Are these data real? Statistical methods for the detection of data fabrication in clinical trials

Sanaa Al-Marzouki, Stephen Evans, Tom Marshall, Ian Roberts

Abstract

Objectives To test the application of statistical methods to detect data fabrication in a clinical trial.
Setting Data from two clinical trials: a trial of a dietary intervention for cardiovascular disease and a trial of a drug intervention for the same problem.
Outcome measures Baseline comparisons of means and variances of cardiovascular risk factors; digit preference overall and its pattern by group.
Results In the dietary intervention trial, variances for 16 of the 22 variables available at baseline were significally different, and 10 significant differences were seen in means for these variables. Some of these P values were extraordinarily small. Distributions of the final recorded digit were significantly different between the intervention and the control group at baseline for 14/22 variables in the dietary trial. In the drug trial, only five variables were available, and no significant differences between the groups for baseline values in means or variances or digit preference were seen.
Conclusions Several statistical features of the data from the dietary trial are so strongly suggestive of data fabrication that no other explanation is likely.

Introduction

Most statistical analyses of clinical trials are undertaken on the presumption that the data are genuine. Large accidental errors can be detected during data analysis,[1 2] but if people are trying to "make up" data they are likely to do it in such a way that it is not immediately obvious, avoiding any large discrepancies. Nevertheless, fraudulent data have particular statistical features that are not evident in data containing accidental errors, and several analytical methods have been developed to detect fraud in clinical trials.[3 4] The *BMJ* has taken a general interest in this field and has published a book on fraud and misconduct, now in its third edition, which has a chapter on statistical methods of detection of fraud.[5]

In this paper we use statistical techniques to examine data from two randomised controlled trials. In one trial, the possibility of scientific misconduct had been raised by *BMJ* referees, based on inconsistencies in calculated P values compared with the means, standard deviations, and sample sizes presented (see p 281). For comparison, we used the same methods to analyse a second trial for which there were no such concerns. We were not involved in either trial.

Methods

The trial about which doubts were raised (the diet trial) was a single blind, randomised controlled trial of the effects of a fruit and vegetable enriched diet in 831 patients with coronary heart disease, including patients with angina pectoris, myocardial infarction, or surrogate risk factors. Study participants were stated to be randomly allocated to the intervention diet (Group I, n = 415) or to the control group, which was the patient's usual diet (Group C, n = 416). The aim was to examine the effect of the intervention diet on risk factors for coronary artery disease after two years. We do not present data from the two year follow-up, because differences between groups could arise as a result of the interventions. After the reviewers had expressed suspicions about the integrity of the data, the *BMJ* requested the original trial data. These were provided by the trial's first author on handwritten sheets, which we entered on to computer, making appropriate checks to avoid transcription errors. The data are considered in the two randomised groups at baseline, Group I and Group C.

The second ("drug") trial was a randomised controlled trial of the effects of drug treatment in 21 750 patients with mild hypertension from 31 centres, from which we randomly selected five centres with 838 patients who had complete data for the selected variables. Study participants were randomly allocated to receive the drug (Group I, N = 403) or a placebo (Group C, N = 435). The aim was to determine whether drug treatment reduced the occurrence of stroke, death due to hypertension and coronary events in men and women aged 35-64 years, when followed for two years (again we do not present data from the follow-up). The drug trial data were provided by the trial investigators as computer files. The data are presented by treatment group (I or C) at baseline, using the same notation as for the diet trial. The variables in this study in common with the diet study are weight, diastolic blood pressure, systolic blood pressure, cholesterol measurements, and height. Further details of the methods and results from that trial have been published.[6]

Statistical methods

We conducted various tests on the baseline data of the randomised groups in both trials, looking for patterns that might indicate that the data in the diet trial were not generated by the normal process of making and recording individual measurements on a series of patients. We used the data from the drug trial for comparison, since we expected them to show patterns typical of data collected normally during a trial.

