ELSEVIER

Journal of Statistical Planning and Inference 96 (2001) 29–40

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Cross-over trials in drug development: theory and practice

Stephen Senn *

Department of Epidemiology and Public Health, Department of Statistical Science, University College London, 1-19 Torrington Place, London WC1E 6BT, UK

Abstract

It is maintained that much research into the design and analysis of cross-over trials has been of little practical relevance to drug development. The point is illustrated using three topics: the AB/BA design, bioequivalence and multi-period designs in two treatments. It is suggested that statisticians should pay more attention to the work of fellow scientists, in particular, in the field of pharmacokinetics, and also that the philosophical–inferential base employed in examining cross-over trials has often been too narrow.

Keywords: Bioequivalence; Carry-over; Clinical trial; Multiperiod design; Pharmacokinetics; Pharmacodynamics

1. Introduction

Clinical cross-over trials are trials in which subjects are given sequences of treatments with the object of studying differences between individual treatments (Senn, 1993). They hold a perplexing position in pharmaceutical statistics. They appear to divide regulators and developers as regards their utility: the former are worried that they will give biased results if affected by carry-over (to be explained below), the latter are enthusiastic about their efficiency. They are often inappropriate in phase III, where the investigation of tolerability is usually as important as that of efficacy. On the other hand, in chronic disease, they are the designs of choice for investigating individual response to treatment. When it comes to analysis, they have been plagued by seemingly endless controversy and approaches that were previously regarded as appropriate are now regarded as misguided. On the one hand the challenge of finding 'optimal' designs appears to have been enthusiastically accepted by statisticians; on the other, the models employed have been criticised as inappropriate and pharmacologically naïve. Table 1 lists some areas of recent past or current controversy.

* Tel.: +44-0-171-391-1698; fax: +44-0-171-813-0280.
E-mail address: stephens@public-health.ucl.ac.uk (S. Senn).

0378-3758/01/$ - see front matter
PII: S0378-3758(00)00318-9

Table 1
Debatable issues concerning the use of cross-over trials in drug development

General area	Topic or controversial area
General	Appropriate role in drug development
The AB/BA design	Is pre-testing for carry-over acceptable?
	Can this design be used at all?
	The use of baselines
Bioequivalence	What are optimal tests and can they be used?
	Choice of hypotheses
	Choice of limits for relative bioavailability
	Appropriate size of tests
	Type of confidence limit to use
	Log transformation versus Fieller's theorem
	Individual bioequivalence
Multi-period designs	General utility of this approach
	Choice of models for carry-over
	Optimal or robust designs?
	Degrees of freedom for error
n-of-1 trials (trials with repeated random treatment allocation for a given individual)	Stage of drug development in which to use
	Predicting effects for individual patients

In this paper, I shall attempt to give an overview of some of these areas of controversy. I shall do this from the perspective of an applied statistician who has worked within the pharmaceutical industry. This will tend to stress practical aspects of these controversies at the expense of theoretical ones. Mathematical complexities will be avoided altogether but some philosophical issues of importance will be stressed.

Since an overall review of cross-over trials has recently been given elsewhere (Senn, 1998) and since, in any case, space here would not permit this to be repeated, specific details of analysis and models will be almost completely avoided since they are not necessary for the discussion which follows. Illustration of controversies in this field will be limited to three specific examples: first, the analysis of the AB/BA cross-over, second, the application of this design to the problem of bioequivalence and third, multi-period designs in two treatments. Multi-treatment designs will not be covered.

2. The AB/BA design

A very simple design in two periods and two treatments (A and B, say) is one whereby patients are allocated in equal numbers, or at the very least with equal probability, to two sequences of treatments, A followed by B ('AB') or B followed by A ('BA'). For continuous data a very simple analysis can be based on the within-patient differences. If the period effect is ignored these can be analysed using the matched-pairs t, an analysis which dates back to Student's famous paper (Student, 1908), and a suitable estimate of the treatment effect is the simple mean of these within-patient differences. A slightly more complex analysis is necessary if the period effect is to be eliminated. (Period effect here refers to any secular difference between periods of

measurement, generally affecting all patients or their measurements, and unrelated to treatment.) This analysis is based on an unweighted average (to be referred to subsequently as CROS) of a treatment estimate taken separately within each sequence. Hills and Armitage (1979) gave a very good description of this approach in a famous paper.

