed acceptably in discriminating between intentionally fabricated summary trial data and data from genuine randomized trials.

What is the implication and what should change now?

- The tools could be used to examine bodies of work about which concerns have been raised, and adapted to aid in screening of journal submissions.

observed and expected distributions are discrepant, this suggests that randomization has been unsuccessful or compromised. Because successful randomization is the fundamental aspect of an RCT, compromised randomization means its results are unreliable and causal inferences cannot be drawn. Such an RCT has compromised publication integrity.

At least a moderate number of variables are usually required for sufficient power to detect differences between observed and expected distributions. Many individual publications report few baseline data (such as patient characteristics), so analysing groups of RCTs can overcome this limitation, and the resulting pattern of observed results may provide compelling evidence of compromised publication integrity. It may be difficult for researchers with nefarious intent to fabricate or alter data so that observed distributions conform to expected distributions [6].

Several statistical techniques for assessing groups of RCTs have been developed recently [7−14]. To our knowledge, they have not been reviewed previously. Here, in a narrative review, we summarize these techniques, give examples of their application to existing groups of RCTs with compromised publication integrity, and report the application of these techniques to data intentionally fabricated by clinicians and statisticians.

2. Literature review and examples

We searched PubMed (until October 31, 2023) for relevant articles to inform this review using the terms ‘‘publication integrity,’’ ‘‘‘research integrity’ and statistics,’’ and the MESH term ‘‘Scientific Misconduct/statistics and numerical data.’’ We hand-searched the references of identified articles for other potentially relevant publications. All relevant identified publications have been included in this review. We have illustrated the techniques using examples from existing datasets of groups of RCTs with and without integrity concerns. Table 1 shows features of three groups of RCTs with known integrity concerns (SatoIwamoto [15], Asemi [12,16], Monticone [14]), two control sets of RCTs without integrity concerns (Auckland control dataset [10], REBALANCE [12]) and the Carlisle dataset [4], which includes 5087 RCTs published in eight journals over 15 years, and contains 72 previously retracted RCTs and some trials with publication integrity concerns identified through Carlisle’s analysis. All analyses were done using the freely available package ‘‘reappraised’’ for the R statistical program (https://CRAN.R-project.org/package=reappraised) which contains functions that carry out all the techniques described here for groups of RCTs, and function names are provided.

3. Techniques

3.1. Baseline continuous variables

The most commonly used technique has been evaluation of baseline continuous variables. In the first study of its kind, Carlisle compared the distribution of individual continuous variable means in 168 RCTs by Fujii et al. with the expected distribution [7]. He also pooled all the continuous variables and repeated the analysis. For comparison, he repeated the analyses in 366 RCTs by other authors. In the 168 RCTs of interest, the observed and expected distributions differed markedly, whereas in the 366 control RCTs, both were similar. Subsequently, it was also found that the Fujii RCTs lacked ethical approval, and by November 2023, 172 publications by Fujii had been retracted.

Carlisle et al. modified the original technique to handle calculation of *P* values from rounded summary statistics (mean, standard deviation [SD]) better [8]. This Monte-Carlo analysis could be used across a whole body of RCTs, or on a single RCT [4]. When applied to groups of RCTs, a uniform distribution is expected for *P* values from the comparison of means between RCT arms calculated using this approach. Figure 1A shows that the cumulative distribution function (CDF) in the SatoIwamoto dataset differs markedly from the expected CDF (reappraised package, anova_fn).

We adapted the original approach of Carlisle to *P* values obtained from the comparison of baseline continuous

Table 1. Integrity concerns in the datasets used

	Groups of RCTs with known integrity concerns			Control datasets		
Dataset name	**SatoIwamoto [15]**	**Asemi [12,16]**	**Monticone [14]**	**Auckland [10]**	**REBALANCE [12]**	**Carlisle [4]**
Trials (*n*)	41	151	17	13	131	5087
Integrity concerns						
Differences in baseline means	Yes	Yes	NA	No	NA	No
Continuous baseline *P*-value distribution	Yes	Yes	NA	No	NA	No
Matching summary statistic	Yes	Yes	NA	No	NA	NA
Differences between reported and calculated *P* values (categorical variables)	Yes	NA	Yes	NA	NA	NA
Categorical baseline *P*-value distribution	Yes	NA	Yes	No	NA	NA
Distribution of participant numbers	Yes	Yes	NA	No	NA	NA
Distribution of withdrawals	Yes	Yes	NA	NA	No	NA
Distribution of frequency counts	Yes	NA	Yes	No	No	NA

