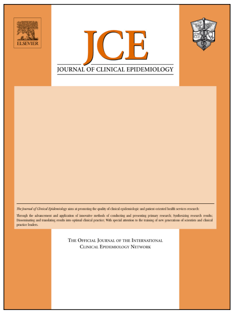


Rounding, but not randomisation method, non-normality, or correlation, affected baseline p-value distributions in randomised trials

Mark J. Bolland, MBChB, PhD, Greg D. Gamble, MSc, Alison Avenell, MBBS, MD, Andrew Grey, MD

PII: S0895-4356(18)31032-1

DOI: https://doi.org/10.1016/j.jclinepi.2019.03.001

Reference: JCE 9839

To appear in: *Journal of Clinical Epidemiology*

Received Date: 15 November 2018

Revised Date: 3 February 2019

Accepted Date: 1 March 2019

Title: Rounding, but not randomisation method, non-normality, or correlation, affected baseline p-value distributions in randomised trials

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

Alison Avenell- Health Services Research Unit, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, Scotland

Word counts:

Text only: 3827
Abstract: 200
Figures: 7
Tables: 2
Appendix: Figures A.1- A.7, Appendix B, Appendix C

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

Abstract:

Objective:

To investigate whether comparing observed with expected p-value distributions for baseline continuous variables in randomised controlled trials (RCTs) might be limited by randomisation 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 randomisation methods were consistent with the theoretical uniform distribution and AUC-CDF respectively, although stratification and minimisation produced smaller-than-expected proportions of p-values <0.10. 77% of 3813 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:

Randomisation 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.

Key words: Statistical Methods, Research Integrity, P-values, Correlation, Randomisation, Rounding

Running title: Factors affecting baseline p-value distribution

What is new?

- Non-normal distribution of baseline continuous variables, eight common randomisation 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.
- Concerns that correlation and non-normality of baseline variables or randomisation methods would impact on baseline p-value distribution in genuine RCTs do not appear to be justified.
- Distribution of baseline p-values calculated from rounded summary statistics should be compared to empirically generated distributions not the uniform distribution.

1.1 Introduction

The distribution of baseline variables in groups of randomised controlled trials (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 a RCT are randomly allocated to study groups, the expected distribution of p-values from comparisons between randomised 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 dataset of individual patient data from 13 placebo-controlled RCTs carried out by our research group over the past 20 years [6]. In this dataset, the average distribution of baseline p-values from comparisons of continuous variables was uniform, with only small differences from the expected distributions. However, it has been suggested that baseline p-values may not be uniform when techniques other than simple randomisation 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 randomisation 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 dataset of our trials and in simulations.

Many groups and journals recommend against reporting of p-values from between-group comparisons of baseline variables in manuscripts, 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 AUC-CDF.

1.2 Methods

1.2.1 Datasets of RCTs:

We pooled anonymised 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, osteoporosis), three in healthy women with osteopenia, and the other six in healthy individuals. Randomisation was by a minimisation algorithm for 1 RCT, using stratification for 1 RCT, and using variable block sizes for 11 RCTs. For these analyses, the pooled dataset 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 dataset contained 319 variables from the 13 RCTs.

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

Study	N	Mean Age (y)	Population	Design	Agent	Baseline variables (N)	Randomisation 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]	1471	74	Older women	2-arm	Calcium	19	Minimisation (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

1.2.2 Calculation of baseline p-values

We compared the means of the baseline variables between randomised 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 non-parametric 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 re-randomised using the original randomisation method (Table 1) and compared the baseline variables with a t-test or one-way ANOVA for each re-randomisation.

1.2.3 Effects of randomisation methods

We generated a dataset with 100 simulated randomisations of the 13 RCTs using eight different randomisation methods. Separate randomisations 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 randomisation were used to form simulated treatment groups. 1. Simple randomisation: 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. Randomisation 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 respectively were used. 5. Stratified one block: 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. Minimisation: treatments were assigned using a minimisation 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 minimisation: there was an 80% chance of treatment being assigned using minimisation and a 20% chance of treatment being assigned using simple randomisation. 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 minimisation analyses.

1.2.4 Effect of normality of distribution

We assessed whether the baseline continuous variables in the pooled original dataset 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 dataset of 100 simulations of normally distributed baseline variables (the ‘simulated normal dataset’) 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 randomised in a single block per treatment group to ensure group numbers were similar. Analyses were repeated in this dataset, in which there were few variables that were highly non-normally distributed.

1.2.5 Effect of correlation of baseline variables

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

To assess the effect of increasing strength of correlation of variables on the distribution of baseline p-values, we generated a dataset 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 randomised 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 randomised groups in the simulated datasets 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.

1.2.6 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 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 manuscript, and secondly an extreme level of rounding (Table 2). We performed these analyses in the dataset of RCTs with 100 simulated randomisations (with individuals randomised in one block per treatment group to ensure group numbers were similar), and in the simulated normal dataset.

Table 2: 30 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	Haemoglobin (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 randomised controlled trials. The rounding columns show two different levels used to round summary statistics from which baseline p-values were calculated.