A common assumption in the analysis of experiments is that there is no interference between units. The units in a cross-over trial are episodes of treatment and a way in which this assumption can break down is if the treatments given during one episode can continue to affect the same patient in a subsequent episode. This sort of residual effect of previous treatments is referred to as 'carry-over' and can bias estimates, for where the trialist believes that (s)he is studying the effect of one treatment only, the effect of two or more may be being studied. It has been well understood, at least since Grizzle (1965), that if carry-over occurs, then unless (which is scarcely plausible) it is identical for each treatment, it will bias the estimate, CROS, described above.

Grizzle (1965) pointed out, however, that an unbiased estimate of the treatment effect is available by using the first period data only as in any parallel group trial. Thus, the estimate is the difference between the two first period means (to be referred to as PAR, hereafter). Of course, no one would design a cross-over trial intending to use this estimate. However, Grizzle also pointed out that a test of carry-over was available. The mean over both periods for any given patient, being not only a mean of the results for both periods but also for both treatments, must reflect the main effects both of period and treatment, irrespective of the *sequence* to which the patient was assigned. It thus follows that any *systematic*, as opposed to random, differences for this statistic between sequences can only be a reflection of the order in which the treatments were administered. Differences due to order will either reflect a treatment by period interaction or carry-over. Both of these, however, may be regarded as troubling the interpretation of the main effect of treatment. A standard two independent sample t-test, therefore, comparing the means over both periods between both sequences provides a method of diagnosing carry-over (albeit, aliased with period by treatment interaction). (The relevant contrast that this test uses will be referred to as CARRY hereafter.)

Grizzle proposed a two-stage procedure whereby a preliminary test was to be carried out on CARRY. Because CARRY was a between-patient contrast for a trial which would probably be designed to have adequate power for within-patient contrasts only, Grizzle recommended a test at a less stringent level of significance, say 10%, than usual. He then proposed that if this test were significant, PAR rather than CROS should be used to evaluate the effect of treatment. In the case of non-significance of CARRY, CROS might be used as originally intended.

This proposal was widely adopted within the pharmaceutical industry, and was even formally endorsed by a Statisticians in the Pharmaceutical Industry (PSI) working group. However, it has had little impact on physicians outside who, if they used any of the analyses described above, would often adopt the matched-pairs approach unadjusted for the period effect.

There were some dissenting voices. For example, Brown (1980), in an important investigation of the power of CARRY under likely circumstances, concluded that it

would give little comfort in practice. He still recommended, however, that Grizzle's two-stage procedure should be followed. Senn (1988), on the other hand, suggested that the procedure should be abandoned altogether since, if the object were to *prove* that carry-over had not taken place, the *null* hypothesis ought to be one of influential carry-over. The significance test employing CARRY ought to seek to reject this hypothesis in an analogous manner to bioequivalence trials (which will be covered in due course). Such rejection would in practice hardly ever be forthcoming. Hence, the grounds for using CROS had to be external to the data, implying that the two-stage procedure was pointless.

A more serious criticism was made by Freeman (1989). He pointed out that PAR was in fact only an unbiased estimator of the treatment effect if it was used in an unconditional sense. However, CARRY and PAR were strongly correlated and this implied that the conditional distribution of PAR, given that CARRY was significant was quite other than statisticians had implicitly supposed. The net result was that under the null hypotheses of no treatment effect and no carry-over, the type I error rate was not 5% as assumed (using the conventional levels) but somewhere between 7 and 9.5%.

Subsequently, it was pointed out that in many ways this understates the problem with the procedure. Only under the extreme version of the Neyman–Pearson philosophy is the average type I error rate of any relevance. It is composed of two parts, a rate of 5% with probability 0.9 (given that CARRY is not significant) associated with CROS and a rate lying between 25% and 50% for PAR with probability 0.1 (given that CARRY is significant) (Senn, 1994). However, in practice, an investigator will know which of the two tests (s)he has used. Thus, to take a more Fisherian view, the overall type I rate is irrelevant: either the two-stage procedure is completely superfluous or the type I error rate is 5–10 times what has been claimed for it (Grieve and Senn, 1998).

