Journal of Clinical Epidemiology 112 (2019) 67–76

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Baseline *P* value distributions in randomized trials were uniform for continuous but not categorical variables

Mark J. Bolland[a,*], Greg D. Gamble[a], Alison Avenell[b], Andrew Grey[a], Thomas Lumley[c]

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

Accepted 15 May 2019; Published online 21 May 2019

Abstract

Objective: Comparing observed and expected distributions of baseline variables in randomized controlled trials (RCTs) has been used to investigate possible research misconduct, although the validity of this approach has been questioned. We explored this technique and introduced a novel metric to compare *P* values from baseline variables between treatment arms.

Study Design and Setting: We compared observed with expected distributions of baseline *P* values using a one-way chi-square test and by comparing the area under the curve (AUC) of the cumulative distribution function in 13 RCTs conducted by our group, two groups of RCTs known to contain fabricated data, and simulations.

Results: In our 13 RCTs, the distribution of *P* values from baseline continuous variables was consistent with the expected theoretical uniform distribution ($P = 0.19$, difference from expected AUC −0.03, 95% confidence interval [−0.04, 0.04]). For categorical variables, the *P* value distribution was not uniform. The distributions of *P* values from RCTs with fabricated data were highly unusual and not consistent with the uniform distribution for continuous variables, nor with the expected distribution for categorical variables, nor with the distribution of *P* values in genuine RCTs.

Conclusions: Assessing baseline *P* values in groups of RCTs can identify highly unusual distributions that might raise or reinforce concerns about randomization and data integrity.

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

1. Introduction

Because allocation of participants in a randomized controlled trial (RCT) is random, baseline characteristics of the randomized groups, on average, should be balanced. Therefore, the expected distribution of *P* values from comparisons between randomized groups for independent variables at baseline is uniform, with an equal likelihood of a *P* value for any decile, that is, <0.1, 0.1–0.2, 0.2–0.3, and so on. Carlisle [1] introduced a novel technique assessing whether the observed distribution of baseline variables in a group of RCTs is consistent with the expected distribution to investigate cases of research misconduct, and we extended the approach to assessing the distribution of *P* values of comparisons between baseline variables [2]. In both cases, subsequent investigations found that at least some of the RCTs assessed were fabricated [3,4]. This suggests that comparing observed and expected distributions of baseline *P* values from a group of RCTs might be a useful way of identifying highly unusual distributions that are unlikely to have arisen by chance through randomization. The technique is likely to have greater utility when there is increased prior probability of irregularities in a group of trials, rather than for a single RCT. Recently, Carlisle assessed

Conflicts of interests: The authors declare that they have no competing interests.

Funding: No specific funding was received for this study. M.B. receives salary support from the Health Research Council of New Zealand. The Health Services Research Unit is funded by the Chief Scientist Office of the Scottish Government Health and Social Care Directorates. The funders had no role in the study design; collection, analysis, and interpretation of the data; writing of the report; and in the decision to submit the paper for publication.

Data access: The specific SAS code used to perform the analyses can be obtained by contacting the lead author (M.B.) by email.

* Corresponding author. Department of Medicine, Bone and Joint Research Group, Faculty of Medical and Health Sciences, University of Auckland, Private Bag 92 019, Auckland 1142, New Zealand. Tel.: +64 9 3737 599×83004; fax: +64 9 3737 677.

E-mail address: m.bolland@auckland.ac.nz (M.J. Bolland).

https://doi.org/10.1016/j.jclinepi.2019.05.006
0895-4356/

What is new?

Key findings

- *P* values calculated from comparisons of baseline continuous variables in genuine randomized controlled trials (RCTs) were approximately uniformly distributed.
- However, *P* values from comparisons of baseline categorical variables were not uniformly distributed.

What this adds to what was known?

- The distribution of *P* values from two sets of RCTs known to contain fabricated data were highly unusual and not consistent with the expected distributions.
- A novel metric, the area under the curve of the cumulative distribution function (AUC-CDF) for baseline *P* values, provides additional useful information when the observed AUC-CDF is compared with the expected AUC-CDF.

What is the implication and what should change now?

- Large departures from expected distributions, particularly for a moderate to large number of continuous variables in a group of RCTs might raise or reinforce concerns about randomization and data integrity and provide supporting evidence for further investigation

the distribution of baseline *P* values by combining all the *P* values from the comparison of baseline variables in a single RCT into a single *P* value using a variety of techniques [5]. Our approach is different because we did not combine *P* values into a single formal test, instead of measuring the distance from the observed distribution to the expected distribution: analogous to the distinction between performing a *t*-test for comparing means and evaluating the clinical significance of the actual mean difference.

Bland considered the hypothesis that the expected distribution of baseline *P* values is uniform in a series of simulations and concluded that it was valid for independent variables that are normally distributed but may not be valid for other distributions or where all variables are highly correlated or for categorical data analyzed with chi-square or Fisher's exact test [6]. Furthermore, Bland's result implies that, even under randomization, the *P* value distribution will fail a test for being exactly uniform, given enough variables. Here, we explore potential limitations of the assessment of baseline *P* values, using a new metric, area under the curve of the cumulative distribution function (AUC CDF), to quantify the *extent* of deviation from the expected distribution of *P* values between randomized groups and to compare this against a more traditional approach. We applied both techniques to individual patient data from 13 placebo-controlled RCTs carried out by our research group over the past 20 years, summary data from two separate groups of RCTs, which are known to contain at least some fabricated data [1,2], and a series of simulations. In this article, we focus on continuous and categorical variables, and in a companion paper, the issues of normality, correlation, rounding, and methods of randomization [7].