Using basic descriptive statistics and conventional statistical significance tests we compared the baseline data in the randomised groups in both trials. In a randomised trial, the data at baseline should be similar in the randomised groups. (The mean, the variability, the shape of the distribution of the data, and the pattern of data resulting from the methods of measurement must be similar since the groups can differ from one another only by chance factors.) This is the reason why in general, tests for statistical significance are not conducted at baseline in genuine trials. If such tests are carried out about one in 20 of such tests will be significant purely

See also p 281, and *Editorial* by Smith and Godlee

Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT
Sanaa Al-Marzouki *research student*
Stephen Evans *professor of pharmacoepidemiology, Medical Statistics Unit*
Tom Marshall *senior lecturer in medical statistics*
Ian Roberts *professor of epidemiology and public heath*
Correspondence to: S Evans stephen.evans@Lshtm.ac.uk

BMJ 2005;331:267–70

Table 1 Baseline variables in the two trials under comparison

	Diet		Drug	
	Intervention	Control	Intervention	Control
Weight (kg):				
Mean	65.74	65.59	70.27	70.08
Median	66	66	70	69
Mode	65	65	70	61
SD	7.89	7.64	11.6	12.4
Min	40	39	40	36
Max	87	85	111	120
Height (cm):				
Mean	165.1	165.28	162.1	162.6
Median	165	165	160	163
Mode	165	165	160	157
SD	6.91	3.93	9.22	9.14
Min	140	140	138	140
Max	179	178	190	188
Systolic blood pressure (mm Hg):				
Mean	134.2	131.9	184.4	184.6
Median	130	130	185	184
Mode	130	130	186	181
SD	18.5	16.9	12.2	12.9
Min	100	100	160	160
Max	200	195	209	210
Diastolic blood pressure (mm Hg):				
Mean	86.5	86.7	91.8	91.2
Median	86	85	92	91
Mode	80	85	101	90
SD	9.98	9.2	10.8	11.4
Min	60	60	46	50
Max	112	120	114	115
Cholesterol (mmol/l):				
Mean	5.46	5.43	6.68	6.57
Median	5.48	5.48	6.6	6.5
Mode	5.43	5.43	6.4	6.1
SD	0.352	0.296	1.26	1.21
Min	4.53	2.95	3.6	3.7
Max	6.52	6.00	12	10.8

by chance. We used *t* tests to compare the means of the randomised groups and F tests to compare the variances (standard deviations).

Data that are recorded (or invented) by people (as opposed to machines) tend to show preferences for certain numbers, such as rounding to the nearest 5 or 10. This is seen in the last recorded digit of numbers, and is called "digit preference." This digit preference should be similar between groups formed just by a chance process—randomisation. We used χ^2 tests to examine whether there was any tendency for the last digit to take on particular values and whether any observed digit preference was the same in the two groups created by randomisation. Digit preference can occur in all legitimate data based on human recording, but any pattern of this preference should be similar between groups formed using randomisation. We used SPSS, version 12.0.1 (Chicago, USA), for our data analysis.

Results

Table 1 shows descriptive summaries of variables common to both trials for both groups in each trial. The drug trial values show what might be expected in a randomised trial, but the diet trial shows notable differences in standard deviations for height and cholesterol measurements.

Table 2 shows for each trial the results of *t* and F tests, for differences in means and also in variances between the intervention and control groups at baseline for all available variables. In a genuine trial, correctly randomised, any such differences would be due to chance. Usually P values should not be quoted to greater precision than $P<0.001$, but because of the extreme nature of these P values, their exact value is given. In the diet trial, differences in variances were significant for 16 of the 22 variables that were available, as were 10 differences in means for these variables. Several of the P values were extraordinarily small. The expectation is that about 5% of such comparisons would have $P<0.05$, and extremely small P values should not occur. In the drug trial, none of the baseline means and none of the baseline variances showed statistically significant differences between the two groups, though only five variables were compared.