The next stage in the development of this on-going story has been to 'correct' the two-stage procedure so that it produces a test of correct size. Similar schemes have been proposed by Senn (1996) and Wang and Hung (1997) but in a rather different spirit. Senn (1996) points out that the power of the corrected procedure is poor compared to the simpler strategy of always using CROS and that the two-stage procedure in any case smacks so strongly of ad hocery that it should be ruled out on those grounds alone. On the other hand, Wang and Hung (1997) seem to find a role for it.

Whatever the eventual fate of the two-stage analysis, whether completely abandoned (as is my hope) or widely adopted in a modified form, its history hitherto has been nothing short of disastrous. It has, in retrospect, been a good thing that the recommendations of statisticians have not been adopted. Ironically, although there are now three books on cross-over trials (Jones and Kenward, 1989; Senn, 1993; Ratkowsky et al., 1993) none of which recommends this procedure, general textbooks on medical statistics continue to be written which endorse it.

However, it should be stressed, that although the two-stage analysis of the AB/BA design cannot be recommended, this does not mean that the design itself is not useful. On the contrary, it can be extremely useful provided that the assumption of no carry-over can be made.

3. Bioequivalence

Once the patent on an innovator drug expires, it is then possible for generic versions to be produced and marketed with a consequent beneficial on the price effect (for health care purchasers!). It also happens that an innovator company may wish to provide differing versions of the same product, for example a dispersible rather than a tablet form. It may also be the case that a treatment can be taken under different circumstances, say with or without a certain food, which may or may not interact with it.

In all these cases it seems intuitively unreasonable to require a full new drug development *ab initio*. For example, if a generic company had to repeat all the steps of the innovator company in developing the product, it would affect the price at which it could compete. It was early recognised that a potential pharmacological solution to this problem was to compare the concentration time curves in the blood of the innovator and generic products using a suitable number of subjects, often healthy volunteers, and nearly always using an AB/BA cross-over. "The blood is a gate through which the drug must pass". It seems implausible that formulations that could be equivalent at this point could subsequently differ in terms of distribution and hence in terms of pharmacodynamic effect. Thus, equivalence in concentration is believed to imply equivalence in effect and side effect.

The pharmacological solution brought with it, however, a statistical problem. Early practice seems to have been to compare concentration–time curves with a conventional significance test carried out on suitable summary measures, nearly always area under the curve (AUC) but sometimes also concentration maximum ($C_{\max}$) and time to reach $C_{\max}$ ($T_{\max}$). (AUC is a measure of the *bioavailability* of the product and is the most important of these measures.) At first, small trials were employed with no thought to power. Gradually, however, it came to be realised that failure to find a significant difference was not proof of equivalence. However, early solutions seem to have involved making a similar error to that with carry-over in the two-stage analysis, namely looking at it in terms of the power of the conventional test to find a difference, rather than evolving a new test altogether.

An important early paper was that of Westlake (1976) who made what was, in my opinion, the crucial observation, namely that the problem is essentially one of *estimating* the relative bioavailability and showing that this is close to 1. He proposed using a 95% confidence interval centred on 1. Kirkwood (1981) proposed instead a confidence interval centred on the observed ratio. As O'Quigley and Baudoin (1988) showed, a fiducial/Bayesian interpretation of the difference between the two approaches is that the first relates to the probability that the relative bioavailability lies within the region of equivalence whereas the second examines whether the most probable region includes the region of equivalence. There has been much technical discussion since of the appropriate approach and there is an enormous literature on this apparently simple problem, some areas of debate are covered in chapter 22 of *Statistical Issues in Drug Development* (Senn, 1997). The general approach now internationally agreed

by regulatory authorities is to analyse log-AUC and, having anti-logged, show that a 90% conventional confidence limit for the ratio of AUCs lies in the invariant range 0.8–1.25.

