Comparisons within randomised groups

Martin Bland
Prof. of Health Statistics
University of York

Douglas G Altman
Centre for Statistics in Medicine
University of Oxford

Presented to the Department of Health Sciences, June 2011.

Testing within randomised groups

When we randomise trial participants into two groups, we do this so that they comparable in every respect except the treatment which we then apply. However, rather than comparing the randomised groups directly, researchers sometimes look at the difference between a baseline measurement and the outcome measurement and test the equality between them, separately in each randomised group. They then report that in one group this difference is significant but not in the other and conclude that this is evidence that the groups, and hence the treatments, are different.

An example: Boots "anti-aging" cream trial

For example, a recent trial received wide media publicity as the first "anti-aging" cream "proven" to work in a randomised controlled clinical trial (Watson *et al.,* 2009). Participants were randomised into two groups, to receive the "anti-ageing" product or the vehicle as a placebo.

The authors report four outcome measures: fine lines and wrinkles, dyspigmentation, overall clinical grade of photoageing, and tactile roughness, each measured on a scale of 0 to 8 at baseline, 1, 6, and 12 months. The data recorded at one month are not mentioned in the paper.

The authors report four outcome measures: fine lines and wrinkles, dyspigmentation, overall clinical grade of photoageing, and tactile roughness, each measured on a scale of 0 to 8 at baseline, 1, 6, and 12 months. They report that "linear regression analysis was used to extrapolate the vehicle response to 12 months, thus allowing direct comparison with the test product", but no details are given. There is a reference (Chakrapani 1994), but this does not mention regression or extrapolation. It would be prudent to concentrate on the six months results.

There was no mention that any of the four measures was a prespecified primary outcome. We might surmise that a significant difference in any variable would be taken to indicate evidence of a treatment effect. The trial was entirely analysed in terms of P values, so prudence should lead us to adjust for multiple testing. We can apply a Bonferroni correction and multiply any P values by 4. If we were to include the 6 and 12 months results in the same analysis, we would multiply by 8.

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). This was interpreted as the cosmetic "anti-ageing" product resulted in significant clinical improvement in facial wrinkles. But we cannot draw this conclusion, because the lack of a significant difference in the vehicle group does not mean that subjects given this treatment do not improve, nor that they do not improve as well as those given the "anti-aging" product. It is the sizes of the differences which is important, they should be compared directly in a two sample test.

The paper includes some data for the improvement in each group, 43% for the active group and 22% for controls. This was what was picked up by the media. No P value is given, but in the discussion the authors acknowledge that this difference was not significant. No confidence interval is given, either.

There is an immediate problem if we try to calculate either P value or confidence interval for ourselves. We cannot reproduce 22%. 6/30 = 20%, 7/30 = 23%, not 22%. So there must have been loss to follow-up. How many people were there at six months?

Watson *et al.* should have followed the CONSORT guidelines. This would have enabled us to see how many participants were present at each stage and would also have required a confidence interval for the difference to be presented.

The nearest I can get is 6 improvements out of 27, with three dropouts, which gives 22.2%. If we take 13/30 and compare this to any of 7/30, 6/30, or 6/27, we do not get a significant chi-squared test, the P values being 0.10, and 0.052, 0.09. One of them almost makes <0.05. However, if we apply the Bonferroni correction, the P value for wrinkles becomes 0.208, clearly not significant.

The authors state that the 12-month clinical assessment data, presumably after the extrapolation, were analysed using a combination of Wilcoxon's matched pairs signed rank and rank sum tests, to give an overall P-value. It is not clear what this means, but they state that for facial wrinkles there was a statistically significant difference between the groups (test product, 70% of subjects improving compared with vehicle, 33% improving; combined Wilcoxon rank tests, P = 0.026). This P value would not stand up to the Bonferroni correction, 4×0.026 = 0.104.

The *British Journal of Dermatology* published my letter (Bland 2009) and a reply by Watson and Griffiths (2009). A different version subsequently appeared in *Significance* (Bland 2009b). This happened, of course, only because the publicity generated by Boots brought the paper to my attention.

The "anti-aging" skin cream trial made me think about this method of analysis. I have come across this several times before. Could I present a clearer explanation for why it is wrong?

An earlier example: information and anxiety

I used this example in my textbook, *An Introduction to Medical Statistics* (Bland 2000). Kerrigan *et al.* (1993) assessed the effects of different levels of information on anxiety in patients due to undergo surgery. They randomized patients to receive either simple or detailed information about the procedure and its risks. Anxiety was again measured after patients had been given the information.