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 1000 simulations for each variable in the dataset of 100 simulated randomisations (with individuals randomised in one block per treatment group) and in the simulated normal dataset. 1000 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, 1000 values uniformly distributed between 29.5 and 30.5 were generated. Likewise, for a standard deviation of 0.15, 1000 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 1000 p-values from the simulations used in place of the p-value calculated from the summary statistics.

1.2.7 Analyses

All analyses were performed with SAS (SAS Institute, Cary, NC 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. 95% confidence intervals (CI) for the AUC for the 319 baseline p-values were calculated from a dataset of 100 re-randomisations using the original trial randomisation method [6]; for other analyses they were calculated from the 2.5 and 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.

1.3 Results:

1.3.1 Effect of randomisation method:

The distribution of p-values from the comparison of the 319 baseline variables between the randomised 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], Figure 1), although some proportions for individual deciles differed from the expected proportions. Figure 1 also shows that in the dataset of 100 re-randomisations using the original randomisation method (Table 1), the distribution of p-values was approximately uniform.

Figure 2 shows the results of 100 simulated randomisations using eight different methods. For simple randomisation and randomisation in blocks, the distribution of baseline p-values was approximately uniform (Figures 2A-2D). When stratification or minimisation was used (Figures 2E-2H), visually there appeared to be a smaller-than-expected proportion of p-values in the lowest decile, consistent with the pattern seen in Figure 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 minimisation; 0.05-0.12 weighted minimisation) included the expected value of 0.10. Figure 2 and Appendix Figure A.1 show that the AUC-CDF was consistent with the uniform AUC for all randomisation methods.

1.3.2 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 non-parametric Wilcoxon or Kruskal-Wallis tests (Figure 3A, Appendix Figure A.2A), 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 (Figure 1). In the analyses of p-values from both the parametric and non-parametric 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 non-parametric p-values). We then generated a dataset of 100 simulations of normally distributed variables (the ‘simulated normal dataset’) based on the means and standard deviations for each variable from each RCT. Figure 3B and Appendix Figure A.2B show that the distribution of p-values from the comparison of baseline variables and the AUC-CDF in the simulated normal dataset is consistent with the uniform distribution.

Next, we restricted the analyses from both the pooled dataset 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 Figure A.2C-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.

1.3.3 Effect of correlation of baseline variables

We determined the correlation between baseline variables in each individual RCT. Of the 3813 pairwise correlations in the individual RCTs, Figure 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 dataset, Figure 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 dataset 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. Figure 5 and Appendix Figures A.3, A.4 show the distribution of correlations and AUC-CDF for each simulated level of correlation and that, despite the increases in correlation, the distribution of baseline p-values is uniform and the AUC of the CDF 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 Figure A.5 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 randomised groups in the simulated datasets 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 consistent with the uniform AUC (Figure 6A, Appendix Figure A.6A). 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 (Figures 6B-D left panel, Appendix Figure A.6B-F, left panel). For the simulations with $P>0.90$ for age, increasing the correlations only leads to a clear non-uniform distribution and change in the AUC-CDF when the increased correlation was large (Figures 6B-6D, Appendix Figure A.6B-F, right panel).

1.3.4 Baseline p-values calculated from rounded summary statistics

In the dataset of RCTs with 100 simulated randomisations, 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 (Figure 7A, Appendix Figure A.7). These effects were more pronounced when extreme rounding was used (Figure 7B), and also in RCTs with two arms compared to those with three or four arms (Figures 7C-7F).

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-7F. To determine whether simulating p-values might produce a more uniform distribution, we used the mean of the 1000 p-values calculated from 1000 simulated means and standard deviations for each variable in each treatment group for each of the 13 RCTs. Figures 7G,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 dataset. Appendix Figure A.8 shows that the results from these analyses are very similar to those from the analyses of the dataset of RCTs with 100 simulated randomisations.

1.4 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 randomisation 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 randomisation and minimisation 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 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 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 randomisation and minimisation algorithms may lead to a smaller-than-expected proportion of p-values <0.10, but other randomisation 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 datasets (Figure 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 randomisation, unless the similarity is expected as in the case of randomisation stratified by a variable or the use of a minimisation algorithm.

Reporting baseline p-values is not a recommended practice, though 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 randomised 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 Figures 7 (our RCTs) and Appendix Figure A.7, A.8 (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 Figure 7 and Appendix Figure A.7, or with distributions empirically generated from simulations using the reported data. The distribution of baseline p-values in two summary datasets known to contain at least some fabricated data (see our previouspaper [6]) differ markedly from the expected distributions in Figure 7 and Appendix Figure A.7. Secondly, the distribution of baseline p-values can be compared with a control dataset 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 datasets using a two sample Kolmogorov–Smirnov test [3, 6].

In summary, randomisation 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 randomisation and minimisation might lead to smaller-than-expected proportion of p-values <0.10. In contrast, calculation of p-values from rounded summary statistics produces a non-uniform 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 analyse at least a moderate number of baseline continuous variables from a body of RCTs. While the results should be interpreted cautiously, large differences between observed and expected distributions of baseline p-values justify further investigation.

Competing Interests

The authors declare that they have no competing interests.

Funding

No specific funding was received for this study. MB 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.

Authors contributions

MB, GG, AA, and AG designed the research. MB extracted the 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.

