Title: Identical summary statistics were uncommon in randomised trials and cohort studies

Authors:

Mark J Bolland, MBChB, PhD; m.bolland@auckland.ac.nz
Greg D Gamble, MSc; gd.gamble@auckland.ac.nz
Alison Avenell, MBBS, MD; a.avenell@abdn.ac.uk
Andrew Grey, MD; a.grey@auckland.ac.nz

Author Affiliation:

Mark Bolland, Greg Gamble and Andrew Grey- Department of Medicine, University of Auckland, Private Bag 92 019, Auckland 1142, New Zealand
Mark Bolland, Department of Endocrinology, ADHB, Private Bag 92 024, Auckland 1142, New Zealand
Alison Avenell- Health Services Research Unit, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, Scotland

Word counts:

Text only: 3166
Abstract: 200
Figures: 0
Tables: 4
Appendix: Reference list, Supplementary Methods, Results, Discussion, Figure, Table

Address for correspondence:

Mark Bolland
Bone and Joint Research Group
Department of Medicine
Faculty of Medical and Health Sciences,
University of Auckland
Private Bag 92019, Auckland, New Zealand
Tel: (+64 9) 3737 599 extn 83004
Fax: (+64 9) 3737 677
email: m.bolland@auckland.ac.nz

Key words: Statistical Methods, Research Integrity, Identical Data, Summary Statistics, Fabricated Data, Data Integrity

What is new?

- In control datasets, the probability of both an identical mean and an identical SD for a baseline variable in separate randomised controlled trials or within an individual trial is low (<~3%)
- Variables rounded to 1 significant figure or with small standard deviations have a higher proportion of identical summary statistics
- We present an example of two randomised controlled trials with an improbably high proportion of identical summary statistics based on simulations and an example of an improbably high proportion of recurrent identical summary statistics in 34 independent cohort studies.
- An unexpectedly high proportion of identical summary statistics may raise a "red flag" for concerns about publication integrity.

Abstract:

Objective:

To examine the proposition that identical summary statistics (mean and/or SD) in different randomised controlled trials (RCTs) or clinical cohorts can be explained by common or homogeneous source populations.

Study design:

We estimated the probability of identical summary data in studies with high proportions of identical summary statistics, in simulations, and in control datasets.

Results:

The probability of both an identical mean and an identical SD for a variable in separate RCTs is low (<~3%), unless the variable is rounded to 1 significant figure. In two RCTs with identical summary statistics for 16/39 shared variables, simulations indicated the probability of the observed matches was <1/100,000. In 34 clinical cohorts with publication integrity concerns, the proportion of summary statistics from variables reported in ≥10 studies that were identical in ≥2 cohorts was high (42% for means, 52% for SDs, and 29% for both), and improbable based on simulations and comparisons to control datasets.

Conclusions:

The likelihood of multiple identical summary statistics within an individual RCT or across a body of RCTs or cohort studies by the same research group is low, especially when both the mean and the SD are identical, unless the variables are rounded to 1 significant figure.

1.Introduction

The detection of unusual patterns in data and results is part of the process of assessing publication integrity, but there is only a small literature on specific methods that can be applied. When raising concerns about the integrity of publications of randomised controlled trials (RCTs), we and others notified journals about unusual distributions or similarities in summary data (mean and standard deviation (SD)) for baseline or outcome variables in separate RCTs, or baseline variables within the same RCT [1-5]. In response, authors stated that this is expected because of the homogeneity of the source population [6-8]. When we raised concerns about a similar issue- the frequent occurrence of identical means and/or SDs for several baseline variables in independent groups of participants in several different observational studies reported by the same group of investigators- the concerns were not addressed in journal responses or correction notices.

In an example of the first situation (example 1.1), separate RCTs were reported as being conducted in five groups of ten 6-week female rats [9, 10] and three groups of ten 8-week female rats, respectively [11]. The senior author confirmed to the journal editors that the trials were two separate RCTs, carried out in separate groups of rats, despite the similarity in the trial reports [7]. Three treatments were the same in each RCT, and 2 baseline and 11 outcome variables were reported in both RCTs for these treatment groups. Therefore, there were 39 baseline or outcome variables in common, of which 16 had both an identical mean and an identical SD (n=6), or an identical mean but different SD (n=4), or an identical SD but different mean (n=6) in the 2 RCTs. The authors stated these results were to be expected because the rat populations were similar [7].

In the second situation (example 1.2), a group of investigators reported data for a range of variables in independent clinical studies whose participants were of the same ethnicity and from the same geographical region. Among 34 cohorts from 33 publications (Supplementary References), 6 variables had both a mean and a SD that were identical in at least 6 different cohorts. For each of these variables, there were also multiple instances in different cohorts of identical means with different SDs, or identical SDs with different means.

We wondered if the explanation provided by the authors is plausible: that is, it is expected that identical summary data (mean and/or SD) occur frequently in independent samples drawn from a common population. We used data from our own and others' RCTs of clinical and basic research into bone health, and simulations, to model the probability of the occurrence of identical summary data in each situation, to identify factors which might affect that probability, and to test whether comparing proportions of identical summary statistics might be used broadly in assessing publication integrity.

2.Methods:

2.1.1 Identical summary statistics in separate animal RCTs (cases)

To assess the first situation (example 1.1), of identical summary statistics in two separate RCTs, we estimated the probability of 16 variables in two independent datasets containing 10 animals having the same baseline or outcome mean and/or SD, using summary statistics from two publications from the first RCT [9, 10], and one publication from the second RCT [11]. The full methods are described in the Supplementary Methods, but briefly, we conducted 100,000 simulations in which summary statistics for two groups were generated using normally distributed random numbers based on the reported summary statistics in the publication and the proportion of identical summary statistics calculated.

2.1.2 Identical summary statistics in different animal RCTs from the same population (controls)

Next, we modelled the scenario that the population from which separate treatment groups was drawn was very similar. We used individual raw data from studies in rats conducted in our own laboratory (Auckland laboratory dataset) [12, 13] for similar bone histomorphometry variables to those reported in the RCTs in section 2.1.1. Briefly, we duplicated data for the original groups of animals to form two identical groups, re-randomised this population into two treatment groups 1,000,000 times, and calculated the proportion of identical summary statistics when common rounding (1-2 decimal places) or more extreme rounding (whole numbers or 1 decimal place) was used (see Supplementary Methods).