Consider the relative bioavailability of the 'test' (for example, generic) treatment compared to the 'reference' (for example, innovator) treatment. Suppose that if this relative bioavailability, ρ, is less than 80% it is referred to as 'sub-availability' and if it is more than 125% it is referred to as 'super-availability' and that anything in between is considered to constitute 'practical equivalence'. In a hypothesis-testing framework, rejection of the hypothesis of super-availability, H_{0A}: $\rho > 1.25$ *and* of the hypothesis of sub-availability, H_{0B}: $\rho < 0.8$ would seem to imply that practical equivalence obtains. It can be shown that carrying out two such one-sided tests independently, each at the 5% level, corresponds to a procedure which accepts bioequivalence if the 90% conventional two-sided confidence limits for the difference in log-AUC, $\delta = \log \rho$, lie between the limits of equivalence $\log 0.8 = -0.223$ and $\log 1.25 = 0.223$.

Now consider, for argument's sake, a bioequivalence trial in which the standard deviation of within-subject log-AUC were known to be σ. (Of course, in practice, this would never be the case but assuming that this *is* so will suffice to make a point.) Suppose that we have m subjects assigned to each sequence and that $n = 2m$. The standard error of the estimate of the difference in log-bioavailability will be $\sigma\sqrt{(2/n)}$. Let the point estimate of the difference in log-bioavailability (test-reference) be d. The conventional approach thus concludes bioequivalence if $d - 1.645 \times \sigma\sqrt{(2/n)} > -0.223$ and $d + 1.645 \times \sigma\sqrt{(2/n)} < +0.223$. A necessary (but not sufficient) condition for this to be so is that

$$1.645 \times \sigma\sqrt{(2/n)} < +0.223. \tag{1}$$

Now suppose that σ is large, then for sufficiently small n, condition (1) is not met. It thus follows that *whatever* the value of d, bioequivalence will not be concluded. However, the treatments might, in fact, be perfectly (or at the very least acceptably) bioequivalent. Hence under these circumstance the procedure has zero power and of course also zero size.

Although the above argument assumes a known value of σ it is intuitively clear that a similar phenomenon can arise where the value of σ has to be estimated and the t-distribution is used for calculating the confidence limits. (Because the estimate of σ can, with small probability, be small even where σ is large, a test of *zero* size will not result but one which is less than the nominal 5% is possible.) This suggests that the now common approach of seeing that the 90% confidence limits lie between the limits of equivalence is not optimal in the Neyman–Pearson sense and that, indeed even where σ is small compared to n, the test is in general conservative. As a consequence, various proposals have been made to improve this approach. An important paper, is that of Mehring (1993), which in addition to covering various theoretical matters in considerable depth, also shows a commendable common-sense which has not been adopted by all who have studied this problem. Other papers include those of Anderson and Hauck (1983), Berger and Hsu (1996) and Brown et al. (1997).

In my view, however, there are good reasons for resisting any attempts to replace the 90% confidence interval approach by anything less conservative and, indeed, it can even be claimed that its is too liberal. The basic problem is that all 'improvements' proposed are improvements in the Neyman–Pearson sense only and do not look so good from the perspective of alternative philosophies. To see why it is necessary to look at the conventional approach when attempting to prove superiority of an active treatment to placebo.

Under such circumstances,the regulator will require the sponsor to show, given a null hypothesis of equality, that a conventional two-sided test has rejected this at the 5% level. If this approach is adopted the following can be claimed *whatever the sample size*.

(1) The type I error rate will be 5%.
(2) If efficacy is accepted, the p-value will be 0.05 or less.
(3) If efficacy is accepted, the conventional 95% confidence limits for the effect of treatment will exclude zero.
(4) If efficacy is accepted, the Bayesian 95% highest posterior density interval corresponding to an uninformative prior will exclude zero.
(5) If efficacy is accepted, the likelihood ratio for the best-supported alternative hypothesis compared the null hypothesis will be greater than 6.

It can thus be seen that in some sense at least consensus is maintained between various approaches to examining efficacy.

When we come to look at bioequivalence, however, this approach breaks down. For example, given large enough σ and small enough n, to produce a test with 'correct' size can require us to accept equivalence even though the point estimate for relative bioavailability lies outside the limits of equivalence. Under such circumstances it is quite clear that the likelihood ratio must be *against* the hypothesis of equivalence and *for* in-equivalence. It is also clear that agreement with Bayesian and other approaches will not be produced. Whatever lip service may be given to the Neyman–Pearson approach from certain quarters within drug development, it is quite clear that the practical consensus that this is the *only* reasonable approach just does not exist. No regulator would accept bioequivalence if the point estimate lay outside the limits of equivalence and (I hope) no pharmaceutical statistician would attempt to convince him or her to do so.