Kerrigan *et al.* calculated significance tests for the mean change in anxiety score for each group separately. In the group given detailed information the mean change in anxiety was not significant (P=0.2), interpreted incorrectly as "no change". In the group given simple information the reduction in anxiety was significant (P=0.01). They concluded that there was a difference between the two groups because the change was significant in one group but not in the other.

This is incorrect. There may, for example, be a difference in one group which just fails to reach the (arbitrary) significance level and a difference in the other which just exceeds it, the differences in the two groups being similar. We should compare the two groups directly.

An alternative analysis tested the null hypothesis that after adjustment for initial anxiety score the mean anxiety scores are the same in patients given simple and detailed information. This showed a significantly higher mean score in the detailed information group (Bland and Altman 1993). In this example, the conclusion from the incorrect and correct analysis would be the same.

An earlier example: ultrasonography screening

Calculating a confidence interval for each group separately is essentially the same error as testing within each group separately. Bland (2000) gave this example. Salvesen *et al.* (1992) reported follow-up of two randomized controlled trials of routine ultrasonography screening during pregnancy. At ages 8 to 9 years, children of women who had taken part in these trials were followed up. A subgroup of children underwent specific tests for dyslexia.

The test results classified 21 of the 309 screened children (7%, 95% confidence interval 3% to 10%) and 26 of the 294 controls (9%, 95% confidence interval 4% to 12%) as dyslexic.

Much more useful would be a confidence interval for the difference between prevalences (–6.3 to +2.2 percentage points) or their ratio (0.44 to 1.34), because we could then compare the groups directly.

A simple simulation

We shall illustrate the inappropriateness of testing within separate groups with a simulation. Table 1 shows simulated data from a randomised trial, two groups of 30 drawn from the same population, so that there is no systematic difference between the groups. There is a baseline measurement, with standard deviation 2.0, and an outcome measurement, equal to the baseline plus an increase of 0.5 and a random element with standard deviation 1.0.

The usual way to analyse such data is to compare the mean outcome between the groups using the two sample t method or, better, to adjust the difference for the baseline measure using analysis of covariance or multiple regression. For this table, using the two sample t method, we get difference in mean differences = 0.20, P=0.27, adjusting the difference for the baseline measure using analysis of covariance we get difference = 0.20, P = 0.45).

The difference is not statistically significant, which is not surprising because we know that the null hypothesis is true, there is no difference in the population.

There are other analyses which we could carry out on the data. For each group, we can compare baseline with outcome using a paired t test. For group A, the difference is 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. We know that these data were simulated with an increase of 0.5 from baseline to outcome, so the significant difference is not surprising. There are only 30 in a group and the power to detect the difference is not great. Only 75% of samples are expected to produce a significant difference, so the non-significant difference is not surprising either.

A bigger simulation

We would not wish to draw any conclusions from one simulation. We repeated it 1000 times. 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. For the 2000 comparisons between baseline and outcome, 1500 had P<0.05, 75%, corresponding to the 75% power noted above. Of the 1000 pairs of t tests for groups A and B, 62 pairs had neither test significant, 562 had both tests significant, and 376 had one test significant but not the other.

So in this simulation, where is no difference whatsoever between the two "treatments", 37.6% of runs produced a significant difference in one group but not the other.

Hence we cannot interpret a significant difference in one group but not the other as a significant difference between the groups.

How many pairs of tests would be expected to have one significant and one not significant difference?

How many pairs of tests will have one significant and one significant difference depends on the power of the paired tests. If the population difference is very large, nearly all 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.

Looking at the problem more mathematically, if there is no difference between groups and power of the paired t test to detect the difference between baseline and outcome is P, the probablility that the first group will have a significant paired test is P, the probability that the second will be not significant is $1 - P$ and the probablity that both will happen is thus $P \times (1 - P)$. Similarly, the probability that the first will be not significant and second significant will be $(1 - P) \times P$, i.e. the same, so the probablity that one difference will be significant and the other not will be twice this, $2P \times (1 - P)$. It will not be 0.05.