Acknowledgements: The authors acknowledge the contribution of Professor Thomas Lumley who provided expert guidance in using the area under the curve of the cumulative distribution function of the baseline p-values for these analyses

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;87:2391-2402.

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;125:1381-1385.

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;109:362-370.

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;98:2301-2307.

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;22:329-343.

25. Wicklin R. The do loop blog: computing the nearest correlation matrix. http://blogssascom/content/iml/2012/11/28/computing-the-nearest-correlation-matrixhtml [accessed 20/7/2017]. 2012

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.

Figure 1: Distribution of 319 p-values for 30 baseline variables in 13 randomised 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 re-randomisations of the original trial data (n=319 variables, 100 randomisations, 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 dataset of 100 original trial re-randomisations.

Figure 2: Effect of randomisation method on baseline p-value distribution. Distribution of 319 p-values for all 30 variables from all 13 randomised controlled trials in 100 simulated randomisations using eight different methods. Panel A simple randomisation. Panel B one block per treatment group. Panel C permuted randomisation 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 respectively. Panel G minimisation algorithm for age, weight and lumbar spine bone density. Panel H weighted minimisation (20% chance of simple randomisation, 80% chance of minimisation 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.

Figure 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 randomised controlled trials (RCTs) using non-parametric 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,D results from baseline variables from the 13 RCTs (Figure 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 dataset of 100 re-randomisations using the original trial randomisation method (Figure 1) in Panels A,C,D, and from the raw data in Panel B.

Figure 4: Distribution of pairwise correlations between baseline values in 13 randomised controlled trials (RCTs) and in the simulated normal dataset.

The left panel shows the results from the 13 RCTs and the right panel from the simulated normal dataset 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.

Figure 5: the effect of increasing correlation of baseline variables on the distribution of p-values.

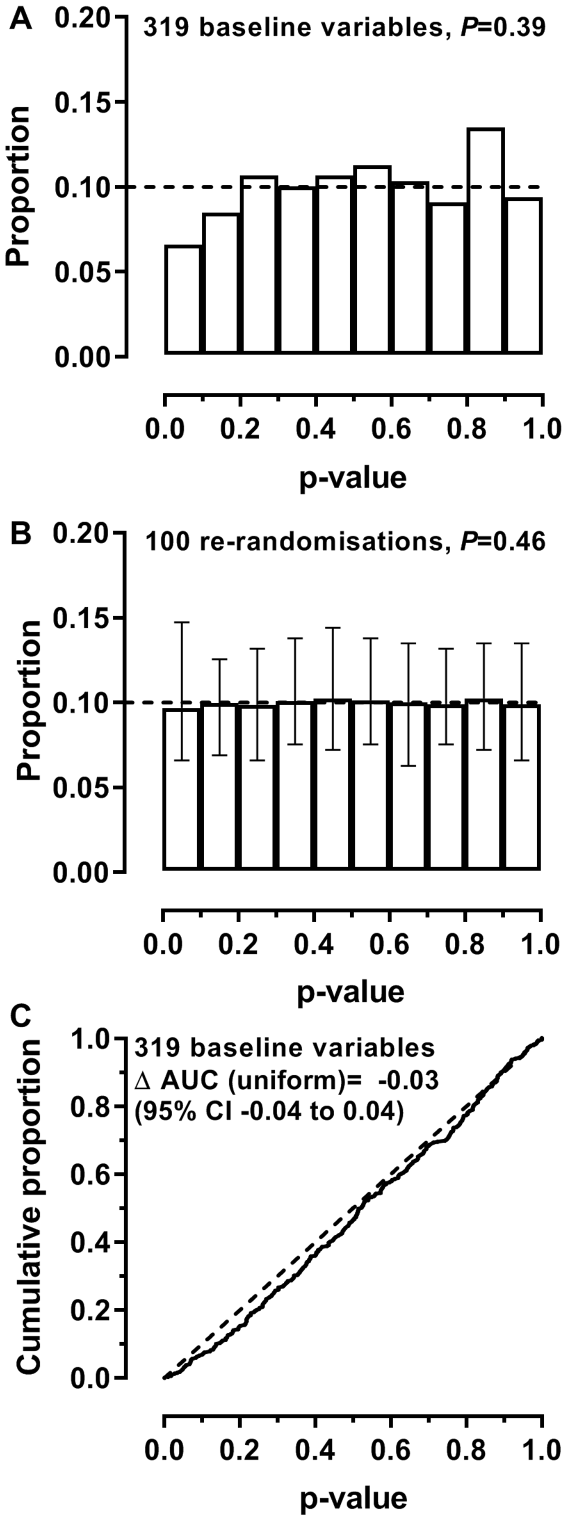
The left panels show the distribution of p-values by decile and the right panels the distribution of correlation statistics in 100 simulations of five variables (age, height, weight, lumbar spine bone density and serum creatinine) based on the mean, standard deviation and covariance matrix from each randomised controlled trial. Panel A uses the covariance matrix from each trial, while 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.

