ORIGINAL ARTICLE

Rounding, but not randomization method, non-normality, or correlation, affected baseline *P*-value distributions in randomized trials

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

[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, Scotland*

Abstract

Objectives: To investigate whether comparing observed with expected *P*-value distributions for baseline continuous variables in randomized controlled trials (RCTs) might be limited by randomization methods, normality and correlation of variables, or calculation of *P*-values from rounded summary statistics.

Study Design and Setting: We assessed how each factor affects differences from expected for *P*-value distributions and area under the curve of the cumulative distribution function (AUC-CDF) of baseline *P*-values in 13 RCTs and in simulations.

Results: The *P*-value distributions and AUC-CDF for variables with possible non-normal distribution and in simulations using eight different randomization methods were consistent with the theoretical uniform distribution and AUC-CDF, respectively, although stratification and minimization produced smaller-than-expected proportions of *P*-values <0.10. Seventy-seven percentage of 3,813 pairwise correlations between baseline variables in the 13 individual RCTs were between −0.2 and 0.2. *P*-value distribution and AUC-CDF remained consistent with the uniform distribution in simulations with incrementally increasing correlation strength. The *P*-value distributions calculated from rounded summary statistics were not uniform, but expected distributions could be empirically generated.

Conclusions: Randomization methods, non-normality, and strength of correlation of baseline variables did not have important effects on baseline *P*-value distribution or AUC-CDF, but baseline *P*-values calculated from rounded summary statistics are non-uniformly distributed. © 2019 Elsevier Inc. All rights reserved.

Keywords: Statistical methods; Research integrity; *P*-values; Correlation; Randomisation; Rounding

* Corresponding author. Department of Medicine, Bone and Joint Research Group, Faculty of Medical and Health Sciences, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand. Tel: +64 9 3737 599x83004; fax: +64 9 3737 677.

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

1. Introduction

The distribution of baseline variables in groups of RCTs has been used in assessment of research integrity and potential research misconduct, when there have been prior concerns about the research [1−4]. Previously, we assessed the distribution of *P*-values from comparisons between baseline variables in a group of RCTs about which concerns had been raised [3]. To date, at least some of these RCTs were determined to be fraudulent [5]. In theory, because participants in an RCT are randomly allocated to study groups, the expected distribution of *P*-values from comparisons between randomized groups for independent baseline continuous variables (baseline *P*-values) is uniform. In an accompanying paper, we assessed the distribution of baseline *P*-values in a data set of individual patient data from 13 placebo-controlled RCTs carried out by our research group over the past 20 years [6]. In this data set, the average distribution of baseline *P*-values from comparisons of continuous variables was uniform, with only small differences

<table>
<tr><td>

What is new?

Key findings

- Non-normal distribution of baseline continuous variables, eight common randomization methods, and correlation of baseline variables did not have important effects on baseline *P*-value distribution.

- However, the distribution of *P*-values calculated from rounded summary statistics is not uniform, although the expected distribution can be empirically generated.

What this adds to what was known?

- Concerns that correlation and non-normality of baseline variables or randomization methods would impact on baseline *P*-value distribution in genuine randomized controlled trials (RCTs) do not appear to be justified.

What is the implication and what should change now?

- Distribution of baseline *P*-values calculated from rounded summary statistics should be compared to empirically generated distributions not the uniform distribution.

</td></tr>
</table>

from the expected distributions. However, it has been suggested that baseline *P*-values may not be uniform when techniques other than simple randomization are used [7]. In addition, concerns have been raised that non-normal or fixed and/or highly correlated baseline variables may lead to non-uniformly distributed baseline *P*-values [7–9]. Here, we extend our previous analyses by exploring the impact of different randomization methods, non-normal distribution of variables and increasingly strong correlation between variables on the distribution of baseline *P*-values, and the area under the curve (AUC) of the cumulative distribution function (CDF) of these *P*-values from continuous variables in the data set of our trials and in simulations.

Many groups and journals recommend against reporting of *P*-values from between-group comparisons of baseline variables in articles, even though it remains a common practice [10]. When baseline *P*-values are not reported, they can be calculated from the reported summary statistics (mean, standard deviation, *n*) using parametric tests. As these summary statistics are invariably rounded, the calculated *P*-values are likely to differ from those calculated from raw data. We also explored what effect rounding has on the distribution of baseline *P*-values and the area under the curve of the cumulative distribution function (AUC-CDF).

2. Methods

2.1. Data sets of RCTs

We pooled anonymized individual patient data from 13 single-centre, placebo-controlled RCTs [11–23] carried out by our group, as previously described [6]. Table 1 shows selected features of the 13 RCTs. All trials were carried out in older people and studied 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. Randomization was by a minimization algorithm for 1 RCT, using stratification for 1 RCT, and using variable block sizes for 11 RCTs. For these analyses, the pooled data set was restricted to 30 commonly presented baseline continuous variables (Table 2) to represent the typical "real world" presentation of baseline data. The number of baseline variables in each trial ranged from 18 to 28, and the entire data set contained 319 variables from the 13 RCTs.

2.2. Calculation of baseline P-values

We compared the means of the baseline variables between randomized groups with a t-test or one-way ANOVA using individual, raw, unrounded data for each RCT in the primary analyses. We repeated these comparisons using nonparametric Wilcoxon or Kruskal-Wallis tests. The distribution of baseline *P*-values by decile was compared to the expected uniform distribution using a one-way chi-square test. We also calculated the AUC of the CDF of the baseline *P*-values and compared the AUC to that of the uniform distribution (0.50) [6]. 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.

2.3. Effects of randomization methods

We generated a data set with 100 simulated randomizations of the 13 RCTs using eight different randomization methods. Separate randomizations were carried out for each simulation and each RCT. First, a uniformly distributed random number was generated in each simulation for each participant in each RCT. Then the different methods of randomization were used to form simulated treatment groups. (1) Simple randomization: groups were formed based on appropriate thresholds (0.5 for two-arm studies; 0.333 and 0.667 for three-arm studies; and 0.25, 0.5, and 0.75 for four arm studies). (2) Randomization in one block per group: groups were formed based on the median, tertile, or quartile of random numbers. (3) Fixed block size: block sizes of 4 for two-arm and four-arm studies and 3 for three-arm studies were used. (4) Variable block size: block sizes between 4 and 20 or 3 and 18 were used. (5) Stratified one block:

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

Study	N	Mean age (y)	Population	Design	Agent	Baseline variables (N)	Randomization method
Reid 1993 [11]	135	58	Older women	2-arm	Calcium	22	Variable blocks
Reid 2000 [12]	185	63	Older women	2-arm	Hydrochlorothiazide	26	Stratification (2 variables) block size 4
Reid 2005 [13]	41	63	Older women	2-arm	Propranolol	18	Variable blocks
Reid 2006 [14]	1,471	74	Older women	2-arm	Calcium	19	Minimization (3 variables)
Bolland 2007 [15]	43	49	HIV-infected men	2-arm	Zoledronate	27	Variable blocks
Grey 2007 [16]	50	67	Older women	2-arm	Rosiglitazone	28	Variable blocks
Reid 2007 [17]	80	65	Women, osteoporosis	2-arm	Fluoride	26	Variable blocks
Reid 2008 [18]	323	56	Older men	3-arm	Calcium	25	Variable blocks
Grey 2009 [19]	50	64	Women, osteopenia	2-arm	Zoledronate	26	Variable blocks
Grey 2012 [20]	180	65	Women, osteopenia	4-arm	Zoledronate	25	Variable blocks
Bolland 2013 [21]	27	57	Sarcoidosis	2-arm	Vitamin D	28	Variable blocks
Grey 2013 [22]	180	69	Women, osteopenia	4-arm	Fluoride	25	Variable blocks
Grey 2014 [23]	86	64	Diabetes	2-arm	Pioglitazone	24	Variable blocks

participants were stratified into 8 groups by the median value for age, weight, and lumbar spine bone density, and then groups were formed based on the median, tertile, or quartile of random numbers for each stratum. (6) Stratified fixed block: block sizes of 4 for two-arm and four-arm studies and 3 for three-arm studies were used for each stratum. (7) Minimization: treatments were assigned using a minimization algorithm that aimed to balance treatment groups for age, weight, and lumbar spine bone density (using median value for each study as threshold). (8) Weighted minimization: there was an 80% chance of treatment being assigned using minimization and a 20% chance of treatment being assigned using simple randomization. The P-values for age, weight, and lumbar spine bone density were not included in analyses of the distribution of baseline P-values for the stratified or minimization analyses.

2.4. Effect of normality of distribution

We assessed whether the baseline continuous variables in the pooled original data set were normally distributed using the Shapiro-Wilk test and identified any variables with $P < 0.05$ and $P < 0.001$, respectively, which suggested a non-normal distribution. We then restricted the analyses of baseline P-values by decile to these variables with possible non-normal distribution.

We generated a data set of 100 simulations of normally distributed baseline variables (the "simulated normal data set") in which each simulated observation for an individual was generated using a normally distributed random number based on the mean and standard deviation for each variable from each of the 13 RCTs, and individuals were randomized in a single block per treatment group to ensure group numbers were similar. Analyses were repeated in this data set, in which there were few variables that were highly non-normally distributed.

2.5. Effect of correlation of baseline variables

Spearman correlations were calculated for baseline variables in the RCTs and Pearson correlations for the simulated normal data set.

To assess the effect of increasing strength of correlation of variables on the distribution of baseline P-values, we generated a data set of 100 simulations of five normally distributed variables (age, height, weight, lumbar spine bone density, and serum creatinine) based on the mean, standard deviation, and covariance matrix of the variable in each of the 13 RCTs using the IML procedure in SAS. We then increased each pairwise correlation between variables away from 0 by 0.1, converted the correlation matrix to a covariance matrix, and repeated the simulation. If correlation matrices were invalid, the nearest valid correlation matrix was estimated and used [24,25]. We repeated these analyses using all baseline variables from each study.

We also assessed the impact of clustering of correlated variables on baseline P-values. By chance, there may be a large difference in a variable between randomized groups. If this variable is highly correlated with other variables, it might be expected that those variables may also differ between groups. The converse argument would apply for closely matched variables. We therefore restricted our analyses to simulations with $P < 0.10$ or $P > 0.90$ for age from the comparison of randomized groups in the simulated data sets with increasing correlation between five variables based on the correlation matrices of the RCTs. We then assessed the distribution of baseline P-values for the other four variables.

2.6. P-values calculated from rounded summary statistics

To assess the effect of rounding of summary statistics on the distribution of P-values, we calculated the mean and

Table 2. Thirty variables commonly presented in baseline trial data

Category	Variable	Mean (SD)	Common rounding	Extreme rounding
Clinical characteristics	Age (y)	68.3 (9.6)	0.1 (0.1)	1 (1)
	Age at menopause (y)	49 (5)	1 (1)	1 (1)
	Height (cm)	162.3 (8.3)	0.1 (0.1)	1 (1)
	Weight (kg)	69.5 (13)	0.1 (0.1)	1 (1)
Full blood count	Hemoglobin (g/L)	136 (10)	1 (1)	1 (1)
	White blood cell count (cells/L)	5.7 (1.6)	0.1 (0.1)	1 (1)
Basic biochemistry	Albumin (g/L)	43 (2.6)	1 (0.1)	1 (1)
	Creatinine (umol/L)	84 (15)	1 (1)	1 (1)
	Glucose (mmol/L)	5.1 (0.7)	0.1 (0.1)	0.1 (0.1)
	Potassium (mmol/L)	4.4 (0.4)	0.1 (0.1)	0.1 (0.1)
	Sodium (mmol/L)	141 (2.2)	1 (0.1)	1 (1)
Liver function	Alkaline phosphatase (U/L)	80 (22)	1 (1)	1 (1)
	Aspartate transaminase (U/L)	23 (6.4)	1 (0.1)	1 (1)
	Bilirubin (umol/L)	12 (6.1)	1 (0.1)	1 (1)
	Gamma-glutamyl transferase (U/L)	23 (18)	1 (1)	1 (1)
Serum calcium and bone parameters	Calcium (mmol/L)	2.33 (0.09)	0.01 (0.01)	0.01 (0.01)
	Phosphate (mmol/L)	1.15 (0.15)	0.01 (0.01)	0.1 (0.1)
	25-hydroxyvitamin D (nmol/L)	63 (26)	1 (1)	1 (1)
	1,25 dihydroxyvitamin D (pmol/L)	105 (32)	1 (1)	1 (1)
	β-C-terminal telopeptide of type I collagen (ug/L)	0.41 (0.20)	0.01 (0.01)	0.01 (0.01)
	Procollagen type-I N-terminal propeptide (ug/L)	48 (20)	1 (1)	1 (1)
	Parathyroid hormone (pmol/L)	3.6 (1.5)	0.1 (0.1)	0.1 (0.1)
	Urine calcium (mmol/L)	2.27 (1.91)	0.01 (0.01)	0.1 (0.1)
	Dietary calcium intake (mg/d)	879 (430)	1 (1)	1 (1)
Dual energy X-ray absorptiometry	Lumbar spine (g/cm^2)	1.09 (0.19)	0.01 (0.01)	0.01 (0.01)
	Total hip (g/cm^2)	0.91 (0.15)	0.01 (0.01)	0.01 (0.01)
	Femoral neck (g/cm^2)	0.86 (0.14)	0.01 (0.01)	0.01 (0.01)
	Total body (g/cm^2)	1.08 (0.12)	0.01 (0.01)	0.01 (0.01)
	Lean mass (kg)	39.6 (9.3)	0.1 (0.1)	1 (1)
	Fat mass (kg)	25.7 (9.5)	0.1 (0.1)	1 (1)

The mean and standard deviation (SD) are the summary data from all 13 randomized controlled trials. The rounding columns show two different levels used to round summary statistics from which baseline *P*-values were calculated.

standard deviation for each variable for each of the 13 RCTs, rounded these summary data, and calculated the *P*-values from them using a t-test or one way ANOVA. We used two levels of rounding: firstly, typical rounding that might be presented in a article, and secondly an extreme level of rounding (Table 2). We performed these analyses in the data set of RCTs with 100 simulated randomizations (with individuals randomized in one block per treatment group to ensure group numbers were similar) and in the simulated normal data set.

