Should we cross off the crossover?

The simple two-treatment, two-period crossover trial is under attack again. In ideal circumstances this design provides an efficient way of comparing the efficacies of two treatments for the short-term alleviation of a chronic condition. Subjects receive the treatments during two periods of administration separated by a wash-out period intended to remove any possible carry-over effect. The two orders of administration, AB and BA, are assigned to subjects at random. Two features of the design account for its popularity. By enabling each subject to provide observations for each treatment, the trial economizes in the number of subjects required. Secondly, the treatments are compared *within* subjects, and the contrast is therefore affected by the random variation of repeated observations on the same subject, which is less than that of observations on different subjects.

These advantages have been long recognized, and simple crossover trials are widely used in a variety of clinical contexts. They are, for instance, common in Phase II studies conducted on behalf of the pharmaceutical industry, many of which remain unpublished. In the early days the correct forms of statistical analysis were not widely understood, and some practitioners no doubt still fail to take proper account of the special features of the design. However, appropriate methods have been described by, for instance, Grizzle (1965), Hills & Armitage (1979), Jones & Kenward (1989) and Everitt (1989). Clayton & Hills (1988) and Sheldon (1990) have recently described a useful graphical approach.

During the 1970s doubts emerged about the validity of the design, and resulted in some degree of discouragement from the US Food and Drug Administration. The point at issue, then and to some extent now, was the possibility that a carry-over effect existed, in spite of the wash-out period, but that it was undetected by the analysis of the data. A carry-over effect (or more strictly a *difference* between the carry-over effects of A and B) results in a different treatment effect in the two periods, i.e. it is one possible reason, amongst others, for a treatment-period (TP) interaction. Any such interaction produces a different mean level of response in the two groups of subjects receiving the different sequences, AB and BA. A straightforward test for TP could be carried out, but it would be affected by random variation *between* subjects and therefore relatively insensitive; in statistical jargon it would have low 'power'. One partial remedy is to make use of baseline observations made before the treatment periods: differences between the responses during treatment and the baseline values are *within* subjects, and may improve sensitivity.

Armitage & Hills (1982) took the view that 'a single crossover trial cannot provide the evidence for its own validity' and suggested that 'crossover trials should be regarded with some suspicion unless they are supported by evidence, from previous trials with similar patients, treatments and response variables, that the interaction is likely to be negligible'.

What procedure should be followed if the TP interaction is clearly demonstrated or even merely suspected of being present? The usual advice is to ignore the second period and analyse the first-period data as though it came from a parallel-groups trial. The second-period observations are less useful because they relate to two groups of subjects who, although separated some time back by random assignment, have in the interim encountered different treatments. This advice is usually sensible. There could, though, be situations where other strategies were preferable. Suppose, for instance, the TP interaction was due not to carry-over but to a radical change in the general level of response between periods, causing treatment effects also to change. In that case, an average treatment effect over the two periods, as provided by the standard crossover analysis, might be entirely appropriate. Less plausibly, perhaps, there might be good reason to believe that the treatments were largely ineffective until the second period; the second-period comparison would then be the best choice.

In his letter in this issue (p. 133), Dr Senn argues forcibly against the 'two-stage' analysis, by which one might first test for TP, and then use the usual crossover estimate of treatment effect if TP is not significant, but the first-period estimate if TP is significant. His point is similar to that of Freeman (1989); see also Senn (1988). If there is really no treatment effect or interaction (i.e. the null hypothesis is true), the two-stage procedure gives an extra opportunity to find a false positive. A misleadingly significant effect might arise either in the crossover analysis (with TP nonsignificant) or in the first-period analysis (with TP significant). The true probability of such a chance finding (a 'Type I' error) is therefore raised above the nominal level of the significance tests. Moreover, the test for interaction is not independent of that for a first-period effect: a high proportion of trials with a false-positive interaction will also show a false-positive first-period effect.