2.2 Identical summary statistics in different clinical studies from similar populations

2.2.1 Cases

We assessed the second situation (example 1.2) of multiple occurrences of identical mean and/or SD values in independent groups of individuals drawn from the same source population in a dataset of clinical bone studies with known concerns about publication integrity. For 34 cohorts in 33 publications by the research group of Y Sato and colleagues (Supplementary References), which has multiple retractions for a wide variety of reasons, we assessed the probability of identical summary statistics in the same variables in different cohorts with, or at risk of, osteoporosis. In these 34 cohorts, 26 baseline variables were reported at least twice, and 10 were reported in ≥10 cohorts. We restricted the analyses to these 10 variables. We used the methods described in section 2.1.1 (see Supplementary Methods) to perform 100,000 simulations and calculate the probability of obtaining multiple occurrences of identical summary data in the different cohorts.

2.2.2 Controls

For comparison, we repeated these analyses in 9 RCTs with osteoporosis outcomes in older women published by our group (Auckland clinical dataset) [14-22]. We treated each randomised group in an RCT as a separate cohort, giving a total of 22 cohorts. Of the 26 variables reported by Sato and colleagues, 16 were reported in these RCTs, and we again restricted analyses to 10 variables reported in ≥10 cohorts. We rounded the summary statistics for each variable in each cohort to the same number of decimal places used by Sato.

2.3 Identical summary statistics within an RCT and in different cohorts

We wondered whether these approaches could be used more broadly to identify datasets with integrity concerns. We assessed the proportions of identical summary statistics within individual RCTs (analogous to section 2.1.2) in a large dataset- the "Carlisle dataset" containing summary data on continuous baseline variables in 5087 RCTs published in 8 general medical and anaesthesia journals between 2000 and 2015 [3] (controls), and in two datasets of RCTs with concerns about publication integrity (cases) (see Supplementary Methods). We then repeated the analyses on identical summary statistics in different cohorts (analogous to section 2.2) in a larger set of variables in the Auckland clinical dataset (controls), and in these two sets of RCTs with concerns about publication integrity (cases).

All simulations and analyses were performed used SAS (SAS Institute, Cary, NC version 9.4) or the R software packages (R 3.5.1, 2019, R Foundation for Statistical Computing, Vienna, Austria).

3.Results

3.1.1 Identical summary statistics in separate animal RCTs (cases)

Table 1 shows the probabilities of obtaining identical means and/or SDs in 100,000 simulations of two independent studies of 10 rats for 16 variables based on the reported summary statistics of two studies. These are extremely conservative estimates because, for each variable, we calculated the probability of *any* identical mean and/or SD, not the probability of the *observed* identical summary statistics. For example, the probability that the HxGH group in both studies had identical means and SDs for periosteal surface MAR is $3*10^{-4}$ (Table 1), but in 10 million simulations, the probability that each treatment group would have the *observed* mean and SD of 3.17 and 0.68, is about 1500 times lower ($2*10^{-7}$).

The observed total number of identical means, SDs, and of both identical mean and identical SD among the 16 variables was 10, 12, and 6 respectively. In 100,000 simulations, the largest number of corresponding matches in a single simulation was 5 (2 simulations), 4 (14 simulations), and 2 (3 simulations) respectively. Thus, the probability of the observed number of matches for each statistic was very small (P<1/100,000). By contrast, the probability of no identical means or SDs among the 16 variables was 0.49 (Table 1).

Table 1 also shows that 6 means or SDs had differences in rounding between publications, and in 4 instances, these means (n=3) or SDs (n=1) became identical when rounded to the smaller number of decimal places. We repeated the analyses after rounding the mean or SD of these variables to the smaller number of decimal places. 10/16, 13/16, and 13/16 variables had both identical mean and identical SD, identical means, or identical SDs respectively. In 100,000 simulations, the largest number of corresponding matches in a single simulation was 2, 6, and 5 respectively (P<1/100,000 for each statistic). The probability of no identical means or SDs among the 16 variables was 0.41.

Table 1: Probability of identical summary statistics from 100,000 simulations of data for 16 variables in two RCTs of 10 animals in each trial treatment group

	Reported Data						**Data from simulations**				
	Trial 1 [9, 10]		**Trial 2 [11]**		**Population**		**Probability of identical summary statistics**			**Probability of no matches**	
Variable[a] (treatment group)[b]	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Means match (%)**	**SDs match (%)**	**Mean and SD match (%)**	**Relevant Probability[c]**	**Cumulative Probability**
Weight (Intact)	**192**	7.7	**192**	11.7	192	9.7	**9.23**	1.2	0.1	0.90	0.90
Weight (HxGH)	190	**10.8**	180	**10.8**	185	10.8	8.16	**1.15**	0.1	0.91	0.81
Final weight (Intact)	255	**18.3**	244.8	**18.3**	249.9	18.3	0.52	**0.66**	<0.01	0.99	0.80
Tibial shaft CSA cortical % (Intact)[d]	*82.26*	**1.77**	*82.3*	**1.77**	82.28	1.77	0.52	**0.66**	0.01	0.99	0.80
Tibial shaft CSA cortical % (HxGH)[d]	*85.35*	**1.89**	*85.4*	**1.89**	85.38	1.89	0.49	**0.62**	0.01	0.99	0.79
Periosteal surface MS/BS (Intact)	**75.8**	**9.7**	**75.8**	**9.7**	75.8	9.7	0.92	1.28	**0.01**	0.98	0.77
Periosteal surface MS/BS (HxGH)	*90.9*	**3.8**	*90.87*	**3.8**	90.89	3.8	0.25	**3.25**	0.01	0.97	0.74
Periosteal surface MAR (Intact)	**2.78**	**0.28**	**2.78**	**0.28**	2.78	0.28	3.16	4.46	**0.14**	0.93	0.69
Periosteal surface MAR (Hx)	**1.02**	0.30	**1.02**	0.19	1.02	0.25	**3.54**	4.84	0.18	0.92	0.63
Periosteal surface MAR (HxGH)	**3.17**	**0.68**	**3.17**	**0.68**	3.17	0.68	1.3	1.75	**0.03**	0.97	0.61
Periosteal surface BFR (Intact)	**210**	29.3	**210**	39.3	210	34.3	**2.59**	0.36	0.01	0.97	0.59
Periosteal surface BFR (HxGH)	**288**	**64.5**	**288**	**64.5**	288	64.5	1.33	0.21	**<0.01**	0.98	0.58
Endocortical surface MS/BS (HxGH)	**56.3**	*11.62*	**56.3**	*11.6*	56.3	11.61	**0.75**	0.11	<0.01	0.99	0.58
Endocortical surface MAR rate (Intact)	**1.45**	**0.21**	**1.45**	**0.21**	1.45	0.21	4.22	5.83	**0.24**	0.90	0.52
Endocortical surface MAR (HxGH)	**1.65**	**0.34**	**1.65**	**0.34**	1.65	0.34	2.66	3.53	**0.1**	0.94	0.49
Endocortical surface BFR (HxGH)	97.58	**32.9**	94.6	**32.9**	96.09	32.9	0.02	**0.36**	<0.01	>0.99	0.49

Data for mean/SD in bold are identical in both trial reports, and in italics indicates data where the means or SDs become identical when rounded to the smallest number of decimal places presented in the different publications. The probability of identical summary statistics in bold indicates the identical summary statistic(s).