More details of the datasets are available in the primary publications.
NA, Not assessed.

variables (continuous baseline *P* values) [10,11]. Because differences between randomized groups arise by chance, the expected distribution of continuous baseline *P* values is random, that is uniform, when calculated from individual patient data. Figure 1B shows that the observed and expected distribution of reported continuous baseline *P* values from the SatoIwamoto dataset are markedly different, and Figure 1D shows the area under the curve of the CDF of

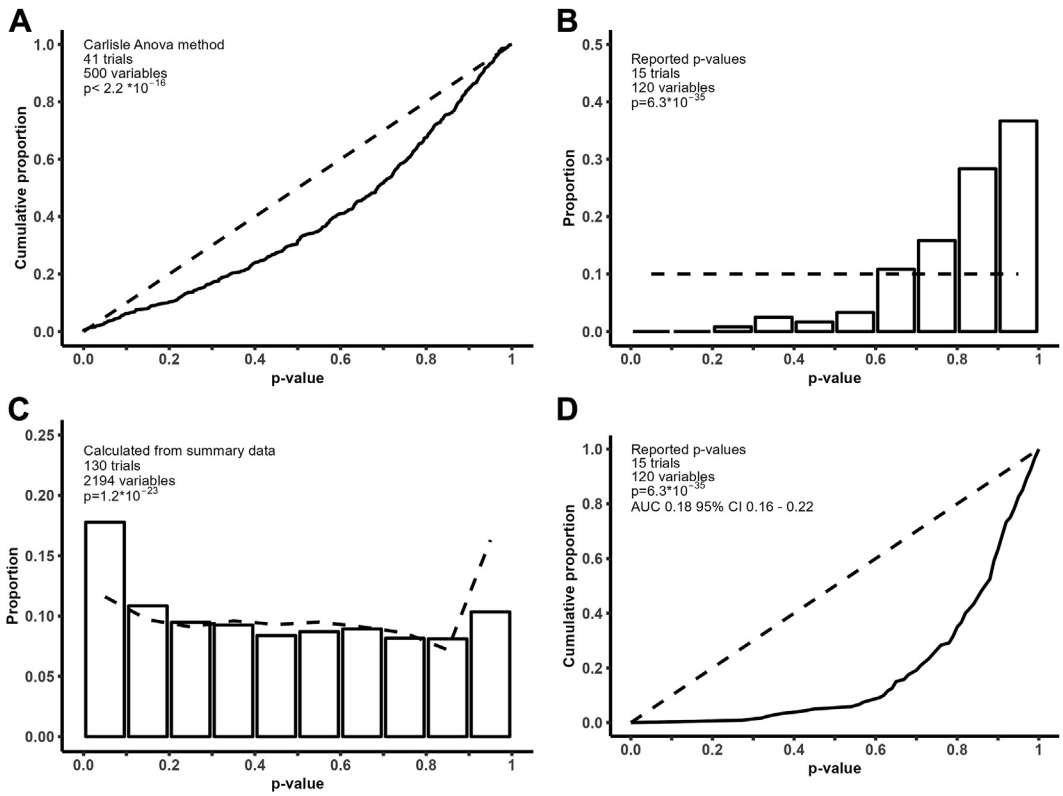

Figure 1. (A) Monte-Carlo analysis of differences in continuous baseline variables in the SatoIwamoto dataset. The expected distribution is the dotted line. (B) Observed and expected (dotted line) distribution of reported continuous baseline *P* values in the SatoIwamoto dataset. (C) Observed and reference (dotted line) distribution of calculated continuous baseline *P* values in the Asemi dataset. (D) Area under the curve (AUC) of the cumulative distribution function of the continuous baseline *P* values in (B). The expected distribution is the dotted line.