When rounded summary means are identical, the *P*-value calculated from summary statistics is 1. To determine whether simulating *P*-values might overcome this issue, we performed 1,000 simulations for each variable in the data set of 100 simulated randomizations (with individuals

randomized in one block per treatment group) and in the simulated normal data set. One thousand simulated means and standard deviations for each rounded mean and standard deviation for each variable in each treatment group for each of the 13 RCTs were calculated using uniformly distributed random numbers that lay within the minimum and maximum rounding of the variable. For example, for a mean of 30, 1,000 values uniformly distributed between 29.5 and 30.5 were generated. Likewise, for a standard deviation of 0.15, 1,000 values between 0.145 and 0.155 were generated. The *P*-value for the difference between groups for each simulation was then calculated from the unrounded simulated means and standard deviations, and the mean of the 1,000 *P*-values from the simulations used in place of the *P*-value calculated from the summary statistics.

2.7. Analyses

All analyses were performed with SAS (SAS Institute, Cary, NC, USA version 9.4). The distributions of *P*-values grouped by decile were compared to the expected uniform distribution using a one-way chi-square test. The AUC for the CDF of *P*-values was calculated using the trapezoidal method. Ninety-five percent confidence intervals (CI) for the AUC for the 319 baseline *P*-values were calculated from a data set of 100 rerandomizations using the original trial randomization method [6]; for other analyses, they were calculated from the 2.5 to 97.5 centiles of the AUCs of the CDF from analyses involving multiple simulations, or from bootstrap resampling (*n* = 500, sampling with replacement) for analyses without multiple simulations.

3. Results

3.1. Effect of randomization method

The distribution of *P*-values from the comparison of the 319 baseline variables between the randomized groups in the 13 placebo-controlled RCTs was approximately consistent with a uniform distribution (*P* = 0.39, difference in AUC from the uniform distribution AUC −0.03, [95% CI −0.04, 0.04], Fig. 1), although some proportions for individual deciles differed from the expected proportions. Fig. 1 also shows that in the data set of 100 rerandomizations using the original randomization method (Table 1) the distribution of *P*-values was approximately uniform.

Fig. 2 shows the results of 100 simulated randomizations using eight different methods. For simple randomization and randomization in blocks, the distribution of baseline *P*-values was approximately uniform (Figs. 2A–D). When stratification or minimization was used (Figs. 2E–H), visually there appeared to be a smaller-than-expected proportion of *P*-values in the lowest decile, consistent with the pattern seen in Fig. 1, although in all cases the calculated 95% confidence interval (0.05–0.12 stratified fixed blocks; 0.06–0.13 stratified one block; 0.05–0.11 minimization; 0.05–0.12 weighted minimization) included the expected value of 0.10. Fig. 2 and Appendix Fig. A1 show that the AUC-CDF was consistent with the uniform AUC for all randomization methods.

3.2. Effect of normality of distribution

Of the 319 baseline variables in 13 placebo-controlled RCTs, 212 (66%) had *P* < 0.05 and 135 (42%) had *P* < 0.001 from the Shapiro-Wilk test, indicating possible non-normal distribution. When the baseline variables were compared using nonparametric Wilcoxon or Kruskal-Wallis tests (Fig. 3A, Appendix Fig. A2A), the distribution of *P*-values and the AUC of the CDF was similar to the distribution of *P*-values and AUC of the CDF from the parametric tests (Fig. 1). In the analyses of *P*-values from both the
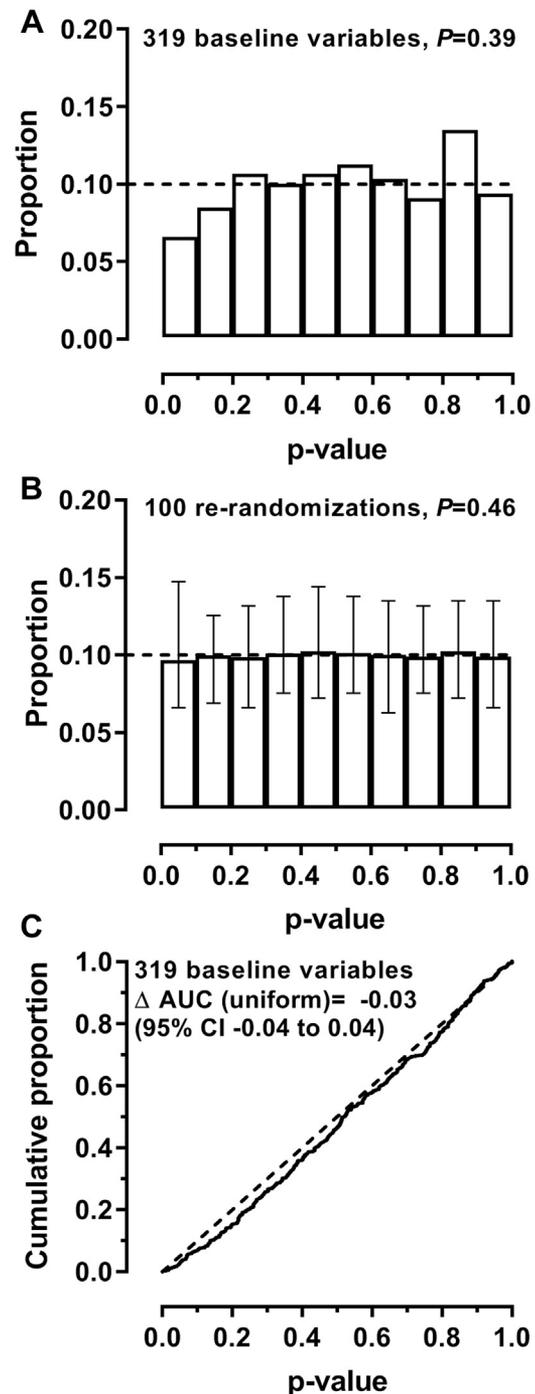

