BMJ

BMJ 2011;342:d561 doi: 10.1136/bmj.d561

RESEARCH METHODS & REPORTING

STATISTICS NOTES

Comparisons within randomised groups can be very misleading

J Martin Bland *professor of health statistics*[1], Douglas G Altman *professor of statistics in medicine*[2]

[1]Department of Health Sciences, University of York, York YO10 5DD; [2]Centre for Statistics in Medicine, University of Oxford, Oxford OX2 6UD

When we randomise trial participants into two or more intervention groups, we do this to remove bias; the groups will, on average, be comparable in every respect except the treatment which they receive. Provided the trial is well conducted, without other sources of bias, any difference in the outcome of the groups can then reasonably be attributed to the different interventions received. In a previous note we discussed the analysis of those trials in which the primary outcome measure is also measured at baseline. We discussed several valid analyses, observing that "analysis of covariance" (a regression method) is the method of choice.[1]

Rather than comparing the randomised groups directly, however, researchers sometimes look at the change in the measurement between baseline and the end of the trial; they test whether there was a significant change from baseline, separately in each randomised group. They may then report that this difference is significant in one group but not in the other, and conclude that this is evidence that the groups, and hence the treatments, are different. One such example was a recent trial in which participants were randomised to receive either an "anti-ageing" cream or the vehicle as a placebo.[2] A wrinkle score was recorded at baseline and after six months. The authors gave the results of significance tests comparing the score with baseline for each group separately, reporting the active treatment group to have a significant difference (P=0.013) and the vehicle group not (P=0.11). Their interpretation was that the cosmetic cream resulted in significant clinical improvement in facial wrinkles. But we cannot validly draw this conclusion, because the lack of a significant difference in the vehicle group does not provide good evidence that the anti-ageing product is superior.[3]

The essential feature of a randomised trial is the comparison *between* groups. Within group analyses do not address a meaningful question: the question is not whether there is a change from baseline, but whether any change is greater in one group than the other. It is not possible to draw valid inferences by comparing P values. In particular, there is an inflated risk of a false positive result, which we shall illustrate with a simulation.

The table shows simulated data for a randomised trial with two groups of 30 participants. Data were drawn from the same population, so there is no systematic difference between the two groups. The true baseline measurements had a mean of 10.0 with standard deviation (SD) 2.0, and the outcome measurement was equal to the baseline plus an increase of 0.5 and a random element with SD 1.0. The difference between mean outcomes is 0.22 (95% confidence interval –0.75 to 0.34, P=0.5), adjusting for the baseline by analysis of covariance.[1] The difference is not statistically significant, which is not surprising because we know that the null hypothesis of no difference in the population is true. If we compare baseline with outcome for each group using a paired *t* test, however, for group A the difference is statistically significant, P=0.03, for group B it is not significant, P = 0.2. These results are quite similar to those of the anti-ageing cream trial.[2]

We would not wish to draw any conclusions from one simulation. In 1000 runs, the difference between groups had P<0.05 in the analysis of covariance 47 times, or for 4.7% of samples, very close to the 5% we expect. Of the 2000 comparisons between baseline and outcome, 1500 (75%) had P<0.05. In this simulation, where there is no difference whatsoever between the two "treatments," the probability of a significant difference in one group but not the other was 38%, not 5%. Hence a significant difference in one group but not the other is not good evidence of a significant difference between the groups. Even when there is a clear benefit of one treatment over the other, separate P values are not the way to analyse such studies.[4]

How many pairs of tests will have one significant and one non-significant difference depends on the size of the change from baseline to final measurement. If the population difference from baseline is very large, nearly all the within group tests will be significant, and if the population difference is small, nearly all tests will be not significant, so there will be few samples with only one significant difference. If the difference is such that half the samples would show a significant change from baseline, as it would be in our simulation if the underlying

Correspondence to: Professor M Bland martin.bland@york.ac.uk

difference were 0.37 rather than 0.5, we would expect 50% of samples to have just one significant difference.

The anti-ageing trial is not the only one where we have seen this misleading approach applied to randomised trial data.[3] We even found it once in the *BMJ*![5]

Contributors: JMB and DGA jointly wrote and agreed the text, JMB did the statistical analysis.

Competing interests: All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous 3 years; no other relationships or activities that could appear to have influenced the submitted work.

1 Vickers AJ, Altman DG. Analysing controlled trials with baseline and follow-up measurements. *BMJ* 2001;323:1123-4.
2 Watson REB, Ogden S, Cotterell LF, Bowden JJ, Bastrilles JY, Long SP, et al. A cosmetic 'anti-ageing' product improves photoaged skin: a double-blind, randomized controlled trial. *Br J Dermatol* 2009;161:419-26.
3 Bland JM. Evidence for an 'anti-ageing' product may not be so clear as it appears. *Br J Dermatol* 2009;161:1207-8.
4 Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ* 2003;326:219.
5 Bland JM, Altman DG. Informed consent. *BMJ* 1993;306:928.

Cite this as: *BMJ* 2011;342:d561

Table

Table 1| Simulated data from a randomised trial comparing two groups of 30 patients, with a true change from baseline but no difference between groups (sorted by baseline values within each group)

	Group A				Group B		
	Baseline	6 months	Change		Baseline	6 months	Change
1	6.4	7.1	0.7	1	6.8	7.9	1.1
2	6.6	5.6	-1.0	2	7.2	7.5	0.3
3	7.3	8.3	1.0	3	7.2	6.9	-0.3
4	7.7	9.1	1.4	4	7.4	6.9	-0.5
5	7.7	9.5	1.8	5	7.5	8.3	0.8
6	7.9	9.6	1.7	6	7.5	9.4	1.9
7	8.0	8.5	0.5	7	8.3	9.0	0.7
8	8.0	8.5	0.5	8	8.4	8.8	0.4
9	8.1	9.1	1.0	9	8.7	8.0	-0.7
10	9.2	9.6	0.4	10	9.0	7.2	-1.8
11	9.3	8.7	-0.6	11	9.2	7.1	-2.1
12	9.6	10.7	1.1	12	9.6	10.6	1.0
13	9.7	9.0	-0.7	13	9.9	11.0	1.1
14	9.8	9.0	-0.8	14	10.1	11.5	1.4
15	9.8	8.0	-1.8	15	10.2	10.4	0.2
16	10.2	11.1	0.9	16	10.3	11.0	0.7
17	10.3	11.5	1.2	17	10.4	9.9	-0.5
18	10.6	9.1	-1.5	18	10.5	11.3	0.8
19	10.6	12.0	1.4	19	10.7	9.9	-0.8
20	10.7	13.2	2.5	20	10.8	10.7	-0.1
21	10.9	9.7	-1.2	21	10.8	11.8	1.0
22	11.1	12.2	1.1	22	11.1	10.0	-1.1
23	11.2	10.8	-0.4	23	11.1	13.2	2.1
24	11.8	11.9	0.1	24	11.4	11.8	0.4
25	12.3	12.2	-0.1	25	11.6	12.1	0.5
26	12.4	12.6	0.2	26	11.7	11.5	-0.2
27	13.1	15.0	1.9	27	12.0	12.7	0.7
28	13.2	13.8	0.6	28	12.3	13.7	1.4
29	13.3	14.1	0.8	29	13.7	12.6	-1.1
30	13.7	14.2	0.5	30	13.9	13.7	-0.2
Mean	10.02	10.46	0.44	Mean	9.98	10.21	0.24
SD	2.06	2.29	1.06	SD	1.90	2.09	1.02