the *P* values, which quantifies the extent of departure from the expected distribution (reappraised package, pval_cont_fn). By contrast, the observed distribution of continuous baseline *P* values is uniform in the Auckland control dataset [10]. In a subsequent series of analyses and simulations, we found that non-normality of data, correlation between baseline variables at the level found in real clinical trials, and method of randomization did not affect *P*-value distribution [11]. However, rounding of data had a visually obvious effect on the expected distribution [11], and therefore needs to be incorporated into reference expected distributions. Reference data for the expected distribution of *P* values from rounded data are available from >5000 RCTs [16] or can be calculated empirically. Figure 1C shows that the observed distribution of continuous baseline *P* values in the Asemi dataset when *P* values are calculated from reported (rounded) summary statistics is very different to the reference expected distribution (reappraised, pval_cont_fn).

3.2. Matching summary statistics for continuous variables

Some studies report identical summary statistics (mean, SD) for a variable. These can be either in different arms within an RCT or in different cohort studies/RCTs. The occurrence of an identical mean and an identical SD for a variable in an RCT is uncommon, unless it is rounded to 1 significant figure, and/or the SD is small [13]. For a single variable, the likelihood of the observed identical summary statistics can be simulated [13]. This approach can be extended to determine the likelihood of identical summary statistics in different RCTs (or cohort studies) conducted in similar populations [13]. Table 2 shows the proportion of identical means/SDs in different studies from the SatoIwamoto dataset compared with the Auckland control trials (reappraised, cohort_fn). Six of 10 variables occurring in ≥10 studies had more identical mean/SDs than expected in the SatoIwamoto dataset compared with 0/10 variables in the Auckland trials.

A similar approach can be applied to the proportion of matching summary statistics within individual RCTs. These calculated proportions can be compared to those from the reference Carlisle dataset [13] or calculated empirically. Table 3 shows proportions for the SatoIwamoto dataset, a fabricated dataset from a validation study described later, and the reference dataset (reappraised, match_fn). In the SatoIwamoto dataset, there are a large proportion of means reported to three significant figures, and a higher proportion of matching means than for the reference dataset. By contrast, in the fabricated dataset, there are fewer than expected matches. The differences between proportions of matches in RCTs with integrity concerns we have analysed and the reference dataset were inconsistent, suggesting this approach needs further study.

3.3. Baseline categorical variables

Unlike continuous variables, the expected distribution of baseline *P* values for categorical variables is not uniform, but it can be calculated empirically, permitting an observed to expected comparison [10]. Figure 2A shows the observed and empirically calculated expected distribution of

Table 2. The proportion of recurring summary statistics for variables in at least 10 studies in the SatoIwamoto and Auckland control datasets

	SatoIwamoto dataset				Auckland control dataset				
Variable	*N*[a]	Mean/SD Match *N* (%)[a]	Largest Number Matches *N* (%)[a]	P Mean/SD Match[b]	Variable	*N*[a]	Mean/SD Match *N* (%)[a]	Largest Number Matches *N* (%)[a]	P Mean/SD Match[b]
Age	28	0 (0)	0 (0)	0.99	Age	20	0 (0)	0 (0)	>0.99
BMI	18	4 (22)	2 (11)	0.49	BMI	22	2 (9)	2 (9)	0.64
BMD	12	3 (25)	3 (25)	0.09	BMD	22	2 (9)	2 (9)	0.11
25OHD	28	11 (39)	9 (32)	0.03	25OHD	22	0 (0)	0 (0)	0.81
CTx	18	8 (44)	4 (22)	0.36	CTx	14	0 (0)	0 (0)	0.40
PTH	26	5 (19)	3 (12)	<0.001	PTH	10	0 (0)	0 (0)	0.99
Vitamin D intake	11	6 (55)	2 (18)	0.03	sCR	22	0 (0)	0 (0)	>0.99
1,25OHD	26	8 (31)	6 (23)	<0.001	Weight	22	0 (0)	0 (0)	0.90
iCa	25	11 (44)	4 (16)	<0.001	YSM	20	0 (0)	0 (0)	0.81
Osteocalcin	20	6 (30)	4 (20)	0.02	Albumin	22	21 (95)	7 (32)	0.81

BMD, bone mineral density; BMI, body mass index; 25OHD, 25-hydroxyvitamin D; CTx, C-telopeptide; PTH, parathyroid hormone; Vitamin D intake, dietary vitamin D intake; 1,25OHD, 1,25-dihydroxyvitamin D; iCa, ionized calcium; SD, standard deviation.