Table 3 shows the analysis of digit preference, assuming a uniform distribution of last digits. In the diet trial, all of the χ^2 values were highly significant, indicating that all the variables showed strong digit preference, although some preference is not unexpected. Digit preference was also evident for the results of a laboratory cholesterol test, which is unexpected since human estimation of the results is not usual. Measurements of height were not supplied for the diet trial (they were derivable from body mass index and weight for means, but this is not relevant for digit preference). In the drug trial, the χ^2 value was highly significant for height (indicating strong digit preference as might be expected) but not for any of the other measures Blood pressure measurement used a random zero machine, intended to remove digit preference. Table 4 shows the results of χ^2 testing for a difference in the pattern of digit preference between the two groups created by randomisation. This allows for the fact that digit preference can occur, but this should show a similar pattern in each of the randomised groups. In the diet trial, the final digit distributions are significantly different between the intervention group and the control group at baseline for all variables apart from cholesterol, fasting blood glucose, caffeine, carotene, and vitamin A. In the drug trial, the two randomised groups are far from being significantly different in terms of the final digit.

Discussion

The data from the diet trial have various anomalous statistical features that are not present in the data from the drug trial. These features are differences in means, and, even more noticeable, in variances at baseline and in differences in pattern of digit preference between randomised groups.

Magnitude of P values

These differences in the means and variances between baseline variables in the diet trial indicate that the two groups simply cannot have been formed as a result of random allocation as the authors claim. The magnitude of the P values derived from *t* tests of these differences for several variables is not compatible with a chance effect. One or two variables might show a small effect, but several of these P values are extreme.

Table 2 Baseline comparison of the two intervention groups, diet trial and drug trial

	Diet trial				Drug trial			
	Levene's F test for equality of variances		*t* test for equality of means		Levene's F test for equality of variances		*t* test for equality of means	
	F	Significance	*t*	Significance (two tailed)	F	Significance	*t*	Significance (two tailed)
Height	71.15	$1.4x10^{-16}$	–0.508	0.612	0.054	0.82	0.82	0.411
Weight	0.204	0.652	0.284	0.776	2.46	0.12	–0.227	0.82
Systolic blood pressure	4.81	0.029	1.89	0.06	2.45	0.12	0.206	0.84
Diastolic blood pressure	4.366	0.037	–0.27	0.788	0.89	0.35	–0.679	0.497
Cholesterol	28.77	$1x10^{-7}$	1.19	0.235	0.27	0.61	–1.22	0.22
Fasting blood glucose	8.21	0.004	–0.57	0.566	—	—	—	—
Total cholesterol	0.043	0.835	–0.35	0.729	—	—	—	—
Triglycerides	21.98	$3x10^{-6}$	0.484	0.628	—	—	—	—
Energy	0.98	0.322	–1.57	0.118	—	—	—	—
Total carbohydrate	1.97	0.161	0.236	0.814	—	—	—	—
Complex carbohydrate	12.86	0.0004	14.8	$6x10^{-44}$	—	—	—	—
Protein	15.18	0.0002	5.02	$6x10^{-7}$	—	—	—	—
Fat	20.5	$7x10^{-6}$	–2.88	0.004	—	—	—	—
Saturated	15.2	0.0001	3.9	0.0002	—	—	—	—
Fibre	94.23	$4x10^{-21}$	–8.47	$2x10^{-16}$	—	—	—	—
Soluble fibre	10.13	0.002	–6.95	$7x10^{-12}$	—	—	—	—
Caffeine	2.41	0.121	0.957	0.339	—	—	—	—
Salt	39.72	$5x10^{-10}$	–.377	0.706	—	—	—	—
Vitamin C	0.007	0.931	–5.6	$3x10^{-8}$	—	—	—	—
Carotene	51.06	$2x10^{-12}$	29.8	$2x10^{-133}$	—	—	—	—
Vitamin E	25.7	$5x10^{-7}$	5.9	$5x10^{-9}$	—	—	—	—
Vitamin A	51.42	$2x10^{-12}$	4.49	$8x10^{-6}$	—	—	—	—

Similarly, the significant difference in the pattern of digit preference between the randomised groups provides additional evidence that this is not a truly randomised trial.