[a] Abbreviations CSA- cross sectional area, MS/BS mineral surface/bone surface, MAR mineral apposition rate, BFR bone formation rate

[b] The three treatment groups common to each trial were control group (intact), hypophysectomy (Hx), hypophysectomy followed by growth hormone treatment (HxGH)

[c] Relevant probability is the probability of there being no identical summary statistic (either mean or SD) for the variable.

[d] Data for Trial 1 from [10]

3.1.2 Identical summary statistics in different animal RCTs from the same population (controls)

Table 2 shows summary statistics for raw individual data from the Auckland laboratory dataset for similar bone histomorphometry variables to those in Table 1. Even with a population formed from two identical groups, in 1,000,000 re-randomisations of the individual raw data, the probability of getting identical summary statistics is generally low, increases when the SD is small and when data have more extreme rounding to fewer decimal places. Table 3 also shows the probability of getting 0, 1, 2 or 3 identical summary statistics among the 3 different variables in each simulated randomisation. The likelihood of ≥1 variable with both identical mean and identical SD in a simulated randomisation was generally low (<5%) except when variables were highly rounded.

Table 2: Probability of identical summary statistics in 1,000,000 re-randomisations of two identical treatment groups from control animal RCTs

		Common rounding **Probability of identical summary statistics**					**Extreme rounding** **Probability of identical summary statistics**			
Variable	**Raw data Mean (SD)**	**Means Match (%)**	**SD match (%)**	**Mean and SD match (%)**	**Mean/SD match raw data (%)**	**Extreme Rounded Mean (SD)**	**Means Match (%)**	**SD match (%)**	**Mean and SD match (%)**	**Mean/SD match raw data (%)**
6 month old rats (n=9)										
BFR/BS	10.1 (4.7)	2.6	4.3	0.23	0.02	10 (5)	24.9	39.6	10.4	3.8
MAR	0.77 (0.19)	6.4	11.4	1.2	0.14	0.8 (0.2)	51.4	67.6	41.1	35.3
MS/BS	13.0 (5.6)	2.2	3.4	0.27	0.03	13 (6)	21.0	30.2	7.3	2.1
Number of identical summary statistics per simulated randomisation										
0 matches		89.2	82.0	98.4	99.8		28.8	13.6	48.9	61.0
1 match		10.5	17.0	1.64	0.19		47.7	43.4	43.7	36.9
2 matches		0.34	0.98	0.01	<0.001		20.8	34.9	7.1	2.09
3 matches		0.004	0.02	<0.001	<0.001		2.67	8.1	0.31	0.03
6 week old mice (n=11)										
BFR/BS	51.0 (7.0)	1.9	4.3	0.11	<0.001	51 (7)	18.9	39.9	8.0	2.6

MAR	1.31 (0.12)	11.3	19.9	2.5	0.35	1.3 (0.1)	87.7	99.9	87.5	87.2
MS/BS	0.68 (0.14)	9.3	19.0	1.9	0.17	0.7 (0.1)	67.2	64.4	42.0	34.6
Number of identical summary statistics per simulated randomisation										
0 matches		78.9	62.1	95.6	99.5		3.3	0.0	6.7	8.2
1 match		19.7	32.8	4.4	0.52		30.8	21.4	52.1	60.1
2 matches		1.41	5.0	0.05	0.001		54.8	52.9	38.3	31.0
3 matches		0.02	0.16	<0.001	<0.001		11.1	25.7	2.9	0.78
37 week old mice (n=9)										
BFR/BS	67.8 (10.5)	1.2	2.0	0.19	0.02	68 (11)	11.5	17.9	2.3	0.28
MAR	3.31 (0.78)	1.6	2.8	0.12	0.01	3.3 (0.8)	15.4	25.6	4.3	0.79
MS/BS	2.23 (0.57)	2.2	4.9	0.60	0.10	2.2 (0.6)	20.6	30.8	7.9	3.0
Number of identical summary statistics										
0 matches		95.1	90.6	99.1	99.9		59.4	42.3	86.1	96.0
1 match		4.9	9.1	0.90	0.14		34.0	42.6	13.3	4.0
2 matches		0.08	0.28	0.002	<0.001		6.2	13.7	0.58	0.03
3 matches		0.00	0.00	<0.001	<0.001		0.36	1.39	0.01	<0.001

Abbreviations BFR/BS bone formation rate/bone surface, MAR mineral apposition rate, MS/BS mineral surface/bone surface.

3.2 Identical summary statistics in different clinical studies from similar populations

3.2.1 Cases

Table 3 shows the number of identical summary statistics for 10 baseline variables reported in ≥10 cohorts from 34 clinical cohorts. Of 226 reported mean values, 95 (42%) had an identical value in ≥1 other cohort; 110/212 (52%) reported SD values and 62/212 (29%) reported combinations of mean and SD values had an identical value or combination in ≥1 other cohort. In 100,000 simulations, the estimated probability of the observed or more extreme number of identical summary statistics for individual variables was frequently low (11 P<0.001; 18 P<0.1). For the 10 variables, the observed distribution of both identical mean and identical SD was different from the expected distribution based on simulations: only 1 variable (age) had no identical mean/SD combinations whereas the expected number (based on simulations) was 4.2, while 9 variables had ≥2 identical mean/SD combinations, whereas the expected number was 5.8 (P = 0.04).

3.2.2 Controls

By comparison, in the Auckland clinical studies conducted in similar populations, of 196 summary statistics, 49 (25%) of means recurred, 87 (44%) of SD recurred, and 25 (13%) of the combination of mean/SD recurred. The estimated probability of the observed identical summary statistics was uniformly distributed (Supplementary Table). In addition, the observed distribution of identical means, identical SDs and both identical mean and identical SD were consistent with the expected distribution from 100,000 simulations, (data not shown, P>0.28 for each statistic). The observed distribution for 2 of the summary data differed between the cases and controls (Supplementary Figure).