[a] The columns represent the number where the mean/SD combination has identical matches for the same variable in different cohorts; and the largest number of matches for a single mean/SD combination.

[b] P refers to the probability that the reported number of matching mean/SD combinations for each variable (or a more extreme number of matches) occurred in 100,000 simulations.

Table 3. Proportion of identical summary statistics in the SatoIwamoto dataset, one fabricated dataset, and the reference Carlisle control dataset

	All variables	1 significant figure	2 significant figures	3 significant figures	4 significant figures	5 significant figures
SatoIwamoto dataset						
Number of variables, *n* (%)						
Mean	464	12 (2.6)	137 (29.5)	299 (64.4)	15 (3.2)	1 (0.2)
SD	464	79 (17.0)	296 (63.8)	86 (18.5)	3 (0.6)	0 (0.0)
Proportion of identical summary statistics in both treatment groups (%)						
Means match	21.8	91.7	30.7	15.7	6.7	0.0
SDs match	12.3	38.0	8.4	2.3	0.0	0.0
Both mean and SD match[a]	5	18.6	2.4	0.0	0.0	0.0
Fabricated dataset						
Number of variables, *n* (%)						
Mean	300	25 (8.3)	220 (73.3)	54 (18.0)	1 (0.3)	0.0
SD	300	148 (49.3)	104 (34.7)	48 (16.0)	0 (0.0)	0.0
Proportion of identical summary statistics in both treatment groups (%)						
Means match	6	12.0	5.9	3.7	0.0	0.0
SDs match	9.7	19.6	0.0	0.0	0.0	0.0
Both mean and SD match[a]	1.3	2.7	0.0	0.0	0.0	0.0
Carlisle reference dataset						
Number of variables, *n* (%)						
Mean	21,948	607 (2.8)	8149 (37.1)	10,978 (50.0)	2082 (9.5)	132 (0.6)
SD	21,145	4320 (20.4)	10,682 (50.5)	5618 (26.6)	552 (2.6)	36 (0.2)
Proportion of identical summary statistics in both treatment groups (%)						
Means match	13.4	66.4	20.6	7.5	1.6	0.8
SDs match	14.8	40.5	11.6	2.7	0.4	0.0
Both mean and SD match[a]	5.1	16.9	2.7	0.2	0.0	0.0

SD, standard deviation.

[a] Where the number of significant figures differed between the mean and SD, we used the number of significant figures for the mean to categorize the combination of mean and SD.

categorical baseline *P* values are markedly different in the Monticone dataset (reappraised, pval_cat_fn).

Because *P* values for categorical variables can be calculated directly from summary statistics, they can be independently calculated for all reported variables and compared with any reported *P* values. Different statistical tests can also be used for categorical data (eg, chi-square, Fisher's exact). When the statistical test is not described, reported *P* values can be compared to those calculated using a range of common tests. Table 4 shows that in the SatoIwamoto dataset, 35% of reported categorical baseline *P* values differed from the calculated *P* value. Where there are important differences, explanations should be sought [14].

