Statistics Notes

Analysing controlled trials with baseline and follow up measurements

Andrew J Vickers, Douglas G Altman

In many randomised trials researchers measure a continuous variable at baseline and again as an outcome assessed at follow up. Baseline measurements are common in trials of chronic conditions where researchers want to see whether a treatment can reduce pre-existing levels of pain, anxiety, hypertension, and the like.

Statistical comparisons in such trials can be made in several ways. Comparison of follow up (post-treatment) scores will give a result such as "at the end of the trial, mean pain scores were 15 mm (95% confidence interval 10 to 20 mm) lower in the treatment group." Alternatively a change score can be calculated by subtracting the follow up score from the baseline score, leading to a statement such as "pain reductions were 20 mm (16 to 24 mm) greater on treatment than control." If the average baseline scores are the same in each group the estimated treatment effect will be the same using these two simple approaches. If the treatment is effective the statistical significance of the treatment effect by the two methods will depend on the correlation between baseline and follow up scores. If the correlation is low using the change score will add variation and the follow up score is more likely to show a significant result. Conversely, if the correlation is high using only the follow up score will lose information and the change score is more likely to be significant. It is incorrect, however, to choose whichever analysis gives a more significant finding. The method of analysis should be specified in the trial protocol.

Some use change scores to take account of chance imbalances at baseline between the treatment groups. However, analysing change does not control for baseline imbalance because of regression to the mean[1][2]: baseline values are negatively correlated with change because patients with low scores at baseline generally improve more than those with high scores. A better approach is to use analysis of covariance (ANCOVA), which, despite its name, is a regression method.[3] In effect two parallel straight lines (linear regression) are obtained relating outcome score to baseline score in each group. They can be summarised as a single regression equation:

follow up score =

constant + $a \times$ baseline score + $b \times$ group

where a and b are estimated coefficients and group is a binary variable coded 1 for treatment and 0 for control. The coefficient b is the effect of interest—the estimated difference between the two treatment groups. In effect an analysis of covariance adjusts each patient's follow up score for his or her baseline score, but has the advantage of being unaffected by baseline differences. If, by chance, baseline scores are worse in the treatment group, the treatment effect will be underestimated by a follow up score analysis and overestimated by looking at change scores (because of regression to the mean). By contrast, analysis of covariance gives the same answer whether or not there is baseline imbalance.

As an illustration, Kleinhenz et al randomised 52 patients with shoulder pain to either true or sham acupuncture.[4] Patients were assessed before and after treatment using a 100 point rating scale of pain and function, with lower scores indicating poorer outcome. There was an imbalance between groups at baseline, with better scores in the acupuncture group (see table). Analysis of post-treatment scores is therefore biased. The authors analysed change scores, but as baseline and change scores are negatively correlated (about $r = -0.25$ within groups) this analysis underestimates the effect of acupuncture. From analysis of covariance we get:

follow up score =

$24 + 0.71 \times$ baseline score $+ 12.7 \times$ group

(see figure). The coefficient for group (b) has a useful interpretation: it is the difference between the mean change scores of each group. In the above example it can be interpreted as "pain and function score improved by an estimated 12.7 points more on average in the treatment group than in the control group." A 95% confidence interval and P value can also be calculated for b (see table).[5] The regression equation provides a means of prediction: a patient with a baseline score of 50, for example, would be predicted to have a follow up score of 72.2 on treatment and 59.5 on control.

An additional advantage of analysis of covariance is that it generally has greater statistical power to detect a treatment effect than the other methods.[6] For example, a trial with a correlation between baseline and follow

Integrative Medicine Service, Biostatistics Service, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA
Andrew J Vickers
assistant attending research methodologist

ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford OX3 7LF
Douglas G Altman
professor of statistics in medicine

Correspondence to: Dr Vickers
vickersa@mskcc.org

Results of trial of acupuncture for shoulder pain[4]

	Pain scores (mean and SD)		Difference between means (95% CI)	P value
	Placebo group (n=27)	Acupuncture group (n=25)		
Baseline	53.9 (14)	60.4 (12.3)	6.5	
Analysis				
Follow up	62.3 (17.9)	79.6 (17.1)	17.3 (7.5 to 27.1)	0.0008
Change score*	8.4 (14.6)	19.2 (16.1)	10.8 (2.3 to 19.4)	0.014
ANCOVA			12.7 (4.1 to 21.3)	0.005

*Analysis reported by authors.[4]
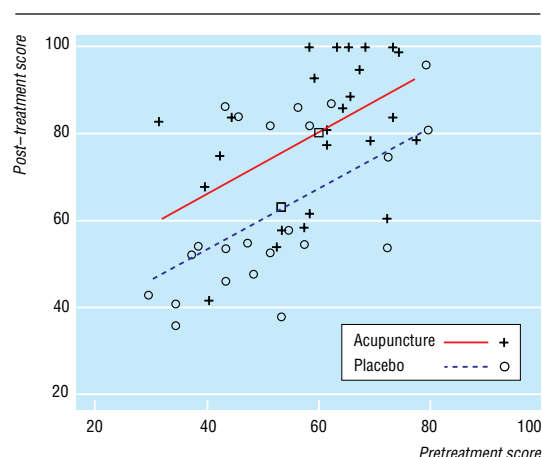
Pretreatment and post-treatment scores in each group showing fitted lines. Squares show mean values for the two groups. The estimated difference between the groups from analysis of covariance is the vertical distance between the two lines

situations, analysis of change scores can be a reasonable alternative, particularly if restricted randomisation is used to ensure baseline comparability between groups.[7] Analysis of covariance is the preferred general approach, however.