Table 3: the proportion of recurring identical summary statistics in 34 cohorts with concerns regarding publication integrity

Variable	Mean (N)[a]	Mean Match N (%)[a]	Largest Number Matches N (%)[a]	SD (N)[a]	SD Match N (%)[a]	Largest Number Matches N (%)[a]	Mean and SD Match N (%)[a]	Largest Number Matches N (%)[a]	Mean (SD) Mode/ Average[b]	Mode Match N[b]	P Mean Match[c]	P SD Match[c]	P Mean and SD Match[c]
Age	31	5 (16)	3 (10)	28	12 (43)	2 (7)	0 (0)	0 (0)	72.14 (5.13)		0.27	0.01	0.99
Body mass index	18	6 (33)	2 (11)	18	8 (44)	4 (22)	4 (22)	2 (11)	22.3 (2)	2	<0.001	<0.001	0.49
Dietary vitamin D	11	6 (55)	2 (18)	11	8 (73)	2 (18)	6 (55)	2 (18)	114 (20)	2	0.29	0.49	0.03
Bone density	13	4 (31)	4 (31)	12	5 (42)	3 (25)	3 (25)	3 (25)	2.55 (0.36)	3	0.11	0.05	0.09
25-hydroxyvitamin D	31	14 (45)	10 (32)	28	17 (61)	10 (36)	11 (39)	9 (32)	21.6 (3.1)	9	<0.001	<0.001	0.03
1,25-dihydroxyvitamin D	27	15 (56)	9 (33)	26	13 (50)	7 (27)	8 (31)	6 (23)	49.6 (9.2)	6	0.13	0.43	<0.001
C-Telopeptide	19	8 (42)	6 (32)	18	11 (61)	6 (33)	8 (44)	4 (22)	7.1 (1.1)	4	<0.001	<0.001	0.36
Ionised Ca	27	19 (70)	11 (41)	25	18 (72)	6 (24)	11 (44)	4 (16)	1.22 (0.04)	4	<0.001	<0.001	<0.001
Parathyroid hormone	28	8 (29)	3 (11)	26	8 (31)	3 (12)	5 (19)	3 (12)	37 (17.4)	3	0.52	0.35	<0.001
Osteocalcin	21	10 (48)	6 (29)	20	10 (50)	6 (30)	6 (30)	4 (20)	7 (4.5)	4	0.27	0.04	0.02

[a] The columns represent the number of reported means and SD; the number of means, SD, and both mean and SD that recurred amongst the different cohorts (eg for age the mean value of 72.4 occurred 3 times, and 68.8 2 times in the 34 cohorts giving a total of 5 matching means); and the largest number of matches for each statistic.

[b] where there was no recurring mean (SD) combination, the average mean and SD in the 34 cohorts weighted by numbers of participant are reported, otherwise the most common mean (SD), ie the mode, is reported together with the number of occurrences of the mode.

[c] P refers to the probability of the reported number of identical means, identical SDs or both identical mean and identical SD for each variable or a more extreme number of matches (relative to the most common number of matches) in 100,000 simulations.

3.3 Identical summary statistics within an RCT and in different cohorts

To test whether these techniques could be used more broadly, we first assessed the proportion of identical summary statistics within a single RCT in the Carlisle dataset [3] (controls). Table 4 shows that the proportion of identical statistics in RCTs is heavily dependent on the degree of rounding, but, consistent with the previous analyses, becomes small when typical degrees of rounding are used, especially when both the mean and the SD are identical. We repeated these analyses in two sets of RCTs for which concerns about publication integrity have been raised (cases), but while there were some differences between the proportions of identical summary statistics compared to the reference Carlisle dataset, the differences were neither consistent or large enough to convincingly support concerns about publication integrity (data not shown).

Table 4: Proportion of identical summary statistics in the Carlisle control dataset of 3599 RCTs by degree of rounding

	All variables	1 significant figure	2 significant figures	3 significant figures	4 significant figures	5 significant figures
Number of variables, n (%)						
Mean	22020	614 (2.8)	8175 (37.1)	11010 (50.0)	2088 (9.5)	133 (0.6)
SD	21247	4310 (20.3)	10725 (50.5)	5621 (26.5)	554 (2.6)	37 (0.2)
Proportion of identical summary statistics in both treatment groups (%)						
Means match	13.4	66.6	20.7	7.5	1.6	0.8
SDs match	14.9	40.8	11.6	2.7	0.4	0
Both Mean and SD match[a]	5.1	17.2	2.7	0.3	0	0

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

Lastly, we repeated the analyses for identical summary statistics in different cohorts in a larger set of variables in the control RCTs, and in 2 sets of RCTs with publication integrity concerns (cases) (see Supplementary Results). For the controls, the probability of the observed number of identical summary statistics was frequently low, and the observed distributions of identical summary data were not consistent with expected distributions. Further, we did not find any consistent or sufficiently large differences between the controls and the cases that could be used to support concerns about publication integrity.

4.Discussion

Using several different techniques, we found the likelihood of multiple instances of identical summary data for baseline or outcome variables in different RCTs or cohorts recruited from similar populations is very low, unless variables are consistently rounded to a small number of significant figures. Firstly, we used reported summary data to generate normally distributed random numbers for individual observations and calculated summary statistics for multiple simulations. Secondly, we used actual raw data and calculated summary statistics after multiple re-randomisations. Thirdly, we calculated the proportions of identical summary statistics in sets of cohort studies and within individual RCTs from a very large set of published RCTs. All 3 approaches produced similar results. In control studies, the probability of both an identical mean and an identical SD for a *single* variable is low (less than about 3%), decreasing when the summary statistics are presented with ≥2 significant figures and increasing when presented with only 1 significant figure. In addition, variables with proportionately small SDs have a higher proportion of identical summary statistics than variables with large SDs. The probability of *multiple* occurrences of identical summary statistics is much lower than that for a single match, especially when both the mean and the SD are identical.

Rounding has the greatest effect on the proportion of identical summary statistics. When a mean and SD are both reported with ≥2 significant figures, the likelihood of both an identical mean and SD becomes very low. In contrast, when summary statistics are presented to only 1 significant figure, high proportions of matching can occur (see Supplementary Discussion for example). Of note, the examples of identical summary statistics in published papers in Tables 1 and 3 did not include highly rounded variables presented using only 1 significant figure. The size of the SD also affects the proportion of identical summary statistics (see example in Supplementary Discussion), but Table 1 shows that the effect of rounding is an order of magnitude greater than that of SD size. Standardisation of rounding in journals [23] would improve both the accessibility of presented data and the usefulness of techniques assessing matching data or baseline p-values [24].

In contrast to the results from control datasets, two animal RCTs stated to be independent studies had 16 variables with identical summary statistics, which we estimated to be extremely improbable (P<1/100,000), especially given the estimated probability of no identical summary statistics among the 16 variables was 0.49. In a group of 34 cohorts from publications with integrity concerns, the proportion of identical summary statistics differed from the expected proportions in simulations and control cohorts, especially for the variables with the highest numbers of identical summary statistics. The least likely occurrence when there are identical summary statistics is that both the mean and SD are identical. However, in both the two animal RCTs and the 34 cohort studies, there were numerous occurrences of both identical means and SDs which is highly improbable with the degree of rounding of the relevant variables. The proportions from the very large Carlisle dataset of RCTs can serve as a reference: about <3% of variables in an individual RCT reported to 2 significant figures will have an identical mean and an identical SD. If variables are presented with more significant figures, this occurs even less frequently.