In his original study, Carlisle found striking differences between the observed distribution of individual categorical variables in the 168 RCTs and the expected binomial distribution [7]. By contrast, the observed and expected distributions were similar for categorical variables from the 366 control RCTs [7]. Briefly, Carlisle's approach was that the number of trial arms with a frequency count for the variable of 0, 1, 2, … participants is summed and this distribution is compared to the expected (binomial) distribution of trial arms with these frequency counts. The expected distribution is calculated using the probability of the variable occurring in the trial and the number of participants in the trial arm, and then the results for all trial arms are summed. Worked examples are available [7,12]. The technique can be applied to baseline variables (eg, gender), or outcome variables unrelated to randomization. There were differences in the observed and expected distributions for participant withdrawals in the Asemi and SatoIwamoto datasets but not in the Auckland or REBALANCE control datasets [12]. Figure 2B shows that there were many more trial arms with three withdrawals and many fewer trial arms with one withdrawal than expected in the Asemi dataset (reappraised, cat_fn). A potentially more powerful approach is to consider between-group differences in two-arm trials rather than frequency counts. Figure 2D shows that for withdrawals in the Asemi dataset, there were far more trials with no difference and fewer trials with differences of $\geq$2 between trial arms than expected (reappraised, cat_fn).

A similar approach can be used to compare differences in numbers of participants between study arms in two-arm trials that use simple randomization [15]. When simple randomization is used, each participant has an equal (50%)