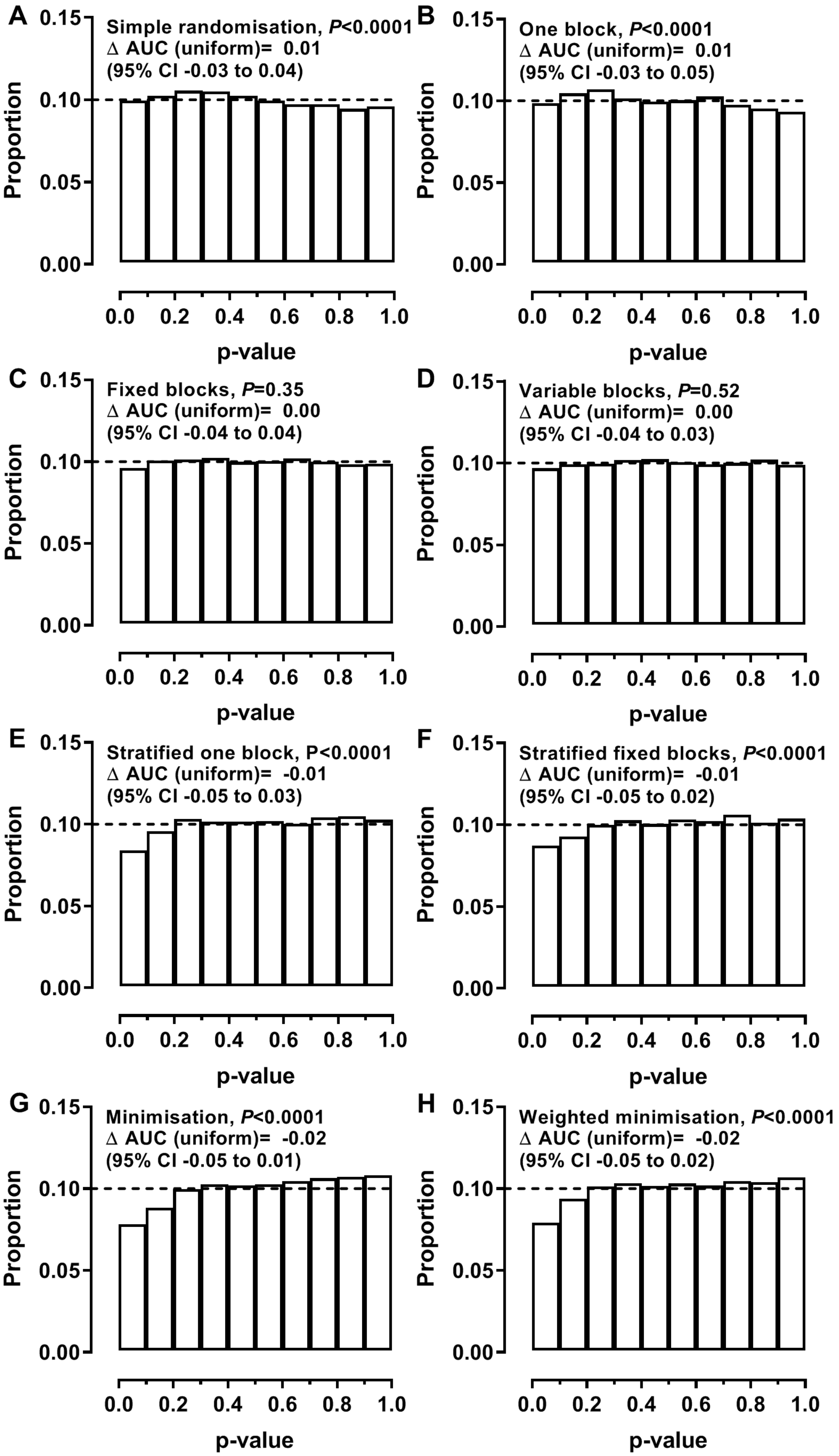
Figure 6: the effect of clustering and increasing correlation of baseline variables on the distribution of p-values.