Fig. 1. Distribution of 319 *P*-values for 30 baseline variables in 13 randomized controlled trials. Panel A shows the distribution of *P*-values by decile for all 319 variables, and Panel B the distribution of *P*-values by decile with 95% confidence intervals from 100 rerandomizations of the original trial data (*n* = 319 variables, 100 randomizations, thus 31,900 *P*-values). The dotted line is the expected uniform proportion of 0.10 in Panels A and B. Panel C shows the cumulative distribution function (CDF) of the baseline 319 *P*-values (solid line) with the CDF of the expected uniform distribution (dotted line). Δ AUC (uniform) is the difference in area under the curve (AUC) of the CDF from the AUC of the uniform distribution CDF, with the confidence intervals (CI) determined from the AUCs of the CDFs from the data set of 100 original trial rerandomizations.
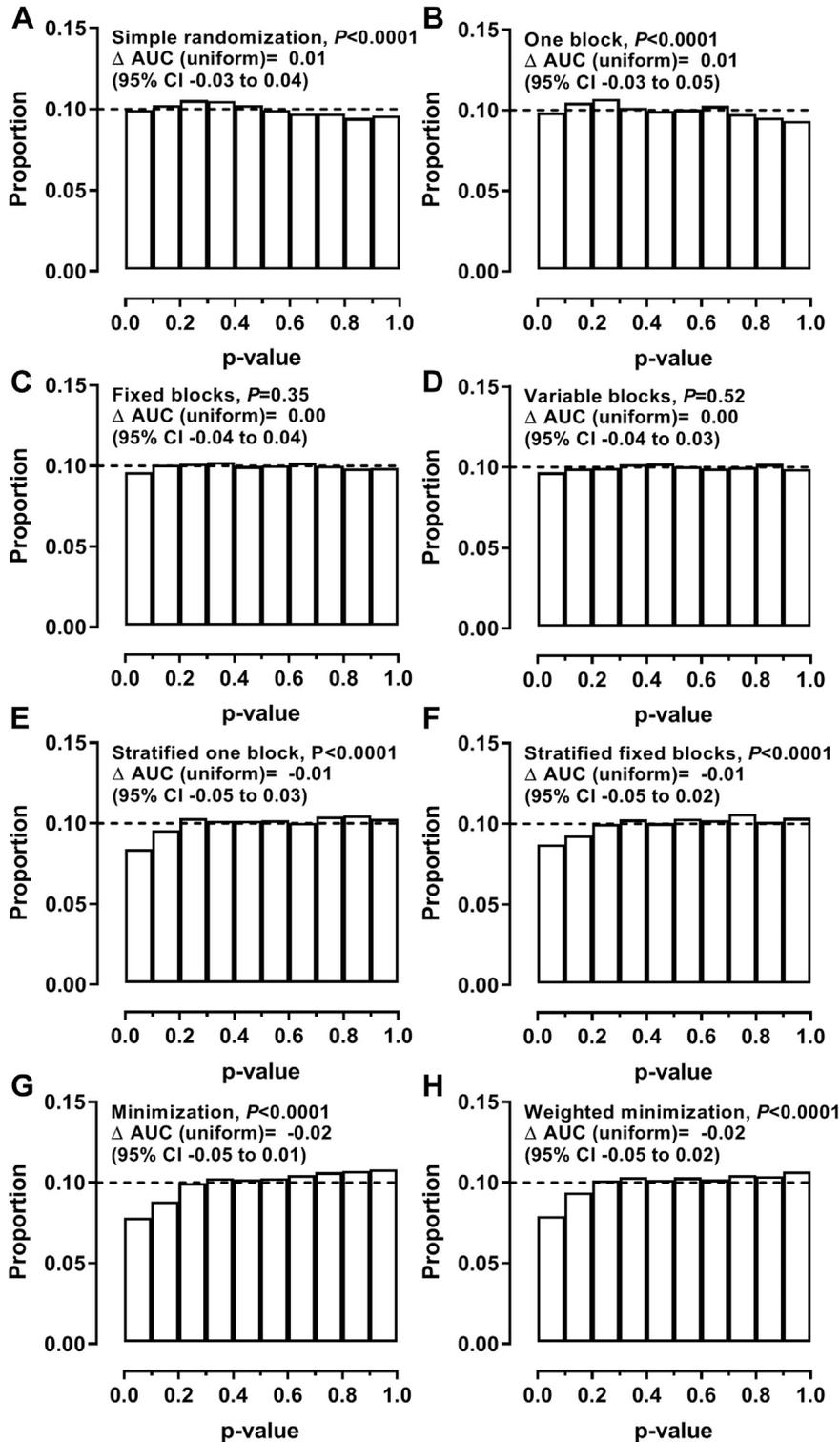
Fig. 2. Effect of randomization method on baseline *P*-value distribution. Distribution of 319 *P*-values for all 30 variables from all 13 randomized controlled trials in 100 simulated randomizations using eight different methods. Panel A simple randomization. Panel B one block per treatment group. Panel C permuted randomization using fixed blocks (4 for 2-arm or 4-arm trials, 3 for 3-arm trials). Panel D variable blocks between 4 and 20 for 2-arm or 4-arm trials and 3 and 18 for 3-arm trials. Panel E stratified by median age, weight, and lumbar spine bone density with one group per stratum. Panel F stratified with fixed block sizes of 4 or 3. Panel G minimization algorithm for age, weight, and lumbar spine bone density. Panel H weighted minimization (20% chance of simple randomization, 80% chance of minimization algorithm). The dotted line is the expected uniform proportion of 0.10. Δ AUC (uniform) is the difference in area under the curve (AUC) of the cumulative distribution function (CDF) from the AUC of the uniform distribution CDF.
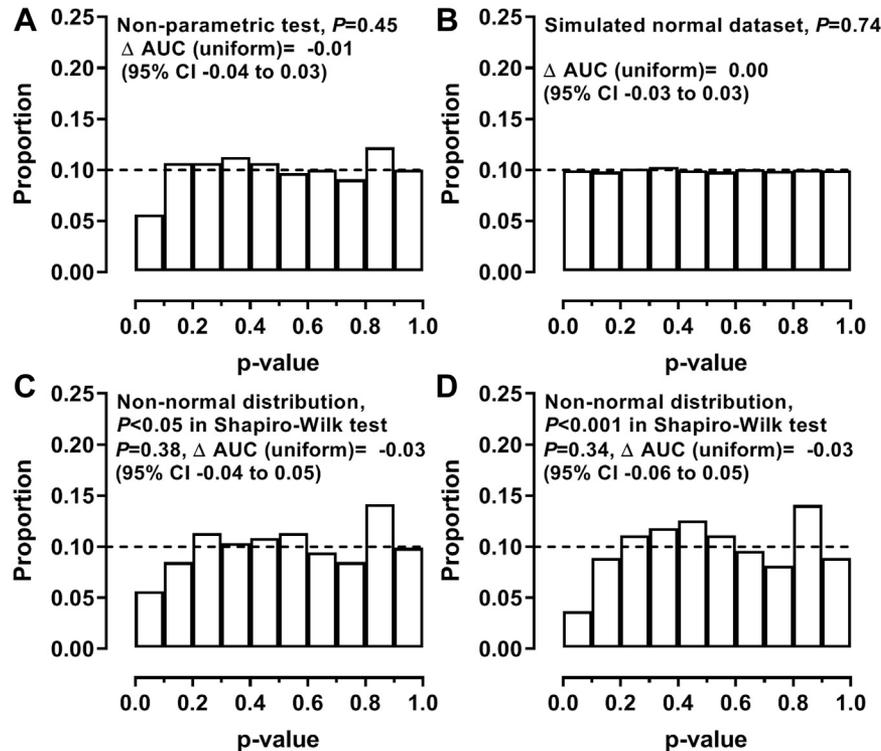
Fig. 3. Effect of normality of distribution on baseline *P*-value distribution. Panel A shows the 319 *P*-values by decile for all 30 variables from all 13 randomized controlled trials (RCTs) using nonparametric tests. Panel B, 100 simulations of normally distributed variables based on the mean and standard deviation from each of the 13 RCTs for all 319 baseline variables. Panels C and D, results from baseline variables from the 13 RCTs (Fig. 1) with $P < 0.05$ (Panel C, $n = 212$) or with $P < 0.001$ (Panel D, $n = 135$) from the Shapiro-Wilk test. The dotted line is the expected uniform proportion of 0.10. Δ AUC (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 (CI) determined from the AUCs of the CDFs from a data set of 100 rerandomizations using the original trial randomization method (Fig. 1) in Panels A, C, and D, and from the raw data in Panel B.