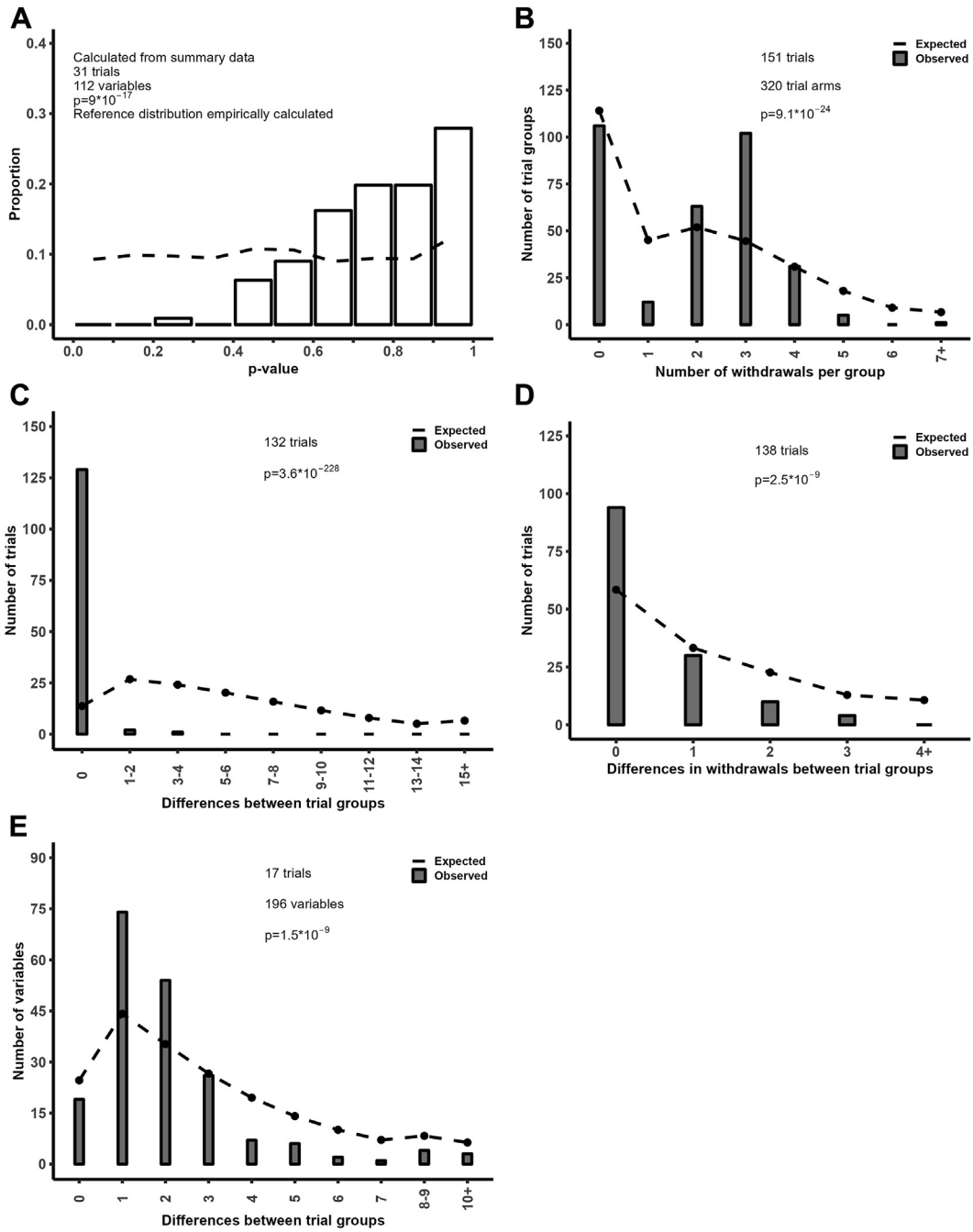

Figure 2. (A) Observed (bars) and empirically calculated expected (dotted line) distribution of categorical baseline *P* values in the Monticone dataset. Observed (bars) and expected (dotted line) distribution of: (B) withdrawals in trial arms in the Asemi dataset, between-arm differences for two-arm trials in: (C) number of participants in the Asemi dataset, (D) withdrawals in the Asemi dataset, and (E) frequency counts for all variables in the Monticone dataset.

Table 4. Difference between reported and calculated categorical baseline *P* values in the SatoIwamoto dataset

Difference between reported and calculated *P* values	*n* (%)	Test		Baseline *P* value				
		Chi-square	Fisher's exact	0.5–0.6	0.6–0.7	0.7–0.8	0.8–0.9	0.9–1
0	37 (65)	35	2	2	9	8	10	8
0–0.1	12 (21)	12		1	3		5	3
0.1–0.2	4 (7)	4		1			2	1
0.2–0.3	4 (7)	4				3		1

chance of being allocated to a treatment group, analogous to a coin toss. For an RCT with 10 people, there are 10 coin tosses. The single most likely outcome is five heads/five tails (five in each arm). However, a between-groups difference of 0 (probability = 0.25) is not the most likely outcome: a difference of 2 (four heads/six tails or six heads/four tails) is nearly twice as likely (probability = 0.41). Thus, the observed and expected between-group differences in numbers of participants can be compared. Figure 2C shows that there are far more trials than expected with the same number of participants in each trial arm in the Asemi dataset (reappraised, sr_fn). A similar approach can be taken for block/permuted randomization, if the final block size and number in the final block are known. Simple randomization can be undertaken using a single block for each trial arm, in which case there will always be no between-groups size difference if the blocks are filled. It would be expected that this randomization method would be detailed in the Methods.

The observed distribution of all categorical variables can be compared to the expected (binomial) distribution, using the same approach as for individual variables [14]. Frequency counts or percentages can be used, but between-group differences for two-arm trials seem to be more powerful. There were marked differences between the observed and expected between-group differences in frequency counts in the Monticone and SatoIwamoto datasets, but not in the Auckland and REBALANCE controls [14]. Figure 2E shows results for the Monticone dataset (reappraised, cat_all_fn).

3.4. Digit preference

Benford's law of the distribution of first digits and final digit preference have been used to examine individual patient data [2,3] or summary data [17]. Mol et al. reported unusual distributions of final digits in groups of RCTs, including one table where the final digit for all 47 variables was an even number [18,19]. Beyond such glaring examples, the role of this technique applied to summary data in groups of studies requires further research before it can be endorsed.

4. Assessing test performance

We assessed the performance of four techniques (distribution of continuous baseline *P* values, a single categorical variable, or numbers of trial participants, and matching summary statistics) in identifying deliberately fabricated data. We invited colleagues to make up summary baseline data for 20 hypothetical RCTs. They were told the purpose was to test the performance of tools developed to detect fabricated data and assess publication integrity, and were given brief, simple instructions: assume each trial used simple randomization; for each trial arm, fill out number of participants (20−2000), females and withdrawals, and the mean and SD for each of 15 continuous variables. Half were randomly assigned to receive reference ranges, and half received no further information. All data used 0−3 decimal places according to participant preference.