2. Methods

2.1. Control RCTs: individual patient dataset of 13 RCTs by our group

We used anonymized individual patient data (n = 2,851, 726 baseline continuous variables, 192 categorical variables) from 13 single-center, placebo-controlled RCTs [8−20] carried out by our group. We extracted baseline data from the original dataset for each individual trial and combined them into a pooled dataset. All baseline variables from the individual trial datasets were included in the pooled dataset, except for categorical variables with an expected value of fewer than 5 in any cell, to avoid problems from the use of sparse data.

2.1.1. Continuous variables

We compared the means of the continuous baseline variables between the randomized groups for each RCT in the pooled dataset using a *t*-test or one-way ANOVA. All analyses were done using raw, unrounded data. The distribution of *P* values by decile was compared with the expected uniform distribution. To estimate the likely random variation in *P* value distribution, we undertook 100 simulations in which each trial was rerandomized using the original randomization method (Table 1) and compared the baseline variables with a *t*-test or one-way ANOVA for each rerandomization.

When the number of variables analyzed become large, it is possible that even small differences from the expected proportions might become statistically significant. Therefore, as an additional analysis, we calculated the AUC of the CDF of the baseline *P* values and compared the AUC with that of the uniform distribution. In contrast to *P* values, which tend to become smaller as sample size increases (unless the null hypothesis is precisely true), these CDF-based comparisons are estimates and do not systematically increase or decrease with sample size. The estimated CDF, like an estimated mean, is unbiased at any sample size. Although the meta-analytic combination of *P* values assumes exact independence, the estimation of the CDF (like estimation of a mean) assumes only that each variable examined provides some incremental information; that is, that collinearity is not close to perfect [21]. The AUC gives a directional summary of departures from a uniform

Table 1. Design and baseline characteristics and variables in 13 randomized controlled trials in the individual patient dataset

Study	*N*	Mean age (y)	Population	Agent	Randomization method[a]	Continuous variables (*N*)	Categorical variables (*N*)
Reid 1993 [8]	135	58	Older women	Calcium	Variable blocks	104	10
Reid 2000 [9]	185	63	Older women	HCTz	Stratification (2 variables) block size 4	82	9
Reid 2005 [10]	41	63	Older women	Propranolol	Variable blocks	32	3
Reid 2006 [11]	1,471	74	Older women	Calcium	Minimization (3 variables)	68	45
Bolland 2007 [12]	43	49	HIV-infected men	Zoledronate	Variable blocks	54	9
Grey 2007 [13]	50	67	Older women	Rosiglitazone	Variable blocks	53	4
Reid 2007 [14]	80	65	Women, osteoporosis	Fluoride	Variable blocks	45	7
Reid 2008 [15]	323	56	Older men	Calcium	Variable blocks[b]	63	11
Grey 2009 [16]	50	64	Women, osteopenia	Zoledronate	Variable blocks	41	14
Grey 2012 [17]	180	65	Women, osteopenia	Zoledronate	Variable blocks[c]	44	23
Bolland 2013 [18]	27	57	Sarcoidosis	Vitamin D	Variable blocks	51	7
Grey 2013 [19]	180	69	Women, osteopenia	Fluoride	Variable blocks[c]	46	22
Grey 2014 [20]	86	64	Diabetes	Pioglitazone	Variable blocks	43	28

Abbreviation: HCTz, hydrochlorothiazide.
[a]All studies were placebo-controlled trials with two arms except one three-arm study[b] and two 4-arm studies[c].

distribution: it is 0.5 for a uniform distribution, >0.5 when *P* values tend to be smaller than expected, and <0.5 when the *P* values tend to be larger than expected.

2.1.2. Categorical variables

Bland showed that there are only a discrete, limited number of *P* values possible when comparing categorical variables with a prevalence of 0.5 between randomized groups where the sample size is small [6]. Therefore, in this situation, the distribution of *P* values will not be uniform. Likewise, in large fixed sample sizes, the distribution of *P* values was also not uniform with clear peaks and troughs. Similar results were also obtained for large samples of varying size [6].

One simulation not assessed by Bland was whether changing the prevalence of the categorical variables impacts on the distribution of *P* values. We therefore undertook a simulation of two treatment groups for 1,000 categorical variables each with a randomly selected prevalence between 0.15 and 0.85 (to avoid sparse data), using different sample sizes, both fixed and varying. For each variable, the number of cases in each treatment group was randomly generated using the binomial distribution and the randomly selected prevalence. Each variable was then compared between the two groups using a chi-square test, and the distribution of *P* values by decile was compared to the uniform distribution, and the AUC of the CDF of *P* values was compared with that of a uniform distribution (i.e., 0.5).