parametric and nonparametric tests, there appeared to be a smaller-than-expected proportion of *P*-values in the lowest (<0.10) decile (95% confidence interval 0.039−0.093 for parametric *P*-values, 0.034−0.086 for nonparametric *P*-values). We then generated a data set of 100 simulations of normally distributed variables (the "simulated normal data set") based on the means and standard deviations for each variable from each RCT. Fig. 3B and Appendix Fig. A2B show that the distribution of *P*-values from the comparison of baseline variables and the AUC-CDF in the simulated normal data set is consistent with the uniform distribution.

Next, we restricted the analyses from both the pooled data set of RCTs to the non-normally distributed variables with $P < 0.05$ or $P < 0.001$ from the Shapiro-Wilk test of normality. Figures 3C−D and Appendix Figures A2C−D show that the distribution of *P*-values and AUC-CDFs in these restricted analyses is similar to the results for all 319 baseline variables.

3.3. Effect of correlation of baseline variables

We determined the correlation between baseline variables in each individual RCT. Of the 3,813 pairwise

correlations in the individual RCTs, Fig. 4 shows that in 49% the correlation statistic was between −0.1 and 0.1; in 77% the correlation was between −0.2 and 0.2; and that the distribution was skewed with a higher proportion of correlation statistics >0.2 (16%) than $< −0.2$ (7%). In the simulated normal data set, Fig. 4 shows that there were fewer moderately or highly correlated variables, the distribution of the correlations was symmetrical, in 66% the correlation statistic was between −0.1 and 0.1, and in 90% the correlation was between −0.2 and 0.2.

Figures 1 and 3B show that the distribution of *P*-values from the comparison of baseline variables is approximately uniform, both in the simulated normal data set with few moderate or strongly correlated baseline variables and in the RCTs which had a higher proportion of more correlated variables.

In 100 simulations of five variables (age, height, weight, lumbar spine bone density, and serum creatinine) based on the mean, standard deviation, and covariance matrix of the variable in each of the 13 RCTs, increasing correlation between variables had little effect on the distribution of baseline *P*-values. Fig. 5 and Appendix Figures A3 and A4 show the distribution of correlations and AUC-CDF for each simulated level of correlation and that, despite the
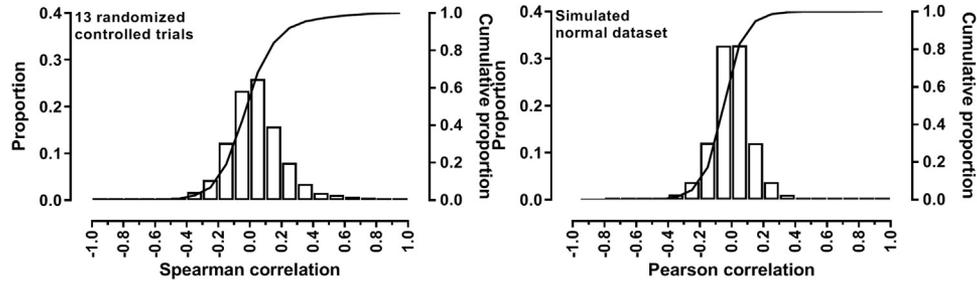
Fig. 4. Distribution of pairwise correlations between baseline values in 13 randomized controlled trials (RCTs) and in the simulated normal data set. The left panel shows the results from the 13 RCTs and the right panel from the simulated normal data set in which 100 normally distributed variables were simulated for each baseline variable based on the mean and standard deviation for each variable from each RCT. The bars show the proportion and the line the cumulative proportion.

increases in correlation, the distribution of baseline *P*-values is uniform, and the AUC of the CDF is consistent with the uniform AUC. Appendix B shows the means, standard deviations, and correlation matrices from the 100 simulations for each increment of correlation. These analyses were repeated using all baseline variables from each study. As larger constants were added to each correlation, there was an increasing number of RCTs for which valid correlation matrices were unable to be produced. Appendix Fig. A5 shows that as the proportion of moderate or highly correlated variables increases, the distribution of *P*-values again remains uniform. Appendix C shows the means, standard deviations, and correlation matrices from the 100 simulations for each increment of correlation.

To assess the impact of clustering of correlated variables on baseline *P*-values, we repeated the analyses restricted to simulations with $P < 0.10$ or $P > 0.90$ for age from the comparison of randomized groups in the simulated data sets of five increasingly correlated variables. For the simulations based on the actual correlation matrices of the RCTs, the distribution of *P*-values for the other four variables was approximately uniform, and the AUC of the CDF was consistent with the uniform AUC (Fig. 6A, Appendix Fig. A6A). However, for the simulations with $P < 0.10$ for age, increasing the correlations leads to a substantial increase of *P*-values <0.2 and a rapid increase in the difference in AUC from the uniform AUC (Figs. 6B−6D left panel, Appendix Figs. A6B−F, left panel). For the simulations with $P > 0.90$ for age, increasing the correlations only leads to a clear nonuniform distribution and change in the AUC-CDF when the increased correlation was large (Figs. 6B−D, Appendix Figs. A6B−F, right panel).

3.4. Baseline P-values calculated from rounded summary statistics

In the data set of RCTs with 100 simulated randomizations, the distribution of baseline *P*-values and AUC of the CDF of these *P*-values calculated from rounded summary statistics is not consistent with the uniform distribution or AUC, with larger than expected proportions of *P*-values >0.9 and <0.1, and smaller-than-expected proportions

between 0.5 and 0.9 (Fig. 7A, Appendix Fig. A7). These effects were more pronounced when extreme rounding was used (Fig. 7B) and also in RCTs with two arms compared to those with three or four arms (Figs. 7C−F).