Randomisation process

If this is not a randomised trial then how did these data arise? One possibility is that the data themselves are genuine but that the randomisation process has been subverted. This might explain, for example, some of the differences between the means of the variables at baseline. Had there been subversion of the randomisation process, in order for example to create differences between the groups at baseline, then smaller differences would have occurred and would also have been more consistent between the variables that are medically related—such as the different meas-

Table 3 χ^2 value (with P value) for the final digit at baseline, diet trial and drug trial

	Diet trial*		Drug trial	
	Intervention	Control	Intervention	Control
Height	—	—	239 ($1.8x10^{-46}$)	251 ($7.2x10^{-49}$)
Weight	128 ($4x10^{-23}$)	23 (0.00655)	7.3 (0.60)	6.5 (0.69)
Systolic blood pressure	1796 (U)	1470 (U)	7.6 (0.58)	9.1 (0.43)
Diastolic blood pressure	763 ($2x10^{-158}$)	820 ($1x10^{-170}$)	8.1 (0.52)	13.8 (0.13)
Cholesterol	554 ($2x10^{-113}$)	430 ($6x10^{-87}$)	16.23 (0.062)	5.76 (0.76)
Fasting blood glucose	478 ($4x10^{-97}$)	538 ($5x10^{-110}$)	—	—
Total cholesterol	1053 ($6x10^{-221}$)	1522 (U)	—	—
Triglycerides	642 ($2x10^{-132}$)	963 ($2x10^{-201}$)	—	—
Energy	2151 (U)	2630 (U)	—	—
Total carbohydrates	207 ($1x10^{-39}$)	927 ($7x10^{-194}$)	—	—
Complex carbohydrates	231 ($1x10^{-44}$)	939 ($3x10^{-196}$)	—	—
Protein	54 ($2x10^{-8}$)	251 ($5x10^{-49}$)	—	—
Fat	229 ($2x10^{-44}$)	437 ($2x10^{-88}$)	—	—
Saturated	123 ($4x10^{-22}$)	98 ($4x10^{-17}$)	—	—
Fibre	263 ($2x10^{-51}$)	1127 ($9x10^{-237}$)	—	—
Soluble fibre	273 ($1x10^{-53}$)	1086 ($6x10^{-228}$)	—	—
Caffeine	613 ($3x10^{-126}$)	694 ($1x10^{-143}$)	—	—
Salt	288 ($9x10^{-57}$)	301 ($2x10^{-59}$)	—	—
Vitamin C	304 ($5x10^{-60}$)	411 ($6x10^{-83}$)	—	—
Carotene	1470 (U)	1156 ($5x10^{-243}$)	—	—
Vitamin E	118 ($3x10^{-21}$)	101 ($8x10^{-18}$)	—	—
Vitamin A	705 ($6x10^{-146}$)	799 ($3x10^{-166}$)	—	—

The χ^2 value has 9 degrees of freedom.
* U means that the P value is too small for calculation.

Table 4 χ^2 value (with P value) for the final digit at the baseline in the diet and drug trials between the two randomised groups

	Diet trial		Drug trial	
	χ^2test (P value)	df	χ^2 test (P value)	df
Height	—	—	5 (0.83)	9
Weight	36 (3×10^{-5})	9	10 (0.31)	9
Systolic blood pressure	26 (0.00019)	6	7 (0.69)	9
Diastolic blood pressure	16 (0.046)	8	10 (0.38)	9
Cholesterol	13 (0.182)	9	7 (0.60)	9
Fasting blood glucose	12 (0.2)	9	—	
Total cholesterol	46 (5×10^{-7})	9	—	
Triglycerides	48 (3×10^{-7})	9	—	
Energy	16 (0.064)	9	—	
Total carbohydrate	154 (2×10^{-28})	9	—	
Complex carbohydrate	135 (1.4×10^{-24})	9	—	
Protein	43 (2×10^{-6})	9	—	
Fat	40 (6.4×10^{-6})	9	—	
Saturated	15 (0.08)	9	—	
Fibre	157 (8×10^{30})	8	—	
Soluble fibre	175 (6.5×10^{33})	9	—	
Caffeine	15 (0.059)	8	—	
Salt	28.5 (0.001)	9	—	
Vitamin C	18 (0.03)	9	—	
Carotene	10 (0.266)	8	—	
Vitamin E	20 (0.017)	9	—	
Vitamin A	9.5 (0.4)	9	—	