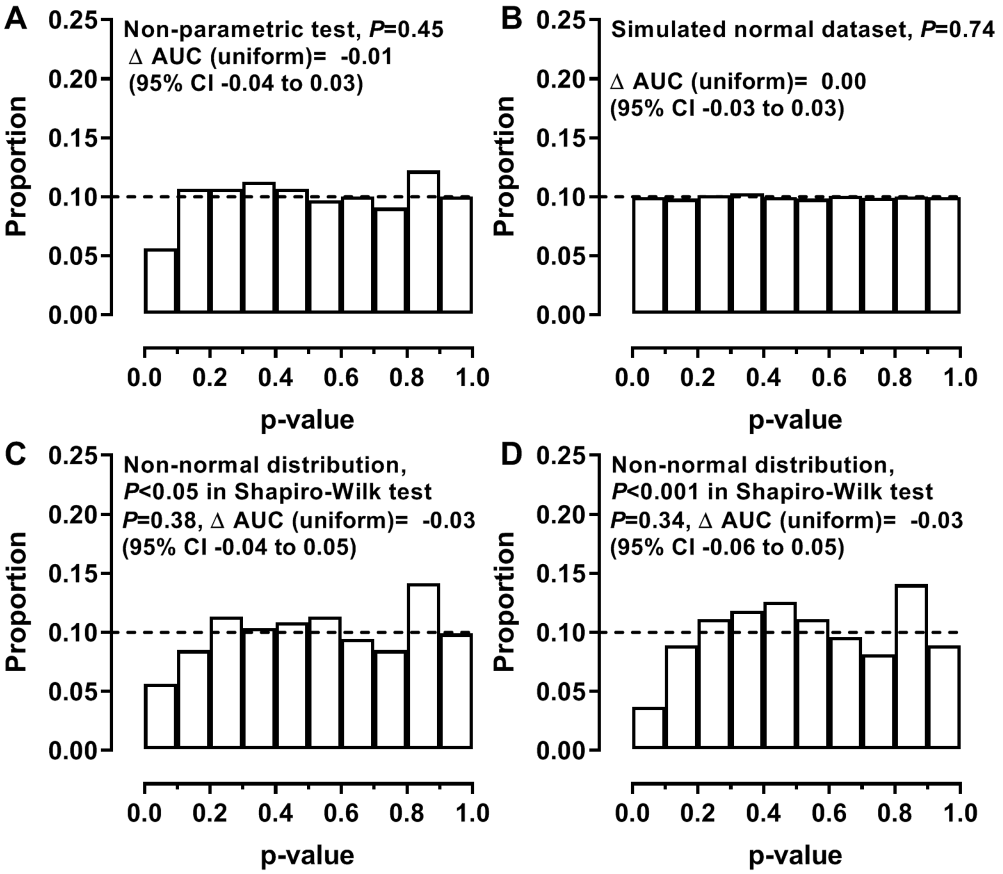
The panels show the distribution of p-values by decile in 100 simulations for four variables (height, weight, lumbar spine bone density and serum creatinine) based on the mean, standard deviation and covariance matrix from each randomised controlled trial (RCT). The left panels are restricted to simulations 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, while 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.

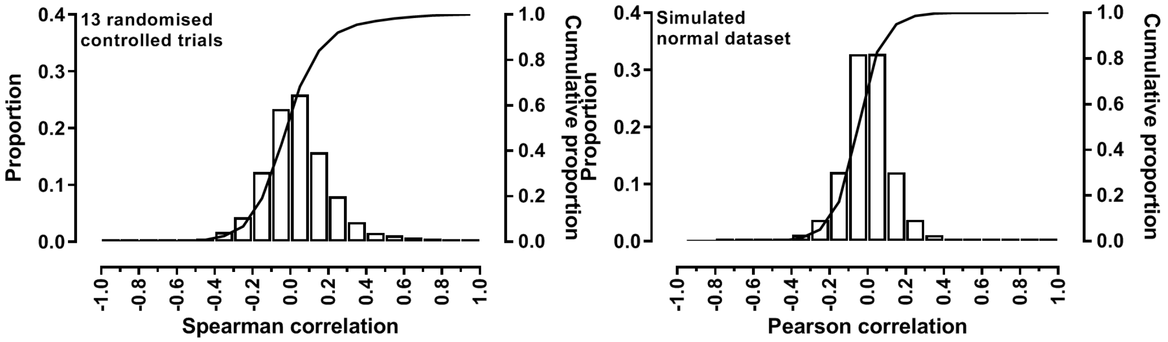
Figure 7: the effect of rounding on the distribution of baseline p-values calculated from summary statistics.

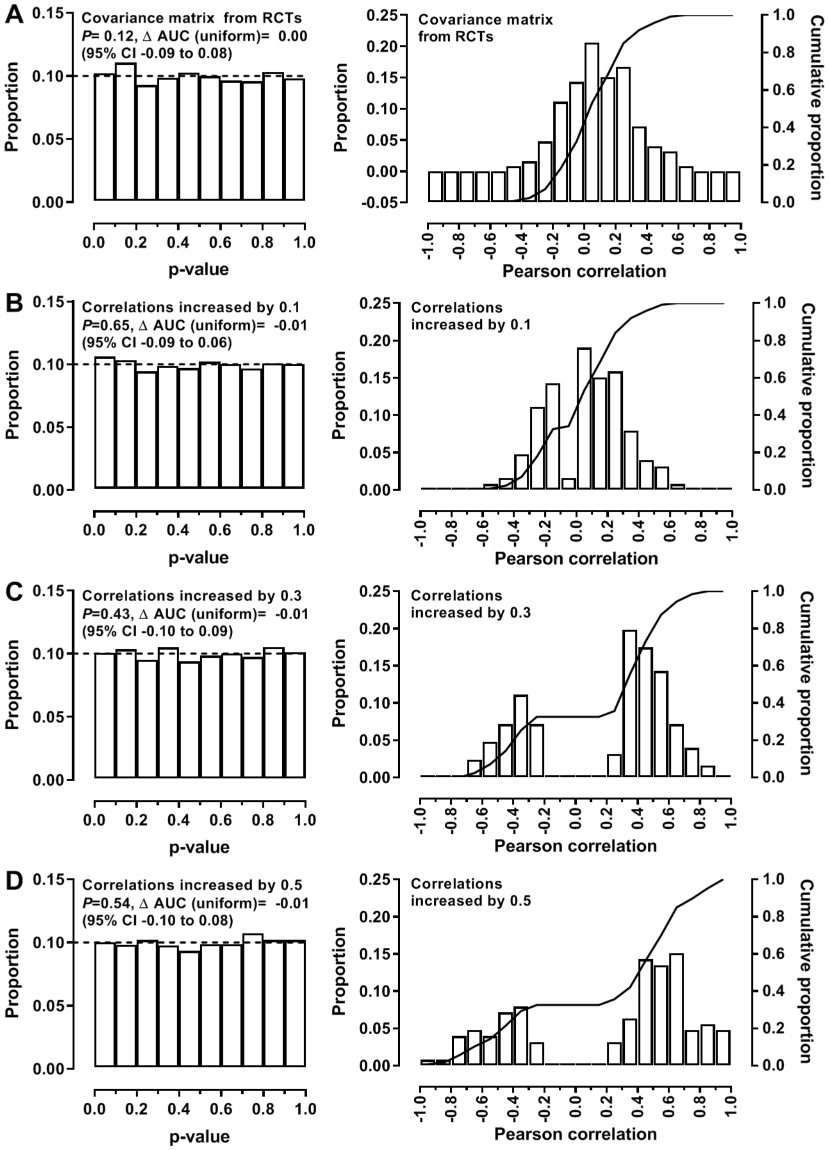
The panels show the distribution of p-values by decile calculated from rounded summary statistics of all 319 baseline variables in 100 simulated randomisations (with individuals randomised in one block per treatment group) of 13 randomised controlled trials (RCTs). Panel A, variables rounded to a common level (Table 2); Panel B, variables rounded to a more extreme level (Table 2); Panels C-D, common (C) or extreme (D) rounding in two arm RCTs; Panels E-F, common (E) or extreme rounding (F) in three or four arm RCTs; Panels G-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. ∆ 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.