In practice, this is not likely to be a serious problem. If it is accepted that a reasonable amount of evidence is needed to conclude anything about a treatment, then in that case we shall never be in the position of being forced or willing to conclude equivalence unless σ is small compared to n, but by the same token there can be no value in improving the conventional approach in a Neyman–Pearson sense. In fact, the various Neyman–Pearson improvements which have been proposed to the conventional approach would only ever be reasonable if the following were the true.

(1) We *had* to make a decision on the basis of the data collected so far so that the decision was between 'accept *now*', 'reject *now*', and a third course of action 'collect more data' was excluded.

(2) The sponsor had the right to determine the amount of evidence which was necessary to come to a conclusion and the regulator was honour bound to accept this decision, provided only that the type I error rate did not exceed the 5% level.
(3) We accepted that maximising power for a given size was a reasonable approach.
(4) We failed to plan trials with adequate numbers of subjects.

But if the first two points are true, what is to prevent a sponsor who fears that his drug may not be bioequivalent adopting the approach that has both highest probability of success and is cheapest, namely that of collecting no data and rolling an icosahedral die?

4. Multi-period trials in two treatments

The various problems which have been outlined above in connection with the analysis of the AB/BA design have led statisticians to seek alternatives in which the effect of carry-over can be eliminated efficiently. One approach has been to add more periods. An extensive review of various designs was made by Matthews (1994a) some years ago in an excellent paper published in this journal. This general field is one characterised by technical brilliance with (from the perspective of an ordinary applied statistician at least) a fair amount of abstract algebra. As an exercise in mathematics this is all a perfectly legitimate exploration of the consequences of various models. To the applied statistician working in drug development, however, this does not justify the application of the results. As will be shown below, there are good grounds for a healthy scepticism regarding the practical utility of these approaches and, in fact, little of this work is of any *direct* relevance to drug development.

Consider as an example a design in four periods and four sequences in which patients are randomised in equal numbers to receive one of the following sequences of treatment: AABB, BBAA, ABBA or BAAB. Any conventional estimator of the treatment contrast for the difference between the effect of A and B can be expressed as a linear combination of the sixteen cell means defined by the cross-classification of the four sequences and four periods. Since there are eight cell means corresponding to treatment A and 8 corresponding to treatment B, a very simple and obvious scheme of weights is to give each of the means corresponding to A a weight of $\frac{1}{8}$ and each corresponding to B a weight of $-\frac{1}{8}$. If the weights are so distributed it will be seen that they add up to zero in any sequence and to zero in any period. This is a simple consequence of the fact that each treatment appears an equal number of times in each sequence and each period. Thus, in addition to estimating the effect of interest, patient and period effects are eliminated.

This design, however has a further property. It is also the case that treatment A is directly followed by itself on three occasions (period 2 of sequence 1, period 4 of sequence 2 and period 3 of sequence 4) and is also directly followed by B on three occasions (period 3 of sequence 1, period 2 of sequence 3 and period 4 of sequence

4). This is a further balancing-type property of this design. (In fact, in the technical literature this property is often referred to as 'balance' as if no other deserved the name.) Now suppose that it were the case that carry-over lasted for one period exactly and depended only on the engendering and not the perturbed treatment (what has been referred to as the 'simple carry-over' model). It would then be the case that we could speak of the carry-over due to A as a *single* effect and that, in the construction of the overall treatment estimate, three cell means associated with this effect would be weighted by $\frac{1}{8}$ and three would be weighted by $-\frac{1}{8}$. Hence, the effect of this carry-over would be eliminated. (A similar phenomenon would apply to the carry-over associated with B.)

Given equal correlation of within patient errors, this design would be fully efficient but, in fact this assumption can be relaxed and further complications such as restricting the number of sequences to two only can be introduced. It may then be the case that there is no fully efficient design but that a best or set of best designs can be found. Various detailed investigations of these and related matter have been carried out by a number of statisticians.