When rounded summary means are identical, the *P*-value calculated from summary statistics is 1. This situation explains a large proportion of the excess *P*-values >0.9 seen in Figures 7A−F. To determine whether simulating *P*-values might produce a more uniform distribution, we used the mean of the 1,000 *P*-values calculated from 1,000 simulated means and standard deviations for each variable in each treatment group for each of the 13 RCTs. Figures 7G and H shows that the distribution of simulated *P*-values is not uniform with a smaller-than-expected proportion of *P*-values >0.9.

We repeated all these analyses using rounded summary statistics from the simulated normal data set. Appendix Fig. A8 shows that the results from these analyses are very similar to those from the analyses of the data set of RCTs with 100 simulated randomizations.

4. Discussion

These results show that any differences in the distribution of *P*-values from the comparison of baseline continuous variables from a group of 13 genuine RCTs from the expected uniform distribution are small and are not substantially affected by the randomization method, the normality of baseline variables, or the degree of correlation between variables. Even when there is a high proportion of non-normally distributed variables or moderate or strongly correlated variables, the distribution of baseline *P*-values remains approximately uniform, and the AUC of the CDF remains consistent with the uniform AUC. Stratified randomization and minimization algorithms may lead to a smaller-than-expected proportion of *P*-values <0.10, but the effect is only small. In contrast to these minor effects, calculation of *P*-values from rounded summary statistics has important effects on the distribution of baseline *P*-values. When all *P*-values are calculated in this way, the distribution of baseline *P*-values is no longer uniform, with
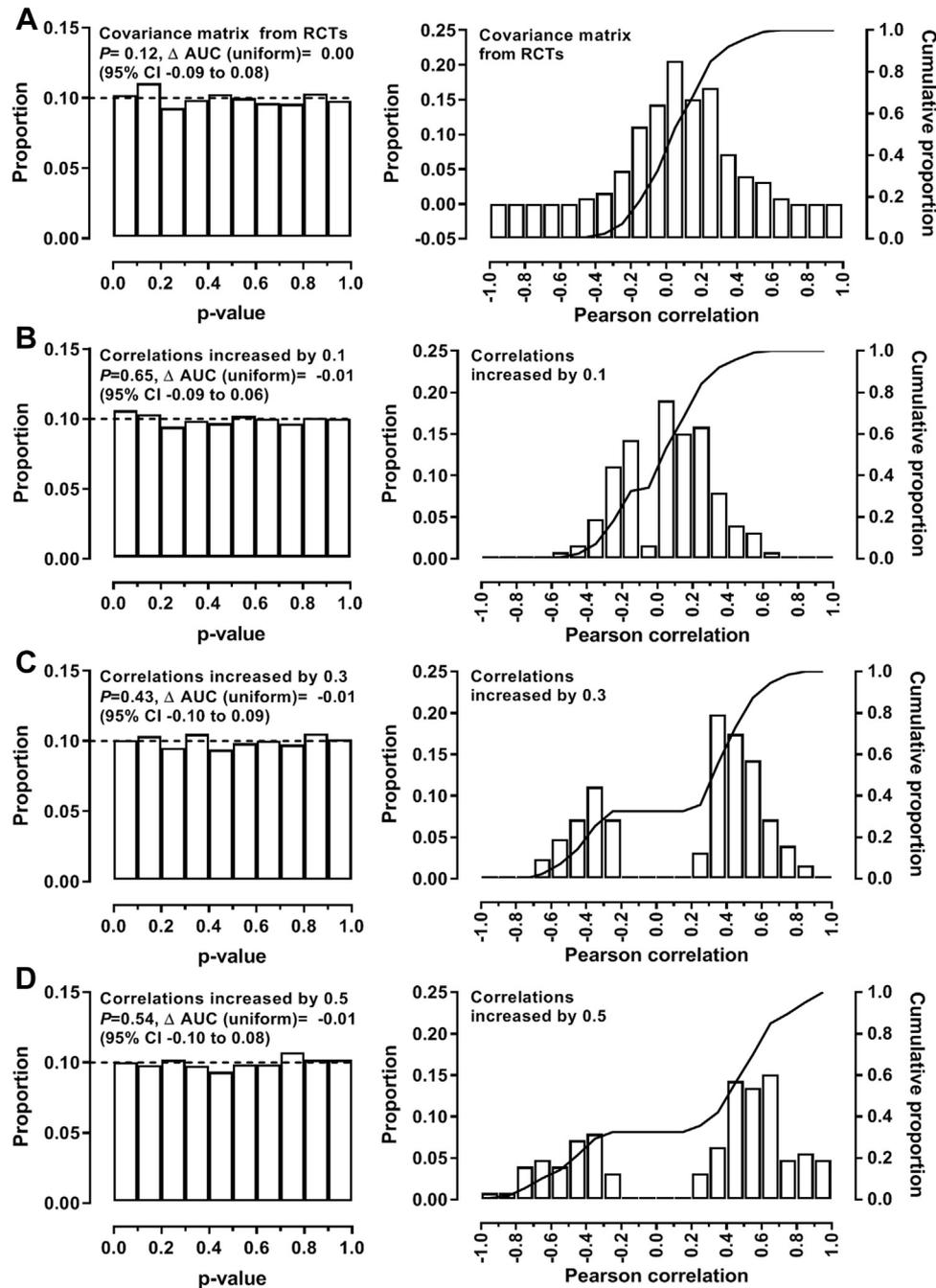
Fig. 5. The effect of increasing correlation of baseline variables on the distribution of *P*-values. The left panels show the distribution of *P*-values by decile and the right panels the distribution of correlation statistics in 100 simulations of five variables (age, height, weight, lumbar spine bone density, and serum creatinine) based on the mean, standard deviation, and covariance matrix from each randomized controlled trial. Panel A uses the covariance matrix from each trial, whereas in Panels B-D, the pairwise correlations are increased by 0.1, 0.3, and 0.5, respectively. The bars show the proportion, the line the cumulative proportion, and the dotted line the uniform distribution proportion of 0.10. Δ AUC (uniform) is the difference in area under the curve (AUC) of the cumulative distribution function (CDF) from the AUC of the uniform distribution CDF.

a large increase in the proportion of *P*-values >0.9, a small increase in *P*-values <0.1, and a small decrease in *P*-values between 0.5 and 0.9 compared to the uniform distribution. The differences from the uniform distribution are greater in two arm RCTs than three or four arm RCTs and greater when rounding is extreme.