There are limitations to the techniques. The simulations generated normally distributed random numbers, but some variables are not normally distributed. The degree of rounding has a substantial impact on the proportions of identical summary data and should be considered in any analysis. Assessment of raw data would be very helpful in such a situation. Since rounding often varies between variables, publications, and individual summary statistics (eg mean values are often reported to more significant values than SDs), comparisons between variables in different studies may be difficult. Comparing proportions of identical summary data in cohort studies also requires a moderate number of studies that report the same variables. Missing data or exclusion of individual results (eg completers analysis) might impact upon the results. Finally, if these techniques were to be routinely used to support concerns about data integrity of publications, validation of the approach and results in other datasets would be important.

We had wondered whether the amount of identical summary data could be used routinely in the assessment of concerns regarding publication integrity, but the differences between control datasets and datasets with concerns were not consistent or large enough to support this. This might be for a few reasons: the assumptions underlying the techniques (eg normally distributed variables; that variables in different cohorts are derived from the same population) might be incorrect; treating individual arms of RCTs as separate cohort studies might affect the results; the inconsistent rounding in variables within and between different studies might obscure unusual proportions of identical summary data for individual variables; and finally, the datasets we assessed might not have unusual proportions of identical summary data.

Considering these limitations, an important question is when the techniques should be applied. It is simple to calculate the proportion of identical summary statistics, so it could be used easily and widely. However, large simulations may need specialist statistical software and moderate computational power. Like other statistical tests used to assess data integrity, these techniques do not produce definitive results. Thus, they would probably be most useful when applied after a high proportion of identical summary statistics was noted incidentally, eg during the assessment of concerns about publication integrity or when a systematic review involving the affected studies is undertaken. If a high proportion of identical summary statistics is found that is not obviously explained by rounding, further investigation is warranted, including examination of raw data. If the unusual proportions of identical summary statistics are confirmed and not explicable, it is reasonable to examine other aspects of publication integrity [25]. Similar approaches have been used previously to find a high proportion of similar summary statistics in publications that subsequently were shown to be compromised by data fabrication [26].

In summary, our results suggest that cohorts being derived from the same source population is not an adequate explanation for a high proportion of identical summary statistics. Furthermore, the likelihood of multiple identical summary statistics (mean and/or SD) for variables within an individual RCT or across a body of clinical studies (RCTs or cohort studies) by the same research group is also low, unless the variables are rounded to only 1 significant figure. This is particularly the case when both the mean and SD are identical. Deviations from the expected patterns derived from simulations and analyses of valid data represent a ‘red flag’ for publication integrity, and in such cases seeking an explanation and examining raw data is warranted.

Acknowledgements

We thank Dorit Naot and Susannah O'Sullivan for providing the animal raw data used in the simulations.

Funding

This research received no specific funding. MB is a recipient of an HRC Clinical Practitioners Fellowship. The Health Services Research Unit is funded by the Chief Scientist Office of the Scottish Government Health and Social Care Directorates. 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.

Competing interest

None of the authors have a conflict of interest to declare.

Authors contributions

MB, GG, AA, and AG designed the research. MB and provided the existing datasets and MB extracted any new data. MB and GG performed the analyses. MB drafted the paper. All authors critically reviewed and improved it. All authors read and approved the final manuscript.

References

1. Carlisle JB. The analysis of 168 randomised controlled trials to test data integrity. Anaesthesia. 2012;67:521-37.

2. Bolland MJ, Avenell A, Gamble GD, Grey A. Systematic review and statistical analysis of the integrity of 33 randomized controlled trials. Neurology. 2016;87:2391-2402.

3. Carlisle JB. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. Anaesthesia. 2017;72:944-952.

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

5. 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.

6. Halbekath JM, Schenk S, von Maxen A, Meyer G, Muhlhauser I. Risedronate for the prevention of hip fractures: concern about validity of trials. Arch Intern Med. 2007;167:513-4; author reply 514-5.

7. Personal correspondence from J Bone Miner Metab Editors. Response from Senior Author to J Bone Miner Metab Editors regarding concerns raised about Iglesias L, Yeh JK, Castro-Magana M, Aloia JF. Effects of growth hormone on bone modeling and remodeling in hypophysectomized young female rats: a bone histomorphometric study. J. Bone Miner. Metab. 2011; 29:159-167. 2019

8. Badawy A. Data integrity of randomized controlled trials: A hate speech or scientific work? Eur J Obstet Gynecol Reprod Biol. 2020;255:259.

9. Chaudhry AA, Castro-Magana M, Aloia JF, Yeh JK. Differential effects of growth hormone and alpha calcidol on trabecular and cortical bones in hypophysectomized rats. Pediatr Res. 2009;65:403-8.

10. Guevarra MS, Yeh JK, Castro Magana M, Aloia JF. Synergistic effect of parathyroid hormone and growth hormone on trabecular and cortical bone formation in hypophysectomized rats. Horm Res Paediatr. 2010;73:248-57.

11. Iglesias L, Yeh JK, Castro-Magana M, Aloia JF. Effects of growth hormone on bone modeling and remodeling in hypophysectomized young female rats: a bone histomorphometric study. J Bone Miner Metab. 2011;29:159-67.

12. O'Sullivan S, Naot D, Callon KE, Watson M, Gamble GD, Ladefoged M et al. Imatinib mesylate does not increase bone volume in vivo. Calcif Tissue Int. 2011;88:16-22.

13. Naot D, Watson M, Callon KE, Tuari D, Musson DS, Choi AJ et al. Reduced Bone Density and Cortical Bone Indices in Female Adiponectin-Knockout Mice. Endocrinology. 2016;157:3550-61.

14. Reid IR, Ames RW, Evans MC, Gamble GD, Sharpe SJ. Effect of calcium supplementation on bone loss in postmenopausal women. N Engl J Med. 1993;328:460-4.

15. Reid IR, Ames RW, Orr-Walker BJ, Clearwater JM, Horne AM, Evans MC et al. Hydrochlorothiazide reduces loss of cortical bone in normal postmenopausal women: A randomized controlled trial. Am J Med. 2000;109:362-370.

16. Reid IR, Lucas J, Wattie D, Horne A, Bolland M, Gamble GD et al. Effects of a beta-blocker on bone turnover in normal postmenopausal women: a randomized controlled trial. J Clin Endocrinol Metab. 2005;90:5212-6.

17. Reid IR, Mason B, Horne A, Ames R, Reid HE, Bava U et al. Randomized controlled trial of calcium in healthy older women. Am J Med. 2006;119:777-85.