Next, we compared baseline categorical variables between the randomized groups for each RCT in the pooled dataset using a chi-square test with exact mid-*P* statistics. The distribution of *P* values by decile was compared with the uniform distribution, and the AUC of the baseline *P* value CDF was compared with that of the uniform distribution. To estimate the likely random variation in *P* value distribution, we used the 100 rerandomizations for each trial and compared the baseline variables using a chi-square test with exact mid-*P* statistics. As the distribution of *P* values was clearly nonuniform, we used the dataset of 100 rerandomizations to generate the expected proportion of *P* values by decile and the expected AUC of the CDF and repeated the analyses.

2.2. Summary datasets of trials known to contain fabricated data

We randomly selected 15 RCTs (Appendix Table A1) from the 168 RCTs reported by Fujii and analyzed by Carlisle [1] and extracted the baseline data. One RCT had two treatment arms, and the remaining 14 trials had three, four, or five arms. *P* values were not reported for any trial; therefore, we calculated these from summary data using *t*-tests or one-way ANOVA for continuous data and using a chi-square test with exact mid-*P* statistics for categorical data (Open Epi Version 2.3.1, www.OpenEpi.com). All data were analyzed using the number of significant figures reported in the text. Categorical variables where the expected value in any cell was fewer than 5 were excluded. In separate analyses, we used the extracted baseline data and *P* values from 25 RCT reports with Sato as the first author (Appendix Table A2) from our previous analysis [2]. The mode of fabrication in these RCTs by Sato has not been publicly reported.

For continuous variables, the distribution of *P* values by decile for both groups of RCTs was compared with the uniform distribution and also to the distribution of *P* values in the control dataset of RCTs, and the AUC of the *P* value CDF was compared with that of the uniform distribution. For categorical variables, the expected distribution of *P* values was generated from 10,000 simulations in which the number of cases in each treatment group for each variable in each trial was randomly generated using the binomial distribution and the prevalence for the variable in the trial. For categorical variables with more than two levels, the levels with the smallest two proportions were pooled, thereby creating a two-level variable. Each variable was then compared between the two groups using a chi-square test with exact mid-*P* statistics to generate the expected *P* value distribution. Finally, the observed distribution of *P* values was compared with the expected distribution and the observed AUC of the *P* value CDF to the expected AUC.

2.3. Statistical analyses

The distributions of *P* values grouped by decile were compared with the expected uniform distribution using a one-way chi-square test with Monte Carlo estimates of exact *P* when there was an expected value of <5 for a decile. The AUC for the CDF of *P* values was calculated using the trapezoidal method. In addition, 95% confidence intervals (CIs) for the AUC were calculated using bootstrap resampling ($n = 500$, sampling with replacement) for the Sato and Fujii trials and from the 2.5 and 97.5 centiles of the AUCs of the CDF from the dataset of 100 rerandomizations for the individual patient dataset. To compare the distribution of *P* values in the summary datasets by Sato and Fujii with the distribution in the control RCTs, we used bootstrap resampling with the replacement of the baseline *P* values from both datasets, performed two-sample Kolmogorov−Smirnov tests on these values, and repeated this 1,000 times. We also compared the AUCs of the *P* value CDFs between the different groups of trials. All analyses on the individual patient and summary datasets were performed with SAS (version 9.4; SAS Institute, Cary, NC) and for the simulations with Excel 2010 or SAS.

3. Results

3.1. Control RCTs

Table 1 shows selected characteristics and design features of the 13 RCTs. All trials were carried out in older people under the broad theme of osteoporosis treatment and prevention. Four were carried out in specific conditions (HIV, sarcoidosis, diabetes, and osteoporosis), three in healthy women with osteopenia, and the other six in healthy individuals. The mean age of participants ranged from 49 to 74 years, and the number of participants per trial ranged from 27 to 1,471. The pooled dataset contained 222 continuous baseline variables, ranging in each trial from 32 to 104 (total 726 variables), and 68 categorical baseline variables, ranging in each trial from 3 to 45.

3.1.1. Continuous variables

We assessed the distribution of *P* values from the comparison of baseline continuous variables between the randomized groups for each RCT. For all 726 baseline variables, Fig. 1A shows that the distribution of *P* values was approximately uniformly distributed ($P = 0.19$), although the proportion of *P* values <0.10 (6.8%) appears less than expected. To estimate the likely random variation in *P* value distribution, we rerandomized each trial 100 times using the original randomization method (Table 1) and compared the 726 baseline continuous values for each rerandomization. Fig. 1B shows that the distribution of *P* values in these 100 rerandomizations was uniform, with greater variation in the decile of *P* values <0.10 compared with other deciles. Fig. 1C shows that the AUC of the CDF of the 726 baseline *P* values was 0.47, a difference of −0.03 (95% CI −0.04 to 0.04) from the uniform distribution AUC of 0.50, indicating *P* values tended to be larger than expected.