However, an absolutely devastating blow to the rationale of all of this work was struck by Fleiss (1986, 1989). He pointed out that the model for carry-over was not in the slightest bit reasonable. Consider a so-called multi-dose trial (a trial in which patients are given regular therapy, for example twice daily, for a period of time). The main object of the trial will be to study the steady-state effect of treatment. If the period has been chosen so as to permit this to be possible there can be no carry-over in terms of effect of a treatment into itself. For example, in the AABB sequence above, steady state will have been reached by the end of period one and hence the effect at the end of period two under A cannot be greater than it would otherwise be had A not preceded it. On the other hand, it is conceivable that there could be a carry-over from A into B. Hence, in contradistinction to what has been supposed by many statisticians working on the so-called 'optimal' cross-over designs, these designs do not necessarily have anything to do with solving the problem of carry-over and the work of the applied statistician in the pharmaceutical industry. In fact, if carry-over occurs, the alternative steady-state theory (carry-over from A into B but not into A) seems more reasonable.

There is an even more fundamental criticism of all of this work. 'Period' is not even a primitive design constraint in many cases. The task of the statistician is not necessarily to find an efficient design in four periods (say) but to find an efficient design in six months (say). One of the tasks of the statistician and physician in working together will be to determine the length of the period of treatment and this can only be done in the light of pharmacodynamic and pharmacokinetic theory. As has been pointed out, a statistician who knows how to design a trial to study the steady-state effect of treatment knows how to design one to eliminate carry-over. Thus, pharmacological theory, rather than deep understanding of algebra and orthogonality, is the key to good design (Sheiner et al., 1991).

Matthews (1994b) has pointed out that designs which are efficient for the simple-carry-over theory are often efficient for the steady-state theory also, so that the designs which

have been found need not necessarily be abandoned in the light of Fleiss's criticism. However, as Senn and Lambrou (1998) have pointed out, Matthews's claim is only relevant provided the form of carry-over is known at the time of analysis. If both steady state and simple carry-over have to be eliminated, alternative designs may be superior.

In any case, this whole discussion raises the issue of the relevance of the usual design criterion. It has been implicitly assumed in most investigations that minimum variance unbiased estimation is attainable and desirable. Consider, however, the 'optimal' design in two periods: the AB/BA/AA/BB design. For a given number of patients allocated in equal numbers to all sequences, the estimator that eliminates simple carry-over has four times the variance of that of CROS for the AB/BA design. Furthermore, unless simple carry-over applies and the carry-over from A into A equals the carry-over from A into B, the estimate is not even unbiased. If realistic carry-over occurs, and an unbiased estimate is essential, there will also be no choice but to use first period data only, in which case a more efficient approach would have been not to collect the second period data. Thus, the statistician who contemplates using the AA/BB/AB/BA design has already implicitly accepted that some bias is acceptable. But if some bias is acceptable, ought not all three designs (two sequence, four sequence and parallel group) be compared in terms of mean-square error. If this is done for almost any realistic combination of carry-over, within- and between-patient variances, the AB/BA design will come out best.

5. Discussion

These examples illustrate that there may be considerable grounds for concern regarding statistical research into the cross-over trial.

First, the distinction between single-dose pharmacodynamic and multi-dose therapeutic trials needs to be understood by all working on these trials. For the latter, the definition of a 'period' is arbitrary. For both types it needs to be appreciated that patients are not recruited simultaneously. In particular, statisticians should be aware that when designing cross-over trials, the length of period and or washout always has to be determined by the trialist. These determinations are made in the light of beliefs about carry-over.

This brings us to the second point. There is a very vigorous pharmacokinetic school doing important work on the effects of treatments. The theories of pharmacodynamic response that they have been developing have their roots at least as far back as Hill's famous paper (Hill, 1910). If statisticians had paid any attention to this work the implausible simple carry-over model would not have been the overwhelming favorite with them which it has proved.

The work on optimal design for simple carry-over would also have been of more practical benefit if more care had been taken over terminology. For example, 'an optimal design for dealing with carry-over', as commonly applied, usually means,

'minimum variance unbiased design for dealing with the sort of carry-over which lasts for one period and depends only on the engendering treatment'. Since it is doubtful that this sort of carry-over is ever encountered, this sort of a design is not necessarily optimal or even useful for the practical experimenter.