When the difference in the population between baseline and outcome is zero, the probability that a group will have a significant difference is 0.05, because the null hypothesis is true. The probability that one group will have a significant difference and the other will not is then $2P \times (1 - P) = 2 \times 0.05 \times (1 - 0.05) = 0.095$, not 0.05. We would expect 9.5% of samples to have one and only one significant difference. If the power is 50%, as it would be here if the underlying difference were 0.37 rather than 0.50, as in our simulation, then $2P \times (1 - P) = 2 \times 0.50 \times (1 - 0.50) = 0.50$. We would expect 50% of samples to have one and only one significant difference.

A few more from the vaults

The next two examples are taken from *Statistical Questions in Evidence-based Medicine* (Bland and Peacock, 2000).

In a randomized trial of morphine vs. placebo for the anaesthesia of mechanically ventilated pre-term babies, it was reported that morphine-treated babies showed a significant reduction in adrenaline concentrations during the first 24 hours (median change −0.4 nmol/L, P<0.001), which was not seen in the placebo group (median change 0.2 nmol/L, P<0.79) (Quinn *et al.* 1993).

In a study of treatments for menorrhagia during menstruation, 76 women were randomized to one of three drugs (Bonnar and Sheppard 1996). The effects of the drugs were measured within the subjects by comparing three control menstrual cycles and three treatment menstrual cycles in each woman. The women were given no treatment during the control cycles. In each subject the control cycles were the three cycles preceding the treatment cycles.

The authors reported that patients treated with ethamsylate used the same number of sanitary towels as in the control cycles. A significant reduction in the number of sanitary towels used was found in patients treated with mefenamic acid (P<0.05) and tranexamic acid (P<0.01) comparing the control periods with the treatment periods.

This made me wonder what happens when we do separate tests within three groups. Suppose there are no treatment differences and the power of the within-group test between outcome and baseline is P. The probability that all three tests will be significant = P^3. The probability that all three will be not significant = $(1–P^3)$. Then the probability at least one test will be significant and one not significant = $1 – P^3 – (1–P^3)$. If all the null hypotheses within the group are true, so that there are no changes from baseline, $P = 0.05$. Then $1 – P^3 – (1–P^3) = 0.14$. If the null hypotheses within the groups are not true and the power to detect the difference is $P = 0.5$, then $1 – P^3 – (1–P^3) = 0.75$. Alpha for this test can be 0.75 rather than 0.05. When the null hypothesis was true, three quarters of such trials would produce a significant difference.

The next example comes from *Practical Statistics for Medical Research* (Altman, 1991).

Patients with chronic renal failure undergoing dialysis were divided into two groups with low or with normal plasma heparin cofactor II (HCII) (Toulon *et al.* 1987). Five months later, the acute effects of haemodialysis were examined by comparing the ratio of HCII to protein in plasma before and after dialysis. The data were analysed by separate paired Wilcoxon tests in each group.

Toulon *et al.* published the data, which appear in Table 2, taken from Altman (1991). They analysed the data using two paired Wilcoxon tests. For the Low HCII group the before to after change was significant, P<0.01. For the normal HCII group the difference was not significant, P>0.05.

What should they have done? They could have done a two sample t test between groups on the ratio before dialysis minus ratio after. This gives $t = 0.16$, 22 d.f., P=0.88. The variability is not the same in the two groups, so they might have done a two sample rank-based test, the Mann Whiney U test. This gives $z = 0.89$, P = 0.37. So either way, the difference is not statistically significant.

Conclusions

- Separate paired tests against baseline is a frequent practice.
- It is highly misleading and invalid.
- Randomised groups should be compared directly by two-sample methods.

The core of this talk was published as number 57 in the Statistics Notes series in the *British Medical Journal*, Bland and Altman (2011).

Recommendations

Trialists should:

- compare randomised groups directly,
- produce estimates with confidence intervals rather than significance tests (Gardner and Altman, 1986),
- follow the CONSORT guidelines (CONSORT).

References

Altman DG. (1991) *Practical Statistics for Medical Research.* London, Chapman and Hall.

Bland M. (2000) *An Introduction to Medical Statistics*, Oxford: University Press.

Bland JM. (2009) Evidence for an 'anti-ageing' product may not be so clear as it appears. *British Journal of Dermatology* **161**, pp1207–1208.

Bland M. (2009b) Keep young and beautiful: evidence for an "anti-aging" product? *Significance* **6**, 182-183.

Bland JM, Altman DG. (1993) Informed consent. *British Medical Journal* 306, 928.

Bland JM, Altman DG. (2011) Comparisons within randomised groups can be very misleading. *BMJ* 2011; **342**: d561.

