Problems with the two stage analysis of crossover trials

A recent paper by Sheldon (1990) provides the opportunity to review a statistical procedure which this journal did much to promote.

Sheldon presents a diagram which bears a close relationship to the two stage (TS) procedure for the analysis of AB/BA crossover trials. The TS procedure has been discussed by a number of authors, notably, Grizzle (1965) who is usually credited with introducing it, and Hills & Armitage (1979) who gave an extremely clear description of it in this journal, covering also general problems affecting crossover trials and the analysis of various types of data. The Hills & Armitage (1979) paper can still be recommended as an excellent introduction to crossover trials but requires one qualification.

The TS procedure requires that a preliminary test for carry-over be carried out. This test is roughly represented on the Zz dimension of Sheldon's graph. If the result of the test is not significant a within patient, or 'crossover' (CROS) test of the treatment effect is carried out. If it is significant it is recommended that the first period values should be compared using a between patient, or parallel (PAR) groups test.

Although this procedure has intuitive appeal it presents a number of difficulties. One, the lack of power of the test of carryover, it is only fair to point out, was recognised by Hills & Armitage (1979) themselves who showed some hesitancy in proposing the TS 'solution' to the problem without actually going as far as Senn (1988) who recommended that therefore the TS procedure should not be used at all. The second difficulty, which was uncovered by Freeman (1989) in an extremely important paper (whose notation, TS, CROS and PAR, this letter follows) is more serious. Since Sheldon (1990) cites Jones & Kenward (1989) but does not note their warnings regarding the TS procedure it is worth quoting what they have to say about Freeman's (1989) result (available to Jones & Kenward (1989) in an earlier unpublished form).

They write: 'However, as Freeman (1988) has shown, the problem with Grizzle's procedure does not just lie in the lack of power of the preliminary test. The problem lies more in the automatic use of the first-period data to provide a treatment comparison in the event of discovering a significant interaction . . . Since the test statistics for the carry-over and direct treatment are correlated, the test statistic for the treatment effect will not then be unbiased' (p 85). The TS procedure leads to a type I error rate which exceeds the nominal size

whether or not carryover is present.

Freeman's (1989) findings regarding the TS procedure are devastating and even Jones & Kenward (1989) underestimate their importance writing: 'We would conjecture that the situation in which Freeman's results are critical is rather limited . . . In the absence of a carry-over difference even the Grizzle procedure deviates only mildly from its nominal properties' (p 86). Presumably they are referring here to the fact that under such conditions the nominal type I error rate for the TS procedure lies typically between 0.07 and 0.095 as opposed to the 0.05 claimed for it but overlooking the fact that this overall error rate is a weighted average of two rates: an error rate of 0.05 for the 90% of cases in which the test of carryover does not show significance and therefore the CROS procedure is used to test for treatment and an error rate of 0.25 to 0.50 for the 10% of cases when the preliminary test does show significance and the PAR test is used instead. Thus the TS procedure gives results which are either identical to the straightforward CROS test or totally unreliable.

I would urge all who are interested in learning about the problems with the TS procedure to read Freeman (1989). An analogy with medicine may help to motivate readers to do so. The TS procedure is a form of statistical screening procedure. It is often assumed in medicine and statistics that screening can never do any harm. In practice, however, whether a programme of screening is harmful or beneficial depends on many things: the sensitivity and specificity of the screening procedure, the prevalence of the disease and the losses incurred by treating the well as if they were ill and the ill as if they were well. When examined in these terms the TS procedure is shown to be as misguided as the policy of treating every case of stomach-ache as an appendicitis.

Thus, the TS procedure must not on any account be used to analyse crossover trials. The implication for the diagrams which Sheldon (1990) presents is that they should be used, if at all, with extreme caution.

S. J. SENN

Medical Department, Pharmaceutical Division
CIBA-GEIGY Ltd, 4002 Basle, Switzerland

Received 29 November 1990
accepted 15 January 1991

ADONIS 030652519100153U

References

- Grizzle, J. (1965). The two-period change-over design and its use in clinical trials. *Biometrics*, **21**, 467-80.
- Freeman, P. (1989). The performance of the two-stage analysis of two-treatment, two-period crossover trials. *Stats. in Med.*, **8**, 1421-1432.
- Jones, B. & Kenward, M. (1989). *Design and analysis of crossover trials*. London: Chapman and Hall.
- Hills, M. & Armitage, P. (1979). The two paired cross-over trial. *Br. J. clin. Pharmac.*, **8**, 7-20.
- Senn, S. J. (1988). Cross-over trials, carry-over effects and the art of self-delusion. *Stats. in Med.*, **7**, 1099-1101.
- Sheldon, T. A. (1990). A graphical presentation of the results of crossover trials. *Br. J. clin. Pharmac.*, **30**, 345-349.