3.1.2. Categorical variables

First, we assessed in simulations whether changing the prevalence of the categorical variables impacts on the distribution of *P* values. The distribution of *P* values for 1,000 variables with a randomly chosen fixed prevalence between 0.15 and 0.85 and a fixed treatment group size of 20 produced a discrete, nonuniform distribution, similar to the distribution reported by Bland for a fixed prevalence. At treatment group sizes larger than approximately 250−300; however, the distribution became more uniform both visually and with fewer one-way chi-square $P < 0.05$. For example, in 92 of 100 repeated simulations with a treatment group size of 300, the one-way chi-square *P* was >0.05, and the mean difference in AUC from the uniform distribution AUC was 0.00 (95% CI −0.02 to 0.01).

Next, we simulated the effect of both varying treatment group sizes and varying prevalence, thereby modeling a ''real-world'' set of trials. We simulated 50 trials each with 20 baseline variables, where each variable had a randomly chosen prevalence between 0.15 and 0.85, and 80% of treatment group sizes were randomly chosen between 20 and 100, and 20% were randomly chosen between 100 and 300. This simulation models a set of small- to moderate-sized clinical trials. The visual distribution of the *P* values from these simulations was approximately uniform with only modest departures from the uniform distribution proportions. For example, in 83 of 100 repeated simulations, the one-way chi-square *P* was >0.05, and the mean difference in AUC from the uniform distribution AUC was 0.00 (95% CI −0.01 to 0.02).

These results suggest that departures from the uniform distribution are small for *P* values distributions of baseline

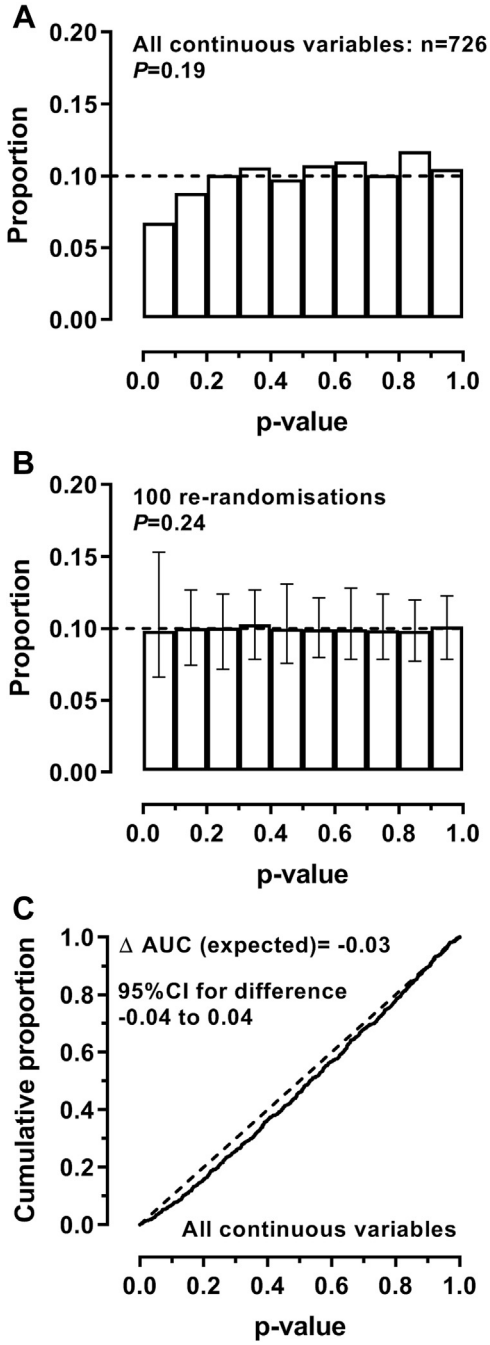

Fig. 1. Distribution of baseline *P* values from continuous variables in 13 randomized controlled trials by our group. Panel A shows the distribution of *P* values by decile for all 726 variables, and Panel B shows the distribution of *P* values by decile with 95% CIs from 100 rerandomizations of the trial data (n = 726 variables, 100 randomizations, thus 72,600 *P* values). The dotted line is the expected proportion of 0.10 in Panels A and B. Panel C shows the cumulative distribution function (CDF) of the baseline 726 *P* values (solid line) with the CDF of the expected uniform distribution (dotted line). ΔAUC is the difference in the area under the curve (AUC) of the CDF from the AUC of the expected uniform distribution CDF, with the confidence intervals (CIs) determined from the AUCs of the CDFs from the 100 rerandomizations. The CDF of the 100 rerandomizations is visually indistinguishable from the plotted CDF of the uniform distribution (data not shown, available from authors on request).