Fifteen people provided data (''cases''): two statisticians used statistical programs, another three individuals used random number generators, and the other 10 reported using no additional resources. The median time to provide data was 4 hours (range 1−8 hours). Twelve of 15 reported expertise of ≥3 on a five-point scale for interpreting clinical trial results (5 = high expertise). Fifteen ''control'' datasets were generated from the Auckland controls. For each control dataset, 20 trials were randomly selected with replacement, 40−4000 participants/trial randomly selected with replacement, treatment groups randomly assigned, and summary statistics calculated. All data were extracted from case and control datasets and run through five assessments (continuous baseline *P* values, withdrawals, gender, numbers of participants, and matching summary statistics) using development versions of reappraised package functions. Output was compiled into a single pdf, with the order of cases and controls randomly selected. This was given to two assessors who independently rated each assessment for each dataset as either integrity concern or not, and concerns overall for the dataset as low, intermediate, or high. Disagreements for overall concerns were resolved by consensus.

Twelve of 15 cases overall were considered as high risk of integrity concerns (three low risk), and 14/15 controls as low risk (one intermediate risk) (sensitivity 0.80, specificity 0.93, and accuracy 0.87).

For the five components of the assessment, the two assessors identified concerns for baseline *P* values in 12/15 and 12/15 cases, respectively, vs 1/15 and 0/15 in controls, gender (4/15, 6/15 cases, 0/15, 0/15 controls), withdrawals (0/15, 0/15 cases, 1/15, 1/15 controls), participants (11/15, 11/15 cases, 0/15, 2/15 controls), and matching (12/15, 11/15 cases, 2/15, 1/15 controls). Agreement between assessors was baseline *P* values (agreement 97%, kappa 0.93), gender (87%, 0.5), withdrawals (93%, 0.93), participants (93%, 0.86), matching (73%, 0.46), and overall (87%, 0.73).

4.1. Summary of test performance

The tests performed acceptably in discriminating cases with integrity concerns from controls without concerns, although the sample size was small. Only one control was rated as moderate concerns in participants and matching by one assessor, but the other assessor rated that control as no concerns for all tests. In 2/3 cases rated low risk, all data were created using random number generators in a statistical program, and in the other case, at least some data were created similarly in a spreadsheet.

Agreement between assessors overall and for individual tests was high. The best performing tests were the distributions of continuous baseline *P* values and participant numbers. Possible reasons for this include that baseline *P* values had the largest number of variables for analysis, that people providing data did not consider that participant numbers should follow an expected distribution, and that these techniques give a single *P* value which might make interpretation simpler.

5. When to use the tests

All the tests described require at least a moderate number of variables to allow reliable conclusions to be drawn. This usually needs a moderate number of RCTs. It is difficult to give definite numbers of variables and studies required, because they vary between tools and with the size of any differences between observed and expected distributions. In general, ≥50 variables are probably needed. Thus, these tools are only likely to be used for groups of RCTs. The data extraction required is time-intensive and labour-intensive. Therefore, it is most likely the tools would be used when concerns have been raised about RCT(s) by a research group. For example, if an RCT is retracted because of questionable research practices, then all RCTs by that group should be examined [20], and these tools could be useful. If systematic reviewers identify concerns about RCTs by investigators, the tools might aid in assessing the body of work by that investigator [7,15,18,21].

There is a potential role for machine-learning models in data extraction. It is straightforward to extract tables from HTML, word processor or spreadsheet files, or PDFs, but tables formatted as pictures cannot be extracted. A bigger problem is that table formats are not standardized, and it is not simple to extract data automatically into a 'tidy' format with one row for each variable and one separate column for each statistic (eg, mean) or other results (*P* values) for each trial arm. Furthermore, tables may contain different statistics within the same column (eg, mean/median) and the statistics presented can be ambiguous (eg, SD/standard error of the mean). Currently, it is often quicker to manually extract data. A machine-learning model tool that automatically extracted data into a 'tidy' format would be very valuable.

While these tools have limited utility for individual trials, journals could adapt the tools and ideas to screen submitted RCTs [5]. Although it is recommended that *P* values are not reported in baseline tables [22], submission of a supplementary baseline data table in a standard format with *P* values would allow automatic calculation and checking of *P* values and matching summary statistics. If concerns were identified, an explanation could be sought before the review process started. If individual data were provided in a standardized format, more sophisticated automated assessments could be undertaken. If full raw data were required at submission, even if not for publication, it could be kept securely by the journal and examined should concerns arise later. Examining individual data is more powerful for assessing publication integrity than using summary data [5].