18. Grey A, Bolland M, Gamble G, Wattie D, Horne A, Davidson J et al. The peroxisome proliferator-activated receptor-gamma agonist rosiglitazone decreases bone formation and bone mineral density in healthy postmenopausal women: a randomized, controlled trial. J Clin Endocrinol Metab. 2007;92:1305-10.

19. Reid IR, Cundy T, Grey AB, Horne A, Clearwater J, Ames R et al. Addition of monofluorophosphate to estrogen therapy in postmenopausal osteoporosis: a randomized controlled trial. J Clin Endocrinol Metab. 2007;92:2446-52.

20. Grey A, Bolland MJ, Wattie D, Horne A, Gamble G, Reid IR. The antiresorptive effects of a single dose of zoledronate persist for two years: a randomized, placebo-controlled trial in osteopenic postmenopausal women. J Clin Endocrinol Metab. 2009;94:538-44.

21. Grey A, Bolland M, Wong S, Horne A, Gamble G, Reid IR. Low-dose zoledronate in osteopenic postmenopausal women: a randomized controlled trial. J Clin Endocrinol Metab. 2012;97:286-92.

22. Grey A, Garg S, Dray M, Purvis L, Horne A, Callon K et al. Low-dose fluoride in postmenopausal women: a randomized controlled trial. J Clin Endocrinol Metab. 2013;98:2301-2307.

23. Cole TJ. Too many digits: the presentation of numerical data. Arch Dis Child. 2015;100:608-9.

24. 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.

25. Grey A, Bolland MJ, Avenell A, Klein AA, Gunsalus CK. Check for publication integrity before misconduct. Nature. 2020;577:167-169.

26. Levelt WJ, Drenth PJ, Noort E. Flawed science: The fraudulent research practices of social psychologist Diederik Stapel. 2012:https://www.tilburguniversity.edu/sites/default/files/download/Final%20report%20Flawed%20Science_2.pdf.

27. Bolland MJ, Gamble GD, Grey A, Avenell A. Empirically generated reference proportions for baseline p values from rounded summary statistics. Anaesthesia. 2020;75:1685-1687.

28. 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. 2020;131:22-29.

29. Asemi Z, Hashemi T, Karamali M, Samimi M, Esmaillzadeh A. An Expression of Concern from the AJCN Editorial Office about: Effects of vitamin D supplementation on glucose metabolism, lipid concentrations, inflammation, and oxidative stress in gestational diabetes: a double-blind randomized controlled clinical trial. Am J Clin Nutr. 2020;112:1406.

Appendix

Supplementary References:

List of 33 Clinical studies by Y Sato

1. Sato, Y., Maruoka, H., Oizumi, K., Kikuyama, M. Vitamin D deficiency and osteopenia in the hemiplegic limbs of stroke patients. Stroke. 1996; 27: 2183-2187
2. Sato Y, Honda Y, Kunoh H, Oizumi K (1997) Long-term oral anticoagulation reduces bone mass in patients with previous hemispheric infarction and nonrheumatic atrial fibrillation. Stroke 28: 2390-2394.
3. Sato Y, Kikuyama M, Oizumi K (1997) High prevalence of vitamin D deficiency and reduced bone mass in Parkinson's disease. Neurology 49: 1273-1278.
4. Sato, Y., Kuno, H., Kaji, M., Ohshima, Y., Asoh, T., Oizumi, K. Increased bone resorption during the first year after stroke. Stroke. 1998; 29: 1373-1377
5. Sato, Y., Asoh, T., Kondo, I., Satoh, K. Vitamin D deficiency and risk of hip fractures among disabled elderly stroke patients. Stroke. 2001; 32: 1673-1677
6. Sato, Y., Kuno, H., Asoh, T., Honda, Y., Oizumi, K. Effect of immobilization on vitamin D status and bone mass in chronically hospitalized disabled stroke patients. Age and Ageing. 1999; 28: 265-269
7. Sato, Y., Oizumi, K., Kuno, H., Kaji, M. Effect of immobilization upon renal synthesis of 1,25-dihydroxyvitamin D in disabled elderly stroke patients. Bone. 1999; 24: 271-275
8. Sato, Y., Kaji, M., Tsuru, T., Oizumi, K. Risk factors for hip fracture among elderly patients with Parkinson's disease. Journal of the Neurological Sciences. 2001; 182: 89-93
9. Sato Y, Kaji M, Tsuru T, Satoh K, Kondo I (2002) Vitamin K deficiency and osteopenia in vitamin D-deficient elderly women with Parkinson's disease. Arch Phys Med Rehabil 83: 86-91
10. Sato Y, Kaji M, Honda Y, Hayashida N, Iwamoto J, Kanoko T, et al. (2004) Abnormal calcium homeostasis in disabled stroke patients with low 25-hydroxyvitamin D. Bone 34: 710-715.
11. Sato Y, Kanoko T, Yasuda H, Satoh K, Iwamoto J (2004) Beneficial effect of etidronate therapy in immobilized hip fracture patients. Am J Phys Med Rehabil 83: 298-303
12. Sato Y, Honda Y, Iwamoto J, Kanoko T, Satoh K (2005) Abnormal bone and calcium metabolism in immobilized Parkinson's disease patients. Mov Disord 20: 1598-1603
13. Sato, Y., Kaji, M., Tsuru, T., Oizumi, K. Carpal tunnel syndrome involving unaffected limbs of stroke patients. Stroke. 1999; 30: 414-418
14. Sato, Y., Honda, Y., Asoh, T., Kikuyama, M., Oizumi, K. Hypovitaminosis D and decreased bone mineral density in amyotrophic lateral sclerosis. European Neurology. 1997; 37: 225-229
15. Sato, Y., Kondo, I., Ishida, S., Motooka, H., Takayama, K., Tomita, Y., Maeda, H., Satoh, K. Decreased bone mass and increased bone turnover with valproate therapy in adults with epilepsy. Neurology. 2001; 57: 445-449
16. Sato, Y., Honda, Y., Kuno, H., Oizumi, K. Menatetrenone ameliorates osteopenia in disuse-affected limbs of vitamin D- and K-deficient stroke patients. Bone. 1998; 23: 291-296
17. Sato, Y., Asoh, T., Oizumi, K. High prevalence of vitamin D deficiency and reduced bone mass in elderly women with Alzheimer's disease. Bone. 1998; 23: 555-557
18. Sato, Y., Fujimatsu, Y., Honda, Y., Kunoh, H., Kikuyama, M., Oizumi, K. Accelerated bone remodeling in patients with poststroke hemiplegia. Journal of Stroke and Cerebrovascular Diseases. 1998; 7: 58-62
19. Sato, Y., Kuno, H., Kaji, M., Saruwatari, N., Oizumi, K. Effect of ipriflavone on bone in elderly hemiplegic stroke patients with hypovitaminosis D. American Journal of Physical Medicine and Rehabilitation. 1999; 78: 457-463
20. Sato, Y., Fujimatsu, Y., Kikuyama, M., Kaji, M., Oizumi, K. Influence of immobilization on bone mass and bone metabolism in hemiplegic elderly patients with a long-standing stroke. Journal of the Neurological Sciences. 1998; 156: 205-210
21. Sato, Y., Kuno, H., Kaji, M., Etoh, K., Oizumi, K. Influence of immobilization upon calcium metabolism in the week following hemiplegic stroke. Journal of the Neurological Sciences. 2000; 175: 135-139
22. Sato, Y., Kaji, M., Higuchi, F., Yanagida, I., Oishi, K., Oizumi, K. Changes in bone and calcium metabolism following hip fracture in elderly patients. Osteoporosis International. 2001; 12: 445-449
23. Sato, Y., Asoh, T., Kaji, M., Oizumi, K. Beneficial effect of intermittent cyclical etidronate therapy in hemiplegic patients following an acute stroke. Journal of Bone and Mineral Research. 2000; 15: 2487-2494
24. Sato Y, Honda Y, Kaji M, Asoh T, Hosokawa K, Kondo I, et al. (2002) Amelioration of osteoporosis by menatetrenone in elderly female Parkinson's disease patients with vitamin D deficiency. Bone 31: 114-118