In short, statisticians involved in developing models and designs for cross-over trials need to look beyond mathematics to wider scientific, philosophical and practical issues, if they are going to do work which is useful to their fellow scientists.

Acknowledgements

I thank two anonymous referees for helpful comments on an earlier version.

References

Anderson, S., Hauck, W.W., 1983. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. Comm. Statist. A 12, 2663–2692.

Berger, R.L., Hsu, J.C., 1996. Bioequivalence trials, intersection-union tests, and equivalence confidence sets. Statist. Sci. 11, 283–302.

Brown, B., 1980. The cross-over experiment for clinical trials. Biometrics 36, 69–79.

Brown, L.D., Hwang, J.T.G., Munk, A., 1997. An unbiased test for the bioequivalence problem. Ann. Statist. 25, 2345–2367.

Fleiss, J.L., 1986. Letter to the editor. Biometrics 42, 449–450.

Fleiss, J.L., 1989. A critique of recent research on the two-treatment cross-over design. Control Clin. Trials 10, 1121–1130.

Freeman, P.R., 1989. The performance of the two-stage analysis of two-treatment, two-period cross-over trials. Statist. Med. 8, 1421–1432.

Grieve, A.P., Senn, S.J., 1998. Estimating treatment effects in clinical cross-over trials. J. Biopharm. Statist. 8, 191–233; discussion 235–247.

Grizzle, J.E., 1965. The two-period change-over design and its use in clinical trials. Biometrics 21 467–480 (Corrigenda see Grizzle (1965), Biometrics 30, 727 and Grieve, A.P., 1982. Biometrics 38, 517).

Hill, A.V., 1910. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. Proc. Physiol. Soc. 40, iv–vii.

Hills, M., Armitage, P., 1979. The two-period cross-over trial. Brit. J. Clin. Pharm. 8, 7–20.

Jones, B.J., Kenward, M.G., 1989. Design and Analysis of Cross-Over Trials. Chapman and Hall, London.

Kirkwood, T.B.L., 1981. Bioequivalence testing a need to rethink. Biometrics 37, 589–591.

Matthews, J.N.S., 1994a. Modelling and optimality in the design of cross-over studies for medical applications. J. Statist. Plann. Inference 42, 89–108.

Matthews, J.N.S., 1994b. Multi-period cross-over trials. Statist. Methods Med. Res. 3, 383–405.

Mehring, G.H., 1993. On optimal tests for general interval-hypotheses. Comm. Statist: Theory Methods 22, 1257–1297.

O'Quigley, J., Baudoin, C., 1988. General approaches to the problem of bioequivalence. Statistician 37, 51–58.

Ratkowsky, D.A., Evans, M.A., Alldredge, J.R., 1993. Cross-over Experiments, Design, Analysis and Application. Marcel Dekker, New York.

Senn, S.J., 1988. Cross-over trials, carry-over effects and the art of self-delusion. Statist. Med. 7, 1099–1101.

Senn, S.J., 1993. Cross-over trials in Clinical Research. Wiley, Chichester.

Senn, S.J., 1994. The AB/BA crossover: past, present and future?. Statist. Methods Med. Res. 3, 303–324.

Senn, S.J., 1996. The AB/BA cross-over: how to perform the two stage analysis if you can't be persuaded that you shouldn't. In: Hansen, B., de Ridder, M. (Eds.), Liber Amicorum Roel van Strik. Rotterdam, pp. 93–100.

Senn, S.J., 1997. Statistical Issues in Drug Development. Wiley, Chichester.
Senn, S.J., 1998. Cross-over trials. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics. Wiley, New York.
Senn, S.J., Lambrou, D., 1998. Robust and realistic approaches to carry-over. Statist. Med. 17, 2849–2864.
Sheiner, L.B., Hasimoto, Y., Beal, S.L., 1991. A simulation study comparing studies for dose ranging. Statist. Med. 10, 303–322.
Student, 1908. The probable error of a mean. Biometrika 6, 1–25.
Wang, S.-J., Hung, H.M.J., 1997. Use of two-stage statistic in the two-period cross-over trials. Biometrics 53, 1081–1091.
Westlake, W.J., 1976. Symmetrical confidence intervals for bioequivalence trials. Biometrics 37, 741–744.