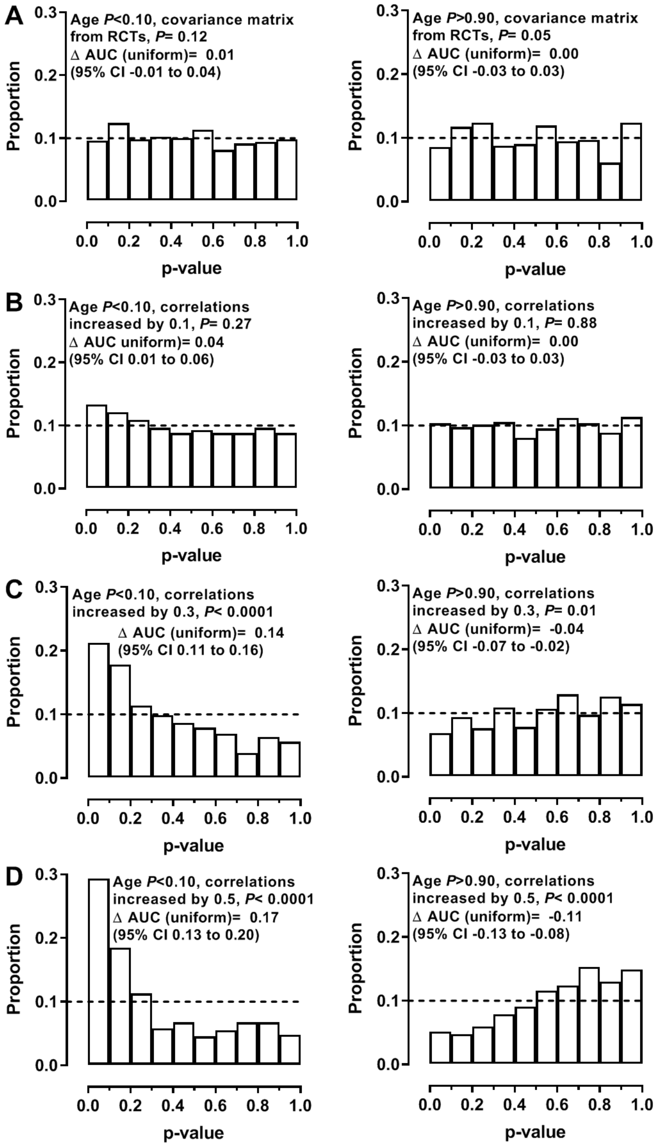

A
0.20
0.15
0.10
0.05
0.00
Proportion
319 baseline variables, P=0.39
0.0
0.2
0.4
0.6
0.8
1.0
p-value
B
0.20
0.15
0.10
0.05
0.00
Proportion
100 re-randomisations, P=0.46
0.0
0.2
0.4
0.6
0.8
1.0
p-value
C
1.0
0.8
0.6
0.4
0.2
0.0
Cumulative proportion
319 baseline variables
Δ AUC (uniform)= -0.03
(95% CI -0.04 to 0.04)
0.0
0.2
0.4
0.6
0.8
1.0
p-value