As with all analyses of continuous data, the use of analysis of covariance depends on some assumptions that need to be tested. In particular, data transformation, such as taking logarithms, may be indicated.[8] Lastly, analysis of covariance is a type of multiple regression and can be seen as a special type of adjusted analysis. The analysis can thus be expanded to include additional prognostic variables (not necessarily continuous), such as age and diagnostic group.

1 Bland JM, Altman DG. Regression towards the mean. *BMJ* 1994;308:1499.
2 Bland JM, Altman DG. Some examples of regression towards the mean. *BMJ* 1994;309:780.
3 Senn S. Baseline comparisons in randomized clinical trials. *Stat Med* 1991;10:1157-9.
4 Kleinhenz J, Streitberger K, Windeler J, Gussbacher A, Mavridis G, Martin E. Randomised clinical trial comparing the effects of acupuncture and a newly designed placebo needle in rotator cuff tendonitis. *Pain* 1999;83:235-41.
5 Altman DG, Gardner MJ. Regression and correlation. In: Altman DG, Machin D, Bryant TN, Gardner MJ, eds. *Statistics with confidence*. 2nd ed. London: BMJ Books, 2000:73-92.
6 Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC Med Res Methodol* 2001;1:16.
7 Altman DG, Bland JM. How to randomise. *BMJ* 1999;319:703-4.
8 Bland JM, Altman DG. The use of transformation when comparing two means. *BMJ* 1996;312:1153.

up scores of 0.6 that required 85 patients for analysis of follow up scores, would require 68 for a change score analysis but only 54 for analysis of covariance.

The efficiency gains of analysis of covariance compared with a change score are low when there is a high correlation (say r > 0.8) between baseline and follow up measurements. This will often be the case, particularly in stable chronic conditions such as obesity. In these

---

A memorable patient
Informed consent

I first met Ivy three years ago when she came for her 29th oesophageal dilatation. She was an 86 year old spinster, deaf without speech from childhood, and the only sign language she knew was thumbs up, which she would use for saying good morning or for showing happiness. She had no next of kin and had lived in a residential home for the past 50 years. She developed a benign oesophageal stricture in 1992 and came to the endoscopy unit for repeated dilatations. The carers in the residential home used to say that she enjoyed her "days out" at the endoscopy unit.

We would explain the procedure to her in sign language. She would use the thumbs up sign and make a cross on the dotted line on the consent form. She would enter the endoscopy room smiling, put her left arm out to be cannulated, turn to her left side for endoscopy, and when fully awake would show her thumbs up again. Every time after her dilatation the nursing staff would question why an expandable oesophageal stent was not being considered. We would conclude that the indications for an expandable stent in benign strictures are not well established.

Her need for dilatation was becoming more frequent, and so on her 46th dilatation we decided to refer her to our regional centre for the insertion of a stent. She had an expandable stent inserted, and in his report the endoscopist mentioned the risk of the stent migrating down in the stomach beyond the stricture. Six weeks later she developed a bolus obstruction. At endoscopy it was noted that the stent had indeed migrated down. She consented to another stent. Four weeks later she had another bolus obstruction that could not be completely removed at the first attempt,

and she was brought back the following day for removal of the bolus by endoscopy.

She came to the endoscopy room but did not have her familiar smile. She looked around for a minute, got off her trolley, and walked out. Everyone in the endoscopy room understood that she was trying to say, "I've had enough."

She did not come back for a repeat endoscopy, and she stayed nil by mouth on intravenous fluids. Two weeks later she died of an aspiration pneumonia. We think she understood all the procedures she had agreed to. We also think it was informed consent. I hope we were right. She gave us a very clear message without saying a word on her last visit to the endoscopy room.

Do we really understand what aphasic patients are trying to tell us when we get informed consent for invasive procedures? We should try to read the non-verbal messages very carefully.

I Tiwari *associate specialist in gastroenterology, Broomfield Hospital, Chelmsford*

We welcome articles of up to 600 words on topics such as *A memorable patient, A paper that changed my practice, My most unfortunate mistake*, or any other piece conveying instruction, pathos, or humour. If possible the article should be supplied on a disk. Permission is needed from the patient or a relative if an identifiable patient is referred to. We also welcome contributions for "Endpieces," consisting of quotations of up to 80 words (but most are considerably shorter) from any source, ancient or modern, which have appealed to the reader.