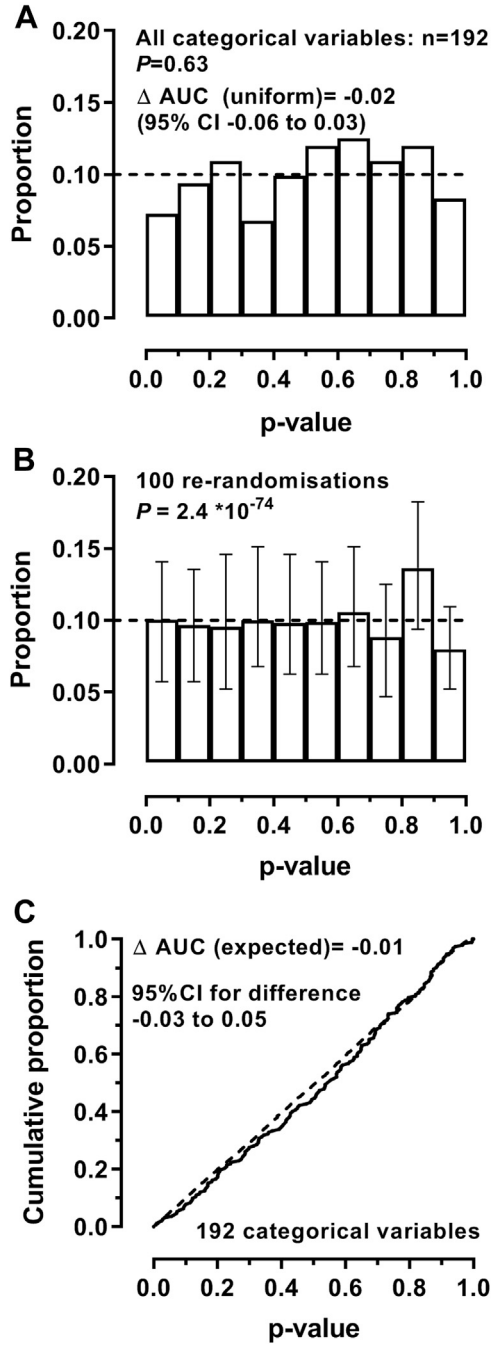

Fig. 2. Distribution of baseline *P* values from categorical variables in 13 randomized controlled trials by our group. Panel A shows the distribution of *P* values by decile for all 192 variables compared with the uniform distribution (dotted line). ΔAUC (uniform) is the difference in area under the curve (AUC) of the cumulative distribution function (CDF) from the AUC of the uniform distribution CDF, with the confidence intervals (CIs) determined from the AUCs of the CDFs from the 100 rerandomizations. Panel B shows the distribution of *P* values with 95% confidence intervals by decile from 100 rerandomizations of the trial data (n = 192 variables, 100 randomizations, thus 19,200 *P* values) compared with the uniform distribution (dotted line). Panel C shows the CDF of the baseline 192 *P* values (solid line) with the CDF of the expected distribution calculated from the 100 rerandomizations (dotted line). ΔAUC (expected) is the difference in CDF AUC from the expected CDF AUC, with the CIs determined from the AUCs of the CDFs from the 100 rerandomizations.

categorical variables with a range of prevalence in moderate to large studies (e.g., treatment group size >300) or from a series of studies of small to moderate size.

We therefore assessed the distribution of *P* values for the baseline categorical variables from our individual patient dataset. There were 192 categorical variables with an expected frequency of 5 or more per cell in the 13 RCTs. The observed prevalence for variables with two outcomes ranged from 0.02 to 0.50, with 15%, 16%, 24%, 25%, and 20% having observed prevalence <0.10, 0.10–0.20, 0.20–0.30, 0.30–0.40, and 0.40–0.50, respectively. Fig. 2A shows that for all 192 variables, the distribution of *P* values could be consistent with a uniform distribution ($P = 0.63$, difference from uniform distribution AUC −0.02), although as for the continuous variables, the proportions in some individual deciles appeared smaller or larger than expected. However, Fig. 2B shows that when each trial was rerandomized 100 times, the resulting distribution of *P* values from these simulations is nonuniform. We therefore used the proportions for each decile of *P* values from the 100 rerandomizations as the expected distribution, but this did not substantially alter the one-way chi-square test result ($P = 0.66$) from the preceding analysis in Fig. 2A. Fig. 2C shows that AUC of the CDF of the 192 baseline *P* values was 0.48, a difference of −0.01 (95% CI −0.03, 0.05) from the expected AUC calculated from the 100 rerandomizations. Taken together, these results suggest that it cannot be assumed that the uniform distribution will be the expected distribution for baseline categorical *P* values.

3.2. Summary datasets of trials known to contain fabricated data

3.2.1. RCTs by Fujii et al.

Fig. 3A shows that the distribution of *P* values for the 90 baseline continuous variables in the 15 randomly selected RCTs was not consistent with a uniform distribution ($P = 2 \times 10^{-18}$), with 76% of *P* values >0.7. Fig. 3B shows that the difference in the CDF AUC from the uniform distribution AUC was −0.27 (95% CI −0.31 to −0.22), indicating *P* values were larger than expected. There were only 13 baseline categorical variables with an expected value of at least 5 in any cell. Fig. 4A shows that the distribution of *P* values from these variables was not consistent with the expected distribution calculated from 10,000 simulations ($P < 0.0001$), and Fig. 4B shows that the difference in CDF AUC from the expected AUC was −0.37 (CI: −0.52 to −0.16).