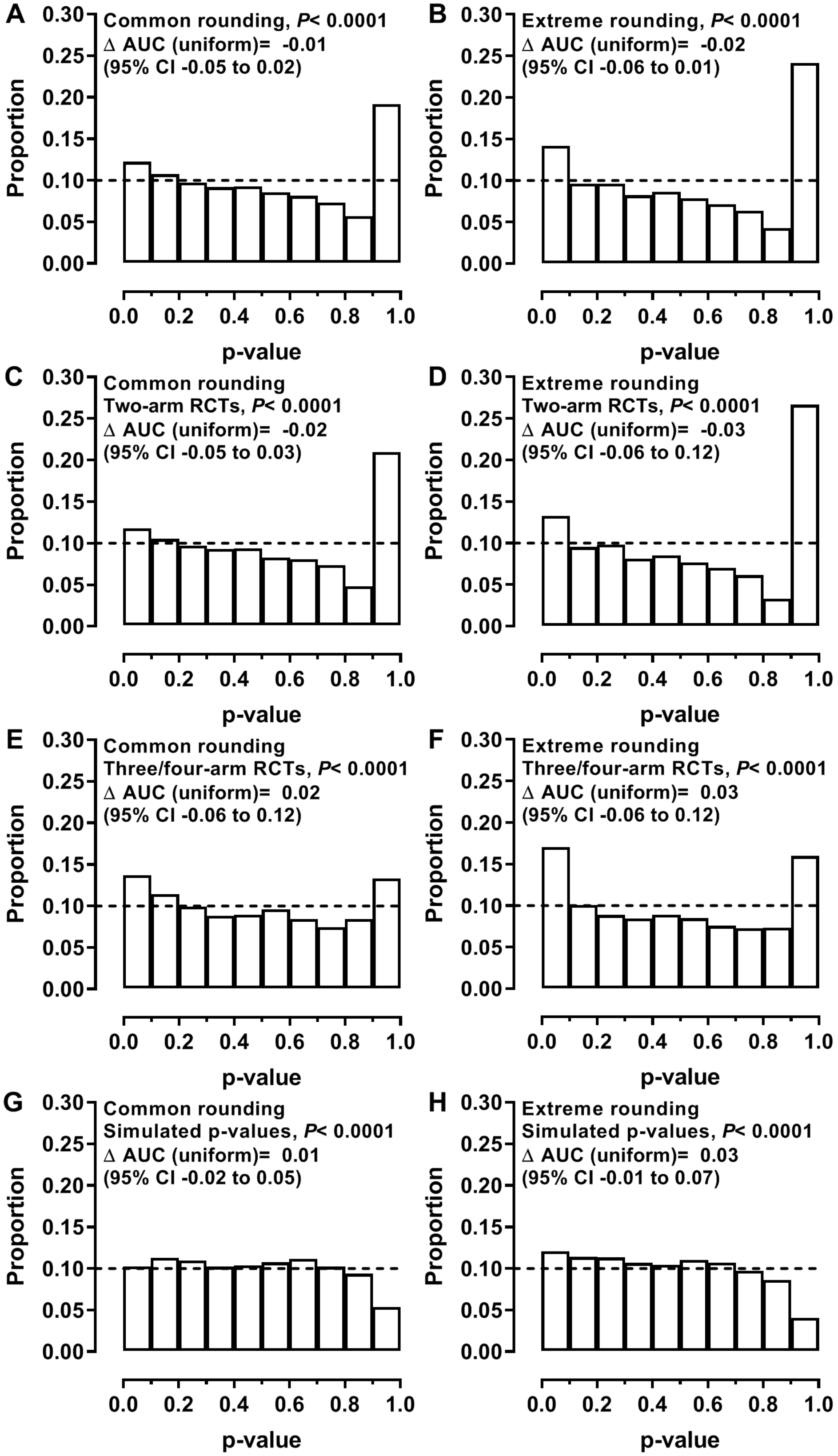

A
Simple randomisation, *P*<0.0001
Δ AUC (uniform)= 0.01
(95% CI -0.03 to 0.04)
B
One block, *P*<0.0001
Δ AUC (uniform)= 0.01
(95% CI -0.03 to 0.05)
C
Fixed blocks, *P*=0.35
Δ AUC (uniform)= 0.00
(95% CI -0.04 to 0.04)
D
Variable blocks, *P*=0.52
Δ AUC (uniform)= 0.00
(95% CI -0.04 to 0.03)
E
Stratified one block, P<0.0001
Δ AUC (uniform)= -0.01
(95% CI -0.05 to 0.03)
F
Stratified fixed blocks, *P*<0.0001
Δ AUC (uniform)= -0.01
(95% CI -0.05 to 0.02)
G
Minimisation, *P*<0.0001
Δ AUC (uniform)= -0.02
(95% CI -0.05 to 0.01)
H
Weighted minimisation, *P*<0.0001
Δ AUC (uniform)= -0.02
(95% CI -0.05 to 0.02)
Proportion
0.15
0.10
0.05
0.00
p-value
0.0
0.2
0.4
0.6
0.8
1.0


A
Non-parametric test, P=0.45
∆ AUC (uniform)= -0.01
(95% CI -0.04 to 0.03)
B
Simulated normal dataset, P=0.74
∆ AUC (uniform)= 0.00
(95% CI -0.03 to 0.03)
C
Non-normal distribution,
P<0.05 in Shapiro-Wilk test
P=0.38, ∆ AUC (uniform)= -0.03
(95% CI -0.04 to 0.05)
D
Non-normal distribution,
P<0.001 in Shapiro-Wilk test
P=0.34, ∆ AUC (uniform)= -0.03
(95% CI -0.06 to 0.05)
Proportion
0.00
0.05
0.10
0.15
0.20
0.25
p-value
0.0
0.2
0.4
0.6
0.8
1.0