The distribution of baseline *P*-values from simulations of variables with highly positively skewed lognormal distributions and simulations in which all variables had fixed or high levels of correlation was not uniform [8,9]. However, these are simulations of extreme situations that are unlikely to be seen in all variables across a group of properly
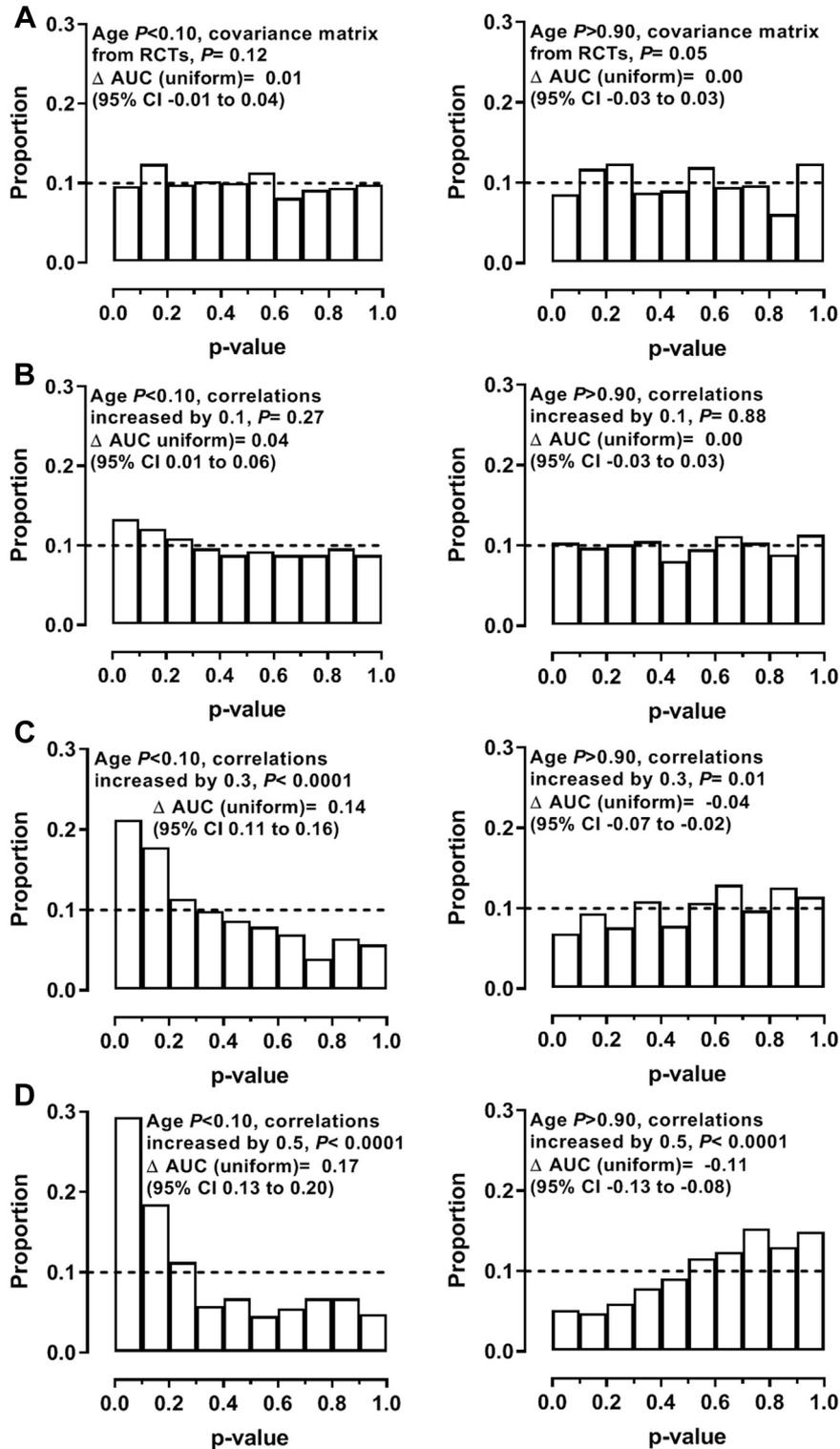
Fig. 6. The effect of clustering and increasing correlation of baseline variables on the distribution of *P*-values. The panels show the distribution of *P*-values by decile in 100 simulations for four variables (height, weight, lumbar spine bone density, and serum creatinine) based on the mean, standard deviation, and covariance matrix from each randomized controlled trial (RCT). The left panels are restricted to simulations with *P* < 0.10 for the between-groups comparison for age; the right panels are restricted to simulations with *P* > 0.90 for age. Panel A uses the covariance matrix from each RCT, whereas in Panels B-D, the pairwise correlations are increased by 0.1, 0.3, and 0.5, respectively. The bars show the proportion and the dotted line the uniform distribution proportion of 0.10. Δ AUC (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 (CI) calculated using bootstrap resampling.
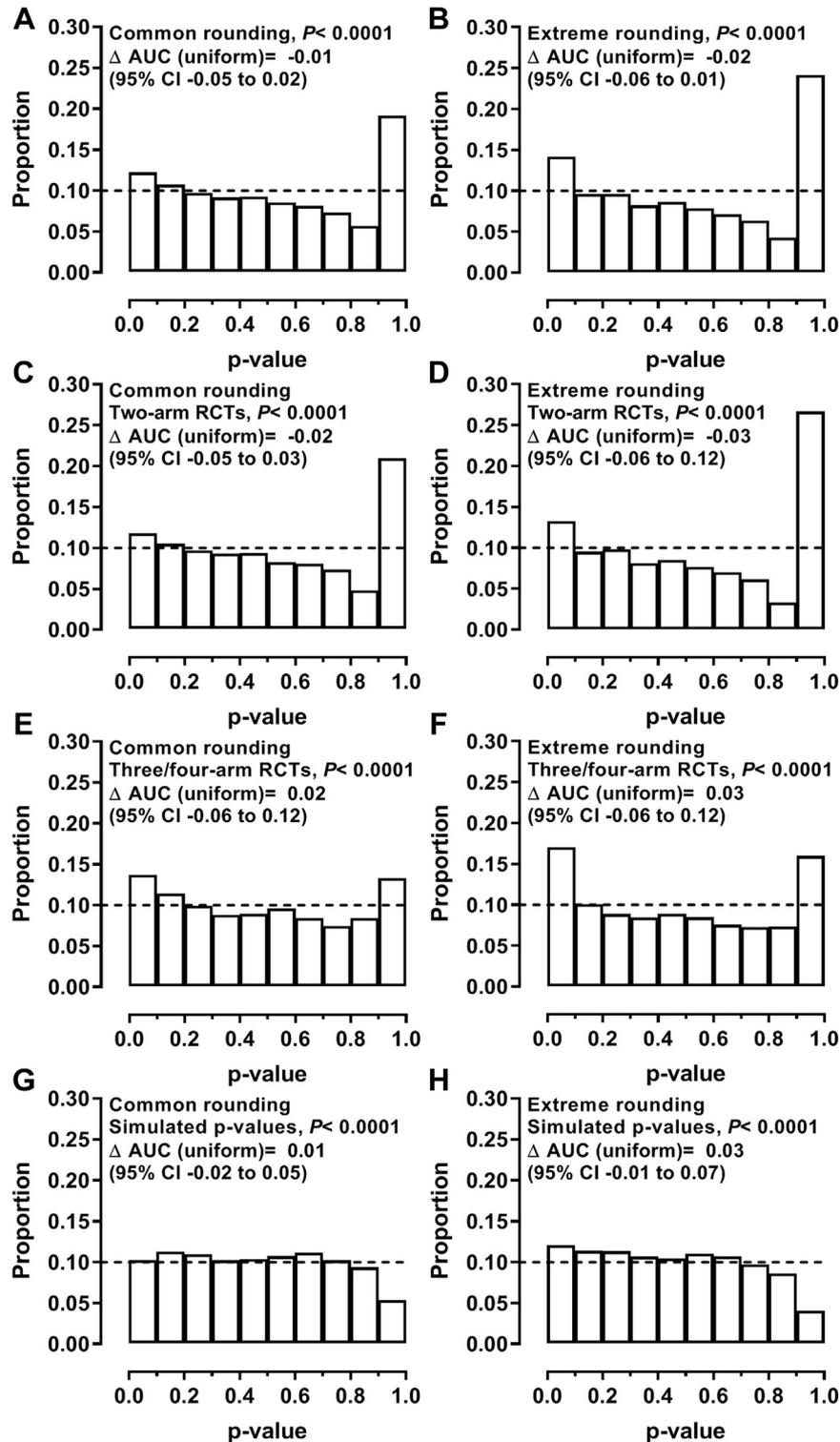
Fig. 7. The effect of rounding on the distribution of baseline *P*-values calculated from summary statistics. The panels show the distribution of *P*-values by decile calculated from rounded summary statistics of all 319 baseline variables in 100 simulated randomizations (with individuals randomized in one block per treatment group) of 13 randomized controlled trials (RCTs). Panel A, variables rounded to a common level (Table 2); Panel B, variables rounded to a more extreme level (Table 2); Panels C and D, common (C) or extreme (D) rounding in two arm RCTs; Panels E and F, common (E) or extreme rounding (F) in three or four arm RCTs; Panels G and H simulated *P*-values from common (G) or extreme (H) rounding—see text for description. The dotted line is the uniform distribution proportion of 0.10. Δ AUC (uniform) is the difference in area under the curve (AUC) of the cumulative distribution function (CDF) from the AUC of the uniform distribution CDF.