Senn (1991) concludes that 'the TS procedure must not on any account be used to analyse crossover trials', confirming Freeman's view that 'the [TS] analysis is so unsatisfactory as to be ruled out of future use'. These opinions seem unduly categorical. The first point, about the enhancement of the Type I error probability, is a common feature in statistical analyses which proceed in stages. Undue concern about significance levels should not be allowed to inhibit the pursuance of a sensible sequence of tests. The second point, about the non-independence of the interaction and first-period effects, is common to other factorial designs, yet it has not prevented a widely recommended practice of testing for interaction and, if this is significant, estimating the effect of one factor separately at different levels of another.

Berry (1990) draws an interesting distinction between the 'Type I' and 'Type II' approaches to statistical analyses. The former is mainly concerned to avoid false inferences when the null hypothesis is true, whereas the latter is concerned more with the detection of true

effects. A Type II statistician might be content with the TS procedure on the grounds that sufficiently large effects, with or without a carry-over, are likely to be detected and correctly described.

Nevertheless, the simple crossover is too fragile an instrument for sorting out all these problems in unfamiliar situations. There seems no escape from the previously expressed advice to use this design only when a good deal of background knowledge is available. Some statisticians (Freeman, 1989; Grieve, 1985; Racine *et al.*, 1986) advocate the use of Bayesian methods, whereby subjective judgements based on prior experience are expressed quantitatively and form an essential part of the analysis. Others will prefer to use prior experience in a less formal manner.

Finally, one should recall that crossover designs can be extended beyond two periods and can compare more than two treatments. One advantage of such extended designs is that carry-over effects can be estimated within subjects. Another is that distinctions can be made between

first-order carry-over effects, depending on the immediately preceding treatment, and those with longer lag periods. The excellent book by Jones & Kenward (1989) gives ample details. Even here, though, all is not plain sailing. The analysis of larger designs requires careful modelling of the effects of the various factors on the response. It raises, for instance, questions of the definition of a carry-over effect. Does this depend only on the identity of the previous treatment, or is its magnitude affected also by the current treatment, or indeed any other factors currently present? More research and more experience are needed before these large designs can be safely recommended for routine use.

PETER ARMITAGE

*Emeritus Professor of Applied Statistics,
University of Oxford,
71 High Street,
Drayton, Abingdon,
Oxon OX14 4JW*

References

- Armitage, P. & Hills, M. (1982). The two-period crossover trial. *Statistician*, **31**, 119–131.
- Berry, D. A. (1990). Correspondence: Subgroup analysis. *Biometrics*, **46**, 1227–1230.
- Clayton, D. & Hills, M. (1988). A two-period crossover trial. In *The statistical consultant in action*, eds Hand, D. & Everitt, B., pp 42–57. Cambridge: Cambridge University Press.
- Everitt, B. (1989). *Statistical methods for medical investigations*. London: Arnold.
- Freeman, P. R. (1989). The performance of the two-stage analysis of two-treatment, two-period crossover trials. *Statistics in Medicine*, **8**, 1421–1432.
- Grieve, A. P. (1985). A Bayesian analysis of the two-period crossover clinical trial. *Biometrics*, **42**, 593–600.
- Grizzle, J. E. (1965). The two-period change-over design and its use in clinical trials. *Biometrics*, **21**, 467–468.
- Hills, M. & Armitage, P. (1979). The two-period cross-over clinical trial. *Br. J. clin. Pharmacol.*, **8**, 7–20.
- Jones, B. & Kenward, M. G. (1989). *Design and analysis of cross-over trials*. London: Chapman & Hall.
- Racine, A., Grieve, A. P., Flühler, H. & Smith, A. F. M. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion). *Applied Statistics*, **35**, 93–150.
- Senn, S. J. (1988). Letter to the Editor: Cross-over trials, carry-over effects and the art of self-delusion. *Statistics in Medicine*, **7**, 1099–1101.
- Sheldon, T. A. (1990). A graphical presentation of the results of crossover trials. *Br. J. clin. Pharmacol.*, **30**, 345–349.
- Senn, S. J. (1991). Problems with two stage analysis of crossover trials. *Br. J. clin. Pharmacol.*, **32**, 133.