13 randomised
controlled trials
Proportion
0.4
0.3
0.2
0.1
0.0
Cumulative proportion
1.0
0.8
0.6
0.4
0.2
0.0
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
Spearman correlation
Simulated
normal dataset
Proportion
0.4
0.3
0.2
0.1
0.0
Cumulative proportion
1.0
0.8
0.6
0.4
0.2
0.0
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
Pearson correlation


A
Covariance matrix from RCTs
P= 0.12, Δ AUC (uniform)= 0.00
(95% CI -0.09 to 0.08)
Proportion
0.15
0.10
0.05
0.00
0.0
0.2
0.4
0.6
0.8
1.0
p-value
Covariance matrix
from RCTs
Proportion
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
Cumulative proportion
1.0
0.8
0.6
0.4
0.2
0.0
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
Pearson correlation
B
Correlations increased by 0.1
P=0.65, Δ AUC (uniform)= -0.01
(95% CI -0.09 to 0.06)
Proportion
p-value
Correlations
increased by 0.1
Proportion
Cumulative proportion
Pearson correlation
C
Correlations increased by 0.3
P=0.43, Δ AUC (uniform)= -0.01
(95% CI -0.10 to 0.09)
Proportion
p-value
Correlations
increased by 0.3
Proportion
Cumulative proportion
Pearson correlation
D
Correlations increased by 0.5
P=0.54, Δ AUC (uniform)= -0.01
(95% CI -0.10 to 0.08)
Proportion
p-value
Correlations
increased by 0.5
Proportion
Cumulative proportion
Pearson correlation


A
Age P<0.10, covariance matrix from RCTs, P= 0.12
∆ AUC (uniform)= 0.01
(95% CI -0.01 to 0.04)
Age P>0.90, covariance matrix from RCTs, P= 0.05
∆ AUC (uniform)= 0.00
(95% CI -0.03 to 0.03)
B
Age P<0.10, correlations increased by 0.1, P= 0.27
∆ AUC uniform)= 0.04
(95% CI 0.01 to 0.06)
Age P>0.90, correlations increased by 0.1, P= 0.88
∆ AUC (uniform)= 0.00
(95% CI -0.03 to 0.03)
C
Age P<0.10, correlations increased by 0.3, P< 0.0001
∆ AUC (uniform)= 0.14
(95% CI 0.11 to 0.16)
Age P>0.90, correlations increased by 0.3, P= 0.01
∆ AUC (uniform)= -0.04
(95% CI -0.07 to -0.02)
D
Age P<0.10, correlations increased by 0.5, P< 0.0001
∆ AUC (uniform)= 0.17
(95% CI 0.13 to 0.20)
Age P>0.90, correlations increased by 0.5, P< 0.0001
∆ AUC (uniform)= -0.11
(95% CI -0.13 to -0.08)
Proportion
0.0
0.1
0.2
0.3
p-value
0.0
0.2
0.4
0.6
0.8
1.0


A
Common rounding, *P*< 0.0001
Δ AUC (uniform)= -0.01
(95% CI -0.05 to 0.02)
B
Extreme rounding, *P*< 0.0001
Δ AUC (uniform)= -0.02
(95% CI -0.06 to 0.01)
C
Common rounding
Two-arm RCTs, *P*< 0.0001
Δ AUC (uniform)= -0.02
(95% CI -0.05 to 0.03)
D
Extreme rounding
Two-arm RCTs, *P*< 0.0001
Δ AUC (uniform)= -0.03
(95% CI -0.06 to 0.12)
E
Common rounding
Three/four-arm RCTs, *P*< 0.0001
Δ AUC (uniform)= 0.02
(95% CI -0.06 to 0.12)
F
Extreme rounding
Three/four-arm RCTs, *P*< 0.0001
Δ AUC (uniform)= 0.03
(95% CI -0.06 to 0.12)
G
Common rounding
Simulated p-values, *P*< 0.0001
Δ AUC (uniform)= 0.01
(95% CI -0.02 to 0.05)
H
Extreme rounding
Simulated p-values, *P*< 0.0001
Δ AUC (uniform)= 0.03
(95% CI -0.01 to 0.07)
Proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
p-value
0.0
0.2
0.4
0.6
0.8
1.0

What is new?

- Non-normal distribution of baseline continuous variables, eight common randomisation 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.
- Concerns that correlation and non-normality of baseline variables or randomisation methods would impact on baseline p-value distribution in genuine RCTs do not appear to be justified.
- Distribution of baseline p-values calculated from rounded summary statistics should be compared to empirically generated distributions not the uniform distribution.

Authors contributions
MB, GG, AA, and AG designed the research. MB extracted the 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.

Acknowledgements: The authors acknowledge the contribution of Professor Thomas Lumley who provided expert guidance in using the area under the curve of the cumulative distribution function of the baseline p-values for these analyses

Declarations of interest: none