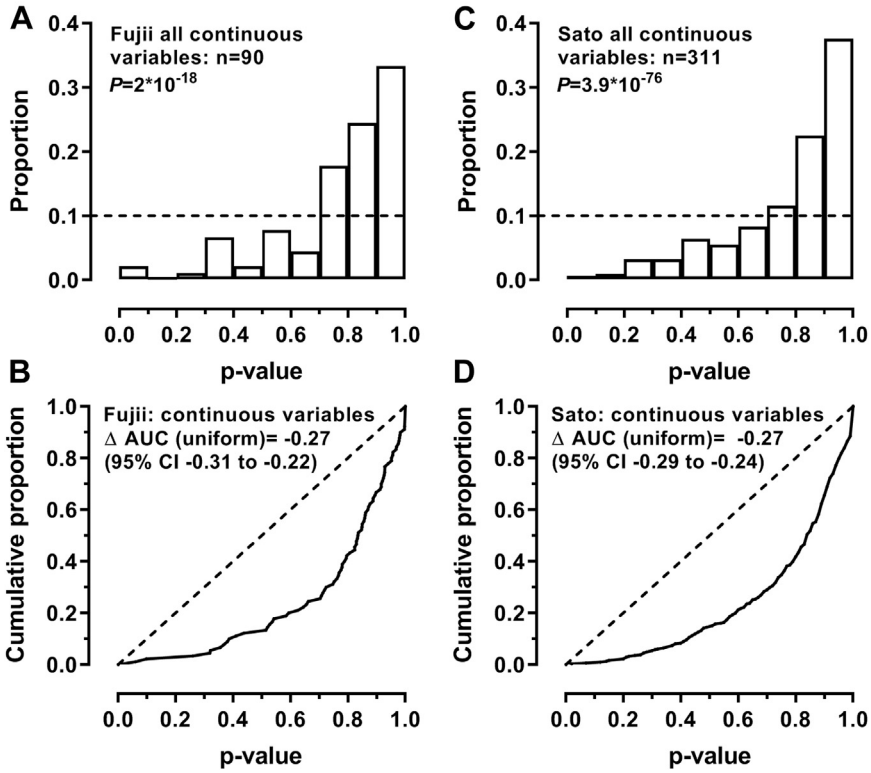

Fig. 3. Distribution of continuous baseline *P* values in randomized controlled trials by Fujii and Sato. Panels A and C show the distribution of *P* values by decile for all continuous variables compared with the expected uniform distribution (dotted line). Panels B and D show the cumulative distribution function (CDF) of the baseline *P* values (solid line) with CDF of the expected uniform distribution (dotted line). ΔAUC (uniform) is the difference in area under the curve (AUC) of the CDF from the AUC of the expected uniform distribution CDF, with the confidence intervals (CIs) determined using bootstrap resampling.

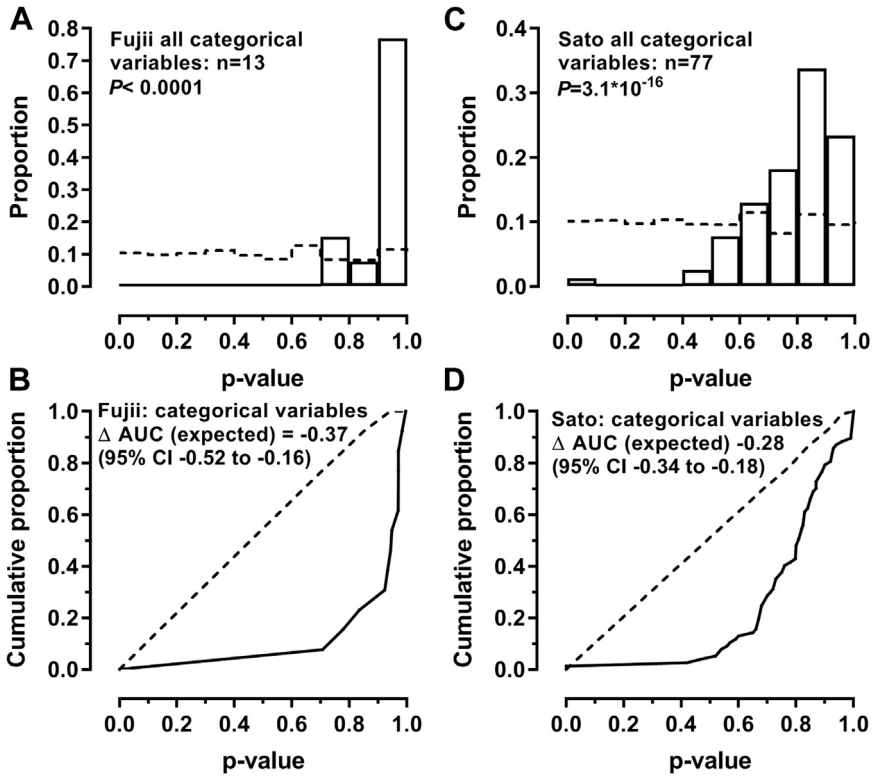

Fig. 4. Distribution of categorical baseline *P* values in randomized controlled trials by Fujii and Sato. Panels A and C show the distribution of *P* values by decile for all categorical variables compared with the expected distribution (dotted line). Panels B and D show the cumulative distribution function (CDF) of the baseline *P* values (solid line) with CDF of the expected distribution (dotted line). ΔAUC (expected) is the difference in area under the curve (AUC) of the CDF from the AUC of the expected distribution CDF. The expected proportions and distributions were generated from 10,000 simulations of the reported trial data, and the 95% confidence intervals (CIs) were calculated using bootstrap resampling.