25. Sato, Y., Kuno, H., Kaji, M., Tsuru, T., Saruwatari, N., Oizumi, K. Serum β2-microglobulin reflects increased bone resorption in immobilized stroke patients. American Journal of Physical Medicine and Rehabilitation. 2001; 80: 19-24
26. Sato Y, Honda Y, Iwamoto J, Kanoko T, Satoh K (2005) Homocysteine as a predictive factor for hip fracture in stroke patients. Bone 36: 721-726.
27. Sato Y, Honda Y, Hayashida N, Iwamoto J, Kanoko T, Satoh K (2005) Vitamin K deficiency and osteopenia in elderly women with Alzheimer's disease. Arch Phys Med Rehabil 86: 576-581.
28. Sato Y, Kanoko T, Satoh K, Iwamoto J (2005) Menatetrenone and vitamin D2 with calcium supplements prevent nonvertebral fracture in elderly women with Alzheimer's disease. Bone 36: 61-68
29. Sato Y, Honda Y, Asoh T, Iwamoto J (2006) Longitudinal study of bone and calcium metabolism and fracture incidence in spinocerebellar degeneration. Eur Neurol 56: 155-161.
30. Sato Y, Kaji M, Metoki N, Satoh K, Iwamoto J (2003) Does compensatory hyperparathyroidism predispose to ischemic stroke? Neurology 60: 626-629.
31. Kuno H (1998) Vitamin D status and nonhemiplegic bone mass in patients following stroke. Kurume Med J 45: 257-263.
32. Sato Y, Kanoko T, Satoh K, Iwamoto J (2004) Risk factors for hip fracture among elderly patients with Alzheimer's disease. J Neurol Sci 223: 107-112.
33. Sato Y, Honda Y, Iwamoto J, Kanoko T, Satoh K (2005) Amelioration by mecobalamin of subclinical carpal tunnel syndrome involving unaffected limbs in stroke patients. J Neurol Sci 231: 13-18.

Supplementary Methods:

2.1.1 Identical summary statistics in separate animal RCTs (cases)

We used the reported means and SDs for each variable in the two common treatment arms to calculate the population mean and SD for the whole cohort of 20 animals. We then generated normally distributed random numbers based on the population mean and SD as observations simulating each animal in each RCT. We calculated the mean and SD of the observations for the 10 simulated animals in each simulated RCT and rounded them to the largest number of decimal places presented in the publications for each variable. We repeated this 100,000 times and calculated the probability from these simulations that the means in the two simulated RCTs were the same, that the SDs were the same, that both the means and SDs were the same, and that neither the mean nor the SD was the same.

For this analysis, we used both baseline and outcome variables from the two studies. Usually, the effects of treatment and chance would make outcome variables less similar both to their respective baseline variables and to outcome variables from other trials. In this example, 14/16 variables were outcome variables. Conceptually, there is no difference as to whether the variable is a baseline variable or an outcome variable in the analysis. However, because the outcome variable represents the baseline variable plus the effect of the treatment, it seems inherently less likely that outcome variables would be identical than the baseline variables would be identical.

2.1.2 Identical summary statistics in different animal RCTs from the same control population (controls)

We duplicated the original groups of either 9 or 11 animals from the 2 original studies to form a population of 18 or 22 animals, respectively. We then re-randomised the animals from this population into two treatment groups 1,000,000 times, representing the two different RCTs. For practical computational reasons, this was a simpler process than the analysis described in 2.1.1, and therefore we did a greater number of simulations. We rounded the means and SDs using common rounding (1-2 decimal places) and more extreme rounding (whole numbers or 1 decimal place), and determined the probability of identical mean and/or SD in the summary statistics of the two re-randomised treatment groups.

2.2 Identical summary statistics in different clinical studies from similar populations

2.2.1 Cases

For the analyses of the 10 variables reported in at least 10 cohorts, we assumed that the most common recurring data for a variable (ie the mode) represented the mean and SD for the source population. This assumption provides the most conservative estimate of the likelihood of more than one occurrence of the same data because the mean and SD of the entire source population are the single most likely values to occur when sampling from that population. When there was no unique mode, we used the average mean and SD for each variable from the 34 cohorts weighted by cohort size to estimate the population mean and SD. We then followed the methods described in supplementary section 2.1.1 and generated normally distributed random numbers based on the population mean and SD as observations simulating each participant in each cohort. We calculated the mean and SD of the observations, rounded them to the largest number of decimal places presented in the publications for each variable, repeated this 100,000 times and calculated the probability of obtaining multiple occurrences of identical summary statistics in the different cohorts.

2.3 Identical summary statistics within an RCT and in different cohorts

These analyses are analogous to those in section 2.1.2. In section 2.1.2, the analyses assumed the randomised groups were for different trials, whereas here we assume they are for the same trial. For the Carlisle dataset analyses, we restricted the trials to two-arm RCTs which left a dataset of 3599 RCTs reporting 22,020 baseline variables. Carlisle extracted data from the source publications, and thus the degree of rounding represents authors' choice rather than any prespecified rule. Because of the important impact of rounding on proportions of identical summary statistics, we analysed proportions of identical summary statistics by the degree of rounding. We used significant figures rather than decimal places to account for differences in units (eg a height of 1.52 m is the same as a height of 152 cm, and both measurements have 3 significant figures even though the number of decimal places differs).