The degrees of freedom are less than 9 when one or more digits do not appear.

ures of cholesterol that show entirely different patterns between the groups. As it is, some are extreme and others are no different between the groups. What is more difficult to explain on the basis of subversion of the randomisation is the difference in the variability at baseline. Here we have highly significant differences in some variables both for the variances and the means, whereas for height, complex cholesterol, and triglyceride, there are highly significant differences in the variances but not in the means. Had there been a tendency to put patients with, say, higher blood pressures into one group, then we might have found significant differences in the mean values but with no difference in variance. However, we did not find this. Furthermore, no clear differences were apparent in the means for variables that would be readily available to a physician or health professional at the time of recruitment.

What is already known on this topic

Data fabrication is a rare form of scientific misconduct in clinical trials, but when it does occur it has serious consequences

Most papers are published without their data being independently verified, and there have been calls for data to be made available for scrutiny

Statistical methods for the detection of misconduct have been described, but few examples of their application have been published

It has been stated that statistical methods alone cannot prove data fabrication

What this study adds

Statistical methods can be applied to detect large scale fabrication of data in a randomised trial where data are available

Certain patterns of data are incompatible with randomisation, especially when a trial is "blind"

This paper shows the fabrication or falsification of data in a particular trial

Digit preference

Digit preference in itself is not evidence of misconduct. It is conceivable that the different patterns of digit preference between the two randomised groups may have arisen had one person recorded data for the treatment group and another recorded data for the control group. However, it is claimed that the trial was single blind, meaning that those recording data should not know to which group patients had been allocated. We would not expect differences therefore in digit preference between the randomised groups. But perhaps the trial was not single blind as described, and those recording the data were separated into groups according to whether they were dealing with patients allocated to either treatment or control. This could lead to differences in digit preference between randomised groups for variables where a human element of judgment was required. This would still not explain the differences in means and variances between the two groups since the effect of digit preference on the means and variances would only be slight. The combination of the differences in means, variances, and digit preference between the randomised groups is strong evidence that data fabrication took place in the diet trial.

Conclusion

We conclude that the data from the diet trial were either fabricated or falsified and that the strength of the evidence is such that appropriate steps should be taken to deal with this matter.

We thank Tom Meade who, on behalf of the Medical Research Council, provided the data for the drug trial and Richard Smith for his encouragement to examine further the data from the diet trial. The *BMJ* provided the data from the diet trial, which were supplied by the original author for further investigation of these data.

Contributors: SE and SAM had the ideas for the analysis, and SAM, SE, TM, and IR all contributed to the planning, conduct, and writing of the paper. SAM planned and carried out the statistical analyses. SAM and SE are jointly responsible for the overall content as guarantors. There are no other contributors.

Competing interests: None declared.

Funding: None.

1 Armitage P, Berry G. *Statistical methods in medical research*. 3rd ed. Oxford: Blackwell Scientific, 1994:386-401.
2 Altman DG. *Practical statistics for medical research*. London : Chapman and Hall, 1991:122-143.
3 Buyse M, George SL, Evans S, Geller NL, Ranstam J, Scherrer B, et al. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Stat Med* 1999;18:3435-51.
4 Taylor RN, McEntegart DJ, Stillman EC. Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. *Drug Inform J* 2002;36:115-25.
5 Evans S. Statistical aspects of the detection of fraud. In: Lock S, Wells F, Farthing M, eds. *Fraud and misconduct in medical research*. 3rd ed. London: BMJ Publishing Group, 2001:186-204.
6 Medical Research Council Working Party. MRC trial of treatment of mild hypertension: principal result. *BMJ* 1985;291:97-104.

(Accepted 15 July 2005)