conducted RCTs. Carlisle reported that non-normal distribution had little effect on an analysis of baseline *P*-values, whereas highly correlated variables could potentially alter the results but was unlikely to explain the results obtained from analysis of fraudulent data [26]. Taken together with the results of our previous work [6], the current analyses show that the distribution of continuous baseline *P*-values in a group of RCTs is approximately uniform and not significantly affected by the presence of non-normally distributed variables or highly correlated variables that occur in real-life RCTs. Stratified randomization and minimization algorithms may lead to a smaller-than-expected proportion of *P*-values <0.10, but other randomization methods produce uniform baseline *P*-values. Therefore, it seems reasonable to conclude that any differences in the distribution of *P*-values from comparison of baseline continuous variables in a group of valid RCTs from the uniform distribution should only be small. One contributing factor to this conclusion regarding correlated variables might be that the restrictive inclusion criteria generally used in RCTs may produce narrower distributions of variables, which in turn would mean highly correlated variables could be uncommon.

If there is a consistently large or small between-groups difference in a baseline variable in a series of RCTs, the distribution of *P*-values no longer remains uniform in highly correlated data sets (Fig. 6). However, this situation is unlikely to occur in practice in independent RCTs because any consistent difference or similarity between variables in independent RCTs suggests a failure of randomization, unless the similarity is expected as in the case of randomization stratified by a variable or the use of a minimization algorithm.

Reporting baseline *P*-values is not a recommended practice, although it is common [10]. Our results show that when baseline *P*-values are calculated from rounded summary data, their distribution is no longer uniform. The most prominent difference from the uniform distribution is the higher-than-expected proportion of *P*-values >0.9, which is more common in two arm RCTs and when rounding is extreme. It largely arises from the situation when the rounded means in the randomized groups are identical, and therefore the *P*-value from the between-groups comparison is 1. Using simulated *P*-values is not able to overcome this issue and produce a uniform distribution. Therefore, when baseline *P*-values are calculated from rounded summary data, it is no longer appropriate to consider the expected distribution as uniform. Instead, the expected distributions are shown in Fig. 7 (our RCTs) and Appendix Figures A7 and A8 (simulated RCTs).

When the baseline *P*-values from RCTs in which concerns about fraudulent data have been raised can only be calculated from reported rounded summary statistics, it is still possible to compare their distribution with the expected distribution. Visually, the distribution of baseline *P*-values and the AUC-CDF can be compared to the relevant panels in Fig. 7 and Appendix Fig. A7, or with distributions empirically generated from simulations using the reported data. The distribution of baseline *P*-values in two summary data sets known to contain at least some fabricated data (see our previous paper [6]) differs markedly from the expected distributions in Fig. 7 and Appendix Fig. A7. Secondly, the distribution of baseline *P*-values can be compared with a control data set of *P*-values calculated using rounded summary statistics from known genuine RCTs. The distribution of *P*-values obtained through bootstrap resampling can then be repeatedly compared in the two data sets using a two sample Kolmogorov−Smirnov test [3,6].

In summary, randomization methods, non-normality, and correlation of baseline variables do not have important effects on the distribution of baseline *P*-values or the AUC-CDF from groups of RCTs, although stratified randomization and minimization might lead to smaller-than-expected proportion of *P*-values <0.10. In contrast, calculation of *P*-values from rounded summary statistics produces a nonuniform distribution of *P*-values. Nevertheless, the observed distribution can still be compared to the expected distribution of baseline *P*-values. Therefore, assessing the distribution of *P*-values from the comparison of baseline variables in a group of RCTs about which concerns have been raised can be helpful in identifying highly unusual distributions that might support concerns about data integrity and lead to further investigations. The limitations have been discussed previously [6], but in general, the technique seems most appropriate to analyze at least a moderate number of baseline continuous variables from a body of RCTs. Although the results should be interpreted cautiously, large differences between observed and expected distributions of baseline *P*-values justify further investigation.

Acknowledgments

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.03.001.

References

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

[2] Yentis SM. Lies, damn lies, and statistics. Anaesthesia 2012;67:455−6.

[3] 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.

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

[5] Gross RA. Statistics and the detection of scientific misconduct. Neurology 2016;87:2388.

[6] Bolland MJ, Avenell A, Gamble GD, Lumley TS, Grey A. Assessment of the distribution of p-values from comparison of baseline values in randomised controlled trials: implications for investigation of data integrity. J Clin Epidemiol 2018. [Submitted].

[7] Mascha EJ, Vetter TR, Pittet JF. An appraisal of the Carlisle-Stouffer-Fisher method for assessing study data integrity and fraud. Anesth Analg 2017;125:1381—5.

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

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

[10] Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[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—7.

[23] 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.

[24] Higham NJ. Computing the nearest correlation matrix-a problem from finance. IMA J Numer Anal 2002;22:329—43.

[25] Wicklin R. The do loop blog: computing the nearest correlation matrix. Available at http://blogssascom/content/iml/2012/11/28/computing-the-nearest-correlation-matrixhtml;. Accessed July 20, 2017.

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