3.2.2. RCTs by Sato et al.

Fig. 3C shows that the distribution of *P* values for 311 baseline continuous variables was not consistent with a uniform distribution ($P = 3.9 \times 10^{-76}$), with 72% of *P* values >0.7, and Fig. 3D shows that the difference in CDF AUC from the uniform distribution AUC was −0.27 (CI: −0.29 to −0.24), indicating that *P* values were larger than expected. There were 77 baseline categorical variables with an expected value of 5 or more in all cells. Fig. 4C shows that the distribution of *P* values for these variables was not consistent with the expected distribution calculated from 10,000 simulations ($P = 3.1 \times 10^{-16}$), and Fig. 4D shows that the difference in CDF AUC from the expected AUC was −0.28 (CI: −0.34 to −0.18).

3.2.3. Comparison to control RCTs

We compared the distribution of baseline *P* values for continuous variables from the RCTs by Fujii et al. and the RCTs by Sato et al. with the distribution of the *P* values for continuous variables in the control RCTs using two-sample Kolmogorov–Smirnov tests on 1,000 datasets of baseline-values created with bootstrap resampling with replacement. To more closely match the presentation of baseline variables between the RCTs by Fujii and Sato and the control RCTs, we restricted the analyses on the control RCTs to 30 commonly presented variables in tables of baseline characteristics in RCTs in the populations we studied (Box 1). There were marked differences in distributions of *P* values between the control RCTs and the RCTs by both Fujii and Sato ($P < 0.0001$ in all 1,000 comparisons for each group of trials). In contrast, the distributions of *P* values from the Fujii and Sato RCTs were similar: in only 1 in 1,000 replicates, the two-sample Kolmogorov–Smirnov *P* was <0.0001, and in 10.6%, the *P* was <0.05.

4. Discussion

The results show that in 13 genuine RCTs of small to moderate size, the distribution of *P* values for baseline continuous variables was consistent with the expected uniform distribution. However, for categorical variables, although the distribution of *P* values was broadly consistent with a uniform distribution in some simulations, in our trials, the distribution was not uniform. This was not obvious in the simple analysis of the baseline variables possibly because there were insufficient categorical variables or RCTs in these analyses, but it was clearly apparent when

Box 1 Thirty commonly presented variables in baseline characteristics tables in 13 randomized controlled trials by our group

Clinical characteristics	Age
	Age at menopause
	Height
	Weight
Full blood count	Hemoglobin
	White blood cell count
Basic biochemistry	Albumin
	Creatinine
	Glucose
	Potassium
	Sodium
Liver function	Alkaline phosphatase
	Aspartate transaminase
	Bilirubin
	Gamma-glutamyl transferase
Serum calcium and bone parameters	Calcium
	Phosphate
	25-Hydroxyvitamin D
	1,25-Dihydroxyvitamin D
	β-C-terminal telopeptide of type I collagen
	Procollagen type-I N-terminal propeptide
	Parathyroid hormone
	Urine calcium
	Dietary calcium intake
Dual-energy X-ray absorptiometry	Lumbar spine
	Total hip
	Femoral neck
	Total body
	Lean mass
	Fat mass

the dataset of 100 rerandomizations of the RCTs was analyzed. In contrast, analyses of two different sets of RCTs known to contain at least some fabricated data showed distributions of baseline values that were markedly different from the uniform distribution for continuous variables and from the expected distribution for categorical variables. In both sets of trials, a much higher proportion of *P* values >0.7 was observed ($>70\%$) than was expected (30%), suggesting that the distribution of *P* values is highly unlikely to have arisen by chance through the process of randomization.

The distribution of *P* values in the two summary datasets known to contain at least some fabricated data differed markedly both from the expected distribution and from the distribution of *P* values from the control RCTs. This empiric evidence suggests that major departures from the expected distributions, as characterized by the differences in AUC of the CDF from the expected AUC, warrant explanation and investigation, particularly if they occur in many studies from the same group of authors. In such situations, provision of the individual patient datasets for confirmation of the reported findings by independent statistical analysis would allow greater confidence in the integrity of the reported results. If the authors are unwilling or unable to do this, it seems reasonable that the unusual distribution of results is flagged to readers as a potential concern regarding possible failure of randomization. Similar analyses of trials known to be fabricated would be helpful in extending our results.

An important related question is when such analyses should be applied. In the absence of automated software to perform these analyses, the extraction of data from multiple RCTs and their analysis is labor intensive and time consuming. Therefore, analyses are likely only to be undertaken when other concerns have arisen about papers. The analyses are more powerful when used for a group of RCTs because there are more *P* values from baseline variables to analyze. Continuous variables are usually reported more frequently than categorical variables, and the limitations in the analysis of categorical data suggest that initial analyses should focus on continuous variables. Generally, there will be insufficient data in a single RCT to draw firm conclusions, but a distribution of *P* values that differs markedly from the expected uniform distribution might flag a concern. For example, in two studies by Sato et al., 21/23 and 10/10 baseline *P* values were >0.8 respectively [2] (references 12,13, Appendix Table A2). We think that focusing on the visual distribution of the *P* values and the descriptive statistics of AUC CDF is more helpful than the one-way chi-square *P*, which may be influenced by the number of variables or between-variable correlation. Both small departures from the expected value for one or more deciles and larger departures from the expected value for a single decile only should be very cautiously interpreted. However, large departures from the expected values, especially in multiple deciles, reflected in a smaller-than-expected AUC, appears to be uncommon in groups of genuine RCTs.