6. Conclusion

A number of tools are now freely available to assist with assessing publication integrity concerns for groups of RCTs. Most of the tools compare whether the observed distributions of variables are consistent with distributions expected if successful randomization occurred. The tools are most likely to be used when concerns have been raised about ≥1 RCT(s), and an assessment of the broader body of work by the researcher(s) is undertaken.

Rights retention statement

CRediT authorship contribution statement

Mark J. Bolland: Writing − review & editing, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alison Avenell:** Writing − review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Andrew Grey:** Writing − review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Data availability

Data will be made available on request.

Declaration of competing interest

M.J.B. is a recipient of an HRC Clinical Practitioners Fellowship. The authors are independent of the HRC. The HRC had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. There are no competing interests for any other author.

References

[1] Cole GD, Nowbar AN, Mielewczik M, Shun-Shin MJ, Francis DP. Frequency of discrepancies in retracted clinical trial reports versus

unretracted reports: blinded case-control study. BMJ 2015;351:h4708.
[2] 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.
[3] Bordewijk EM, Li W, van Eekelen R, Wang R, Showell M, Mol BW, et al. Methods to assess research misconduct in health-related research: a scoping review. J Clin Epidemiol 2021;136:189−202.
[4] Carlisle JB. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. Anaesthesia 2017;72:944−52.
[5] Carlisle JB. False individual patient data and zombie randomised controlled trials submitted to Anaesthesia. Anaesthesia 2021;76:472−9.
[6] Pandit JJ. On statistical methods to test if sampling in trials is genuinely random. Anaesthesia 2012;67:456−62.
[7] Carlisle JB. The analysis of 168 randomised controlled trials to test data integrity. Anaesthesia 2012;67:521−37.
[8] 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.
[9] Carlisle JB, Loadsman JA. Evidence for non-random sampling in randomised, controlled trials by Yuhji Saitoh. Anaesthesia 2017;72:17−27.
[10] 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.
[11] 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.
[12] Bolland MJ, Gamble GD, Avenell A, Cooper DJ, Grey A. Participant withdrawals were unusually distributed in randomized trials with integrity concerns: a statistical investigation. J Clin Epidemiol 2021;131:22−9.
[13] Bolland MJ, Gamble GD, Avenell A, Grey A. Identical summary statistics were uncommon in randomized trials and cohort studies. J Clin Epidemiol 2021;136:180−8.
[14] Bolland MJ, Gamble GD, Avenell A, Cooper DJ, Grey A. Distributions of baseline categorical variables were different from the expected distributions in randomized trials with integrity concerns. J Clin Epidemiol 2023;154:117−24.
[15] 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.
[16] Bolland MJ, Gamble GD, Grey A, Avenell A. Empirically generated reference proportions for baseline p values from rounded summary statistics. Anaesthesia 2020;75:1685−7.
[17] 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.
[18] Bordewijk EM, Wang R, Askie LM, Gurrin LC, Thornton JG, van Wely M, et al. Data integrity of 35 randomised controlled trials in women' health. Eur J Obstet Gynecol Reprod Biol 2020;249:72−83.
[19] Muriithi FG, Gurrin LC, Mol BW, Thornton JG. An investigation of 51 publications by a single author due to doubts about data integrity. PREPRINT (Version 1) available at Research Square. 2022. Available at: https://doi.org/10.21203/rs.3.rs-1539633/v1. Accessed May 3, 2024.
[20] Sox HC, Rennie D. Research misconduct, retraction, and cleansing the medical literature: lessons from the Poehlman case. Ann Intern Med 2006;144:609−13.
[21] O'Connell NE, Moore RA, Stewart G, Fisher E, Hearn L, Eccleston C, et al. Investigating the veracity of a sample of divergent published trial data in spinal pain. Pain 2023;164:72−83.
[22] Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.