Bland M, Peacock J. (2000) *Statistical Questions in Evidence-based Medicine*. Oxford, University Press.

Bonnar, J and Sheppard, BL. (1996) Treatment of menorrhagia during menstruation: randomised controlled trial of ethamsylate, mefenamic acid, and tranexamic acid. *British Medical Journal* **313**: 579-82.

Chakrapani C. Meta Analysis 1: How to Combine Research Studies. Magazine of the PMRS, 1994. Available at http://www.chuckchakrapani.com/Articles/PDF/94020314.pdf [accessed 30 April 2009].

CONSORT. The CONSORT Statement. http://www.consort-statement.org/consort-statement/

Gardner MJ and Altman DG. (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal* **292**: 746-50.

Kerrigan DD, Thevasagayam RS, Woods TO, McWelch I, Thomas WEG, Shorthouse AJ, Dennison AR. (1993) Who's afraid of informed consent? *British Medical Journal* **306**: 298-300.

Quinn MW, Wild J, Dean HG, Hartley R, Rushforth JA, Puntis JW, Levene MI. (1993) Randomised double-blind controlled trial of effect of morphine on catecholamine concentrations in ventilated pre-term babies. *Lancet* **342**: 324-7.

Salvesen KA, Bakketeig LS, Eik-nes SH, Undheim JO, Okland O. (1992) Routine ultrasonography in utero and school performance at age 8-9 years. *Lancet* **339**: 85-89.

Toulon P, Jacquot C, Capron L, Frydman MO, Vignon D, Aiach M (1987) Antithrombin-III and heparin cofactor-II in patients with chronic-renal-failure undergoing regular hemodialysis. *Thrombosis and Haemostasis* **57**: 263-268.

Watson REB, Griffiths CEM. (2009) Evidence for an 'anti-ageing' product may not be so clear as it appears: reply from authors. *British Journal of Dermatology* **161**, 1208–1209.

Watson REB, Ogden S, Cotterell LF, Bowden JJ, Bastrilles JY, Long SP, Griffiths CEM. (2009) A cosmetic 'anti-ageing' product improves photoaged skin: a double-blind, randomized controlled trial. *British Journal of Dermatology* DOI 10.1111/j.1365-2133.2009.09216.x

Table 1. Simulated data from a randomised trial comparing two groups of 30, with no real difference.

Group A				Group B			
Baseline	**Outcome**	**Baseline**	**Outcome**	**Baseline**	**Outcome**	**Baseline**	**Outcome**
11.2	10.8	10.6	12.0	12.3	13.7	8.3	9.0
8.0	8.5	13.1	15.0	7.5	8.3	9.6	10.6
7.3	8.3	6.6	5.6	7.2	7.5	10.4	9.9
9.8	9.0	9.7	9.0	10.3	110	7.5	9.4
7.7	9.1	10.3	11.5	10.8	10.7	8.4	8.8
8.0	8.5	10.9	9.7	7.4	6.9	11.1	10.0
13.2	13.8	12.4	12.6	11.7	11.5	8.7	8.0
11.8	11.9	7.7	9.5	13.9	13.7	10.2	10.4
9.8	8.0	7.9	9.6	12.0	12.7	6.8	7.9
13.3	14.1	9.2	9.6	9.0	7.2	9.9	11.0
10.6	9.1	13.7	14.2	10.8	11.8	10.1	11.5
12.3	12.2	10.7	13.2	10.5	11.3	11.4	11.8
10.2	11.1	11.1	12.2	13.7	12.6	11.1	13.2
6.4	7.1	9.6	10.7	9.2	7.1	10.7	9.9
9.3	8.7	8.1	9.1	11.6	12.1	7.2	6.9

Table 2. HCII/protein ratio in two groups of patients (Toulon et al. 1987, reported by Altman 1991)

Group 1 (low HCII)		Group 2 (normal HCII)	
Before	After	Before	After
1.41	1.47	2.11	2.15
1.37	1.45	1.85	2.11
1.33	1.50	1.82	1.93
1.13	1.25	1.75	1.83
1.09	1.01	1.54	1.90
1.03	1.14	1.52	1.56
0.89	0.98	1.49	1.44
0.86	0.89	1.44	1.43
0.75	0.95	1.38	1.28
0.75	0.83	1.30	1.30
0.70	0.75	1.20	1.21
0.69	0.71	1.19	1.30