There are limitations to this technique and scope for further research. The analyses are best performed on a group of RCTs, for which there are sufficient data reported on baseline continuous variables. Although baseline continuous variables are summarized in most RCT reports, future requirements to provide greater access to individual patient data may enable more comprehensive comparisons to be made. The technique appeared valid in RCTs carried out by our group, but independent

confirmation in other datasets is important. In analyses of single (or few) RCTs with only a small number of baseline values, the results should be viewed very cautiously. Likewise, if categorical variables are analyzed, the limitations discussed previously by Bland [6] and identified in these analyses need to be considered and again results should be viewed cautiously. The technique identified substantial departures from the expected in two moderately large sets of trials known to contain fabricated data, but whether it would successfully identify less extreme examples of fabricated data is not known.

An important area for further research is the statistical test and the distance measure to use for comparing the distribution of baseline *P* values with the expected uniform distribution. Throughout the article, we used the AUC and the one-way chi-square test with Monte Carlo estimates of exact *P* when the expected value of all cells was fewer than 5 for the distributions of *P* values grouped by decile. The chi-square test is valid for independent variables, but a small proportion of baseline variables are at least moderately correlated. The impact of both sparse data and correlation of variables on the results of the chi-square test, as well as other statistical approaches to the analysis, such as other goodness-of-fit tests (such as the one-sample Kolmogorov−Smirnov test or the Anderson-Darling test) are worth exploring. The maximum-difference metric underlying the Kolmogorov-Smirnov test is expected to be more sensitive to discrete data than the AUC, but other averaged metrics such as that underlying the Anderson-Darling test should be considered. *P* values calculated using different software packages can vary, but any differences are usually minor. We have assessed the impact of non-normality, strength of correlation, rounding, and methods of randomization on *P* value distribution in a companion paper [7].

5. Conclusions

Comparing the distribution of *P* values of baseline continuous variables in RCTs with the expected uniform distribution seems to be a useful technique for identifying highly unusual distributions that might raise or support concerns about data integrity and prompt further investigation. Previous concern about the validity of this approach may have been exaggerated, in part, by assuming greater collinearity between baseline variables that is seen in an empirical real-world collection of RCT data. The technique is labor intensive and seems best suited for the analysis of data from a moderate to large number of continuous variables from a body of RCTs where possible RCT irregularities have already been raised. The results should be interpreted cautiously, but large departures from the expected distributions warrant further investigation of the integrity of the data.

CRediT authorship contribution statement

Mark J. Bolland: Conceptualization, Data curation, Formal analysis, Methodology, Writing - original draft. **Greg D. Gamble:** Conceptualization, Formal analysis, Methodology, Writing - review & editing. **Alison Avenell:** Conceptualization, Methodology, Writing - review & editing. **Andrew Grey:** Conceptualization, Methodology, Writing - review & editing. **Thomas Lumley:** Conceptualization, Methodology, Writing - review & editing.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.05.006.

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;87:2391−402.
[3] Yentis SM. Lies, damn lies, and statistics. Anaesthesia 2012;67:455−6.
[4] Gross RA. Statistics and the detection of scientific misconduct. Neurology 2016;87:2388.
[5] 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.
[6] Bland M. Do baseline P-values follow a uniform distribution in randomised trials? PLoS One 2013;8:e76010.
[7] Bolland MJ, Gamble GD, Avenell A, Grey A. Rounding, but not randomisation method, non-normality, or correlation, affected baseline p-value distributions in randomised trials. J Clin Epidemiol 2019;110:50−62.
[8] 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.
[9] 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;109:362−70.
[10] 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.
[11] 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.
[12] Bolland MJ, Grey AB, Horne AM, Briggs SE, Thomas MG, Ellis-Pegler RB, et al. Annual zoledronate increases bone density in highly active antiretroviral therapy-treated human immunodeficiency virus-infected men: a randomized controlled trial. J Clin Endocrinol Metab 2007;92:1283−8.
[13] 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.
[14] 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.

[15] Reid IR, Ames R, Mason B, Reid HE, Bacon CJ, Bolland MJ, et al. Randomized controlled trial of calcium supplementation in healthy, nonosteoporotic, older men. Arch Intern Med 2008;168:2276−82.

[16] 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.

[17] 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.

[18] Bolland MJ, Wilsher ML, Grey A, Horne AM, Fenwick S, Gamble GD, et al. Randomised controlled trial of vitamin D supplementation in sarcoidosis. BMJ Open 2013;3:e003562.

[19] 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;98:2301−7.

[20] Grey A, Bolland M, Fenwick S, Horne A, Gamble G, Drury PL, et al. The skeletal effects of pioglitazone in type 2 diabetes or impaired glucose tolerance: a randomized controlled trial. Eur J Endocrinol 2014;170:257−64.

[21] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100:9440−5.