There were two datasets with concerns about publication integrity. The first dataset contains 41 RCTs by Sato and colleagues, with multiple retractions for a wide variety of reasons including concerns about data integrity and fabricated data [2]. The second dataset of 172 RCTs by the research group of Z Asemi also has numerous concerns regarding their integrity [27, 28] and several expressions of concerns have been published eg [29].

Supplementary Results:

3.3 Identical summary statistics within an RCT and in different cohorts

We repeated the analyses for identical summary statistics variables in different cohorts in a larger set of 29 commonly reported baseline variables in the 9 Auckland clinical control RCTs, again treating each RCT arm as an individual cohort. A total of 540 variables from the 9 RCTs and 22 cohorts were included in the analysis, and 2000 simulations were run. The probability of the observed number of identical summary statistics was frequently low with 40-50% of p-values <0.1. In addition, the observed distribution of identical means and both identical mean and identical SD were inconsistent with the expected distribution (data not shown).

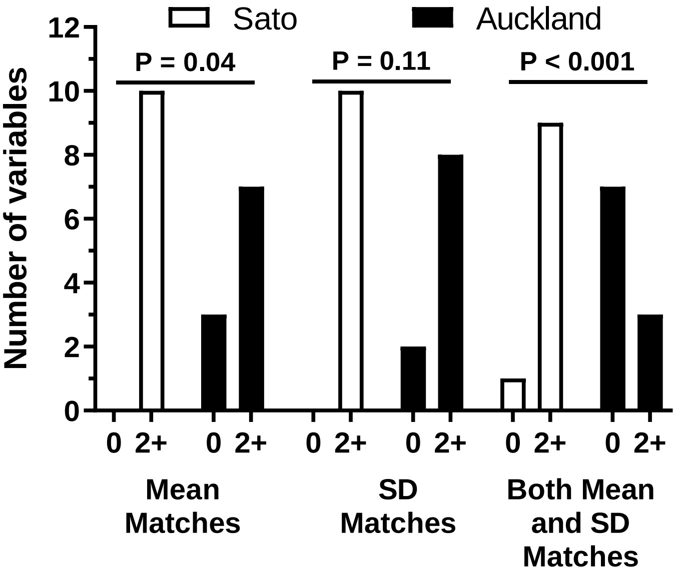
We repeated these analyses in the two sets of RCTs about which concerns have been raised, treating each RCT arm as an individual cohort while restricting the analyses to variables that were in at least 10 cohorts. We also conducted the analyses in the entire groups of RCTs, and then in subgroups of RCTs conducted in similar population groups. As with the previous analysis on identical summary statistics within an individual RCT, we did not find any consistent or sufficiently large differences between the control studies and those with integrity concerns that could be used to support concerns about publication integrity (data not shown).

Supplementary Discussion

A high proportion of identical summary statistics when they are presented with few significant figures is nicely illustrated with data for albumin in Supplementary Table. The mean (43) and SD (2) are presented with 2 and 1 significant figures, respectively, (and both to 0 decimal places) which means that there are a very limited range of possible summary statistics. In the 100,000 simulations only 5 mean values for a simulated cohort occurred (41, 42, 43, 44, 45) and about 89% of the simulations had a mean of 43. In the 22 cohorts, none of the mean values, and only 1 of the SD, and 1 mean and SD combination were unique- all the remaining values recurred in different cohorts. Thus, it is important to consider the degree of rounding when assessing the proportions of identical summary statistics. Assessing summary statistics using raw data would be helpful in this situation.

The size of the SD also affects the proportion of identical summary statistics. Using the example for albumin above, if the SD is increased to 9 in simulations, the range of mean values increases from 41-45 to 34-52, the proportion of simulations with a mean of 43 (the mode) drops from 89% to 34%, the median number of identical matches/simulation drops from 21 to 20, and consequently the probability that all 22 cohorts have an identical mean drops by about 2/3 (from 0.40 to 0.13).

Supplementary Figure: the observed distributions of identical summary statistics for variables in cohorts by Sato and, for comparison, in cohorts from the Auckland clinical control dataset.



Supplementary Table: the proportion of recurring identical summary statistics in 22 cohorts from 9 RCTs from the Auckland clinical control database

Variable	Mean (N)[a]	Mean Match N (%)[a]	Largest Number Matches N (%)[a]	SD (N)[a]	SD Match N (%)[a]	Largest Number Matches N (%)[a]	Mean and SD Match N (%)[a]	Largest Number Matches N (%)[a]	Mean (SD) Mode/ Average[b]	Mode Match N[b]	P Mean Match[c]	P SD Match[c]	P Mean and SD Match[c]
Age	20	0 (0)	0 (0)	20	2 (10)	2 (10)	0 (0)	0 (0)	70.68 (5.21)		0.49	0.63	>0.99
Body mass index	22	12 (55)	2 (9)	22	12 (55)	4 (18)	2 (9)	2 (9)	23.9 (3.6)	2	0.27	0.04	0.64
Spine bone density	22	13 (59)	4 (18)	22	17 (77)	4 (18)	2 (9)	2 (9)	1.07 (0.18)	2	0.01	0.10	0.11
25-hydroxyvitamin D	22	6 (27)	2 (9)	22	8 (36)	3 (14)	0 (0)	0 (0)	22.8 (7.7)		0.25	0.21	0.81
C-telopeptide	14	6 (43)	2 (14)	14	11 (79)	3 (21)	0 (0)	0 (0)	4.4 (1.9)		0.10	0.50	0.40
Parathyroid hormone	10	2 (20)	2 (20)	10	0 (0)	0 (0)	0 (0)	0 (0)	31.1 (12)		0.50	0.37	0.99
Creatinine	22	0 (0)	0 (0)	22	0 (0)	0 (0)	0 (0)	0 (0)	82.89 (12.96)		0.71	0.62	>0.99
Weight	22	4 (18)	2 (9)	22	8 (36)	2 (9)	0 (0)	0 (0)	67 (11.2)		0.27	0.53	0.90
Years since menopause	20	0 (0)	0 (0)	20	8 (40)	2 (10)	0 (0)	0 (0)	21.4 (7)		0.002	0.31	0.81
Albumin	22	22 (100)	9 (41)	22	21 (95)	17 (77)	21 (95)	7 (32)	43 (2)	7	0.40	0.48	0.81

[a] The columns represent the number of reported means and SD; the number of means, SD, and both mean and SD that recurred amongst the different cohorts; and the largest number of matches for each statistic.

[b] where there was no recurring mean (SD) combination, the average mean and SD in the 34 cohorts weighted by numbers of participant are reported, otherwise the most common mean (SD), ie the mode, is reported together with the number of occurrences of the mode.

[c] P refers to the probability of the reported number of identical means, identical SDs or both identical mean and identical SD for each variable or a more extreme number of matches (relative to the most common number of matches) in 100,000 simulations