BIOMETRICS 36, 69–79
March, 1980

The Crossover Experiment for Clinical Trials

Byron Wm. Brown, Jr.
Stanford University Medical Center, Stanford, California 94305, U.S.A.

SUMMARY

The two-period crossover or changeover design for clinical trials is compared with other simple designs in terms of statistical precision and cost. The sensitivity of the crossover to bias due to carryover effects is examined. The feasibility of using the crossover data to test for the existence of carryover effects is investigated and found to be uneconomical. A numerical example is presented.

1. Introduction

In comparing the efficacy of two drugs in the treatment of a disease, there is great appeal in the idea of using each patient as his own control by trying both drugs on each patient at different times and comparing the results, patient by patient. This idea has strong appeal to physicians especially, as seen, for example, in the discussion in the minutes of the Food and Drug Administration, Biostatistics and Epidemiology Advisory Committee meeting, 23 June 1976, chaired by J. Cornfield, the minutes being submitted by Dr. R. T. O'Neill as secretary. The idea has led to the design called the crossover or changeover design.

In the simple two-treatment crossover experiment, each experimental unit (patient) has two periods of treatment available (first and second). The treatments to be compared (A and B) are randomly allocated to the two periods for each patient. Equivalently, the patients are allocated randomly to two groups, the first group to get the treatments in the order AB, the second group in the reverse order. Usually, equal numbers of patients are assigned to the two groups.

Crossover experiments are extremely popular in clinical pharmacology research. For example, McNair (1971) reported that in a survey of numerous studies of the effects of antianxiety drugs on human performance, 68% of the studies used the crossover approach. The majority of studies used males as subjects, were double-blind and compared one dose level of an active medication with placebo.

Many of the studies submitted to the Bureau of Drugs of the Food and Drug Administration in support of claims of efficacy of new drugs are designed as crossover experiments. Recently, statisticians at the Bureau, led by Dr. R. T. O'Neill, have spurred new interest in the evaluation of the criteria for judging the validity of these studies, and this paper is in part the result of this renewed interest.

2. Review of Past Work

Suppose that individuals are randomly allocated to Groups I and II, n_1 and n_2 individuals to the respective groups. Group I receives treatments A and B in Periods 1 and 2 respectively, while Group II receives the treatments in the reverse order, namely Treatment B in the first period and Treatment A in the second. Note that in most clinical trials

Key words: Crossover; Changeover; Clinical trials; Residual effects; Cost efficiency; Bias.

Periods I and II refer to times relative to the admission of each patient to the study and vary in calender time from patient to patient. For example, in an arthritis study the patient may be admitted to the study when he first seeks and needs help for pain in his joints, and the periods may be the first and second months after admission to the study.

Grizzle (1965) has published the most widely read statistical paper on the use of the crossover experiment in clinical trials, though Koch (1972) has published a parallel nonparametric version. We follow Grizzle's model and notation in our discussion.

One measurement or observation is obtained per subject per period in the standard crossover experiment, although this measurement might itself be an average of several measurements of response taken during the period. The following table shows the layout for the data:

Table 1
Notation and layout for the simple crossover experiment

	Group I		Group II	
Period	Treatment	Subjects $S_{11}\ S_{12} \cdots S_{1n_1}$	Treatment	Subjects $S_{21}\ S_{22} \cdots S_{2n_2}$
1	A	$Y_{111}\ Y_{121} \cdots Y_{1n_11}$	B	$Y_{211}\ Y_{221} \cdots Y_{2n_21}$
2	B	$Y_{112}\ Y_{122} \cdots Y_{1n_12}$	A	$Y_{222}\ Y_{222} \cdots Y_{2n_22}$

The model for the observation is

$$Y_{ijk} = \mu + \pi_k + \phi_u + \lambda_v + \xi_{ij} + \varepsilon_{ijk}, \tag{1}$$

where

μ = overall mean;
π_k = effect of the kth period, $k = 1, 2$;
ϕ_u = effect of the uth drug, $u = A, B$;
λ_v = residual effect of the vth drug in the first period on the response in the second period, $v = A, B$; ($\lambda_v = 0$ for all first-period measurements);
ξ_{ij} = the effect of the jth subject in the ith group, $i = 1, 2$; $j = 1, 2, \ldots, n_i$;
ε_{ijk} = within-subject deviation for the kth period.

We take μ, π_k, ϕ_u and λ_v to be fixed and ξ_{ij}, ε_{ijk} to be random and mutually independent with means zero and with variances σ_ξ^2 and σ_ε^2, respectively.

A comment concerning the meaning or interpretation of the λ_v is in order. One could distinguish between two situations in the case of non-null treatment effects, i.e., the case in which $\phi_A \neq \phi_B$. In this case the expectations for the two groups may be different at the start of Period 2, and if the treatment effects depend on the starting level or mean then this in itself will cause differential residual effects for the treatments (i.e. $\lambda_v \neq 0$) in the second period. However, there is a hope of eliminating this type of interaction through transformation to a more appropriate scale. More serious residual effects would be those in which a treatment has a reverse effect or a very marked effect when preceded by the other treatment but not otherwise. Transformation of scale will not allow a model with $\lambda_v = 0$ in such situations. In practice, as Hills and Armitage (1979) point out, the two situations may not be easy to distinguish and *any* situation in which λ_v may not be zero is a cause for concern.

It should be noted that the discussion and results obtained under the parametric model given here could be developed as well by use of the randomization approach, described so clearly by Kempthorne (1977) in his discussion of the need for randomization in experiments. The conclusions of previous writers and the present writer would be essentially the same, whether the parametric or permutation approach is adopted. The advantage of the latter approach is that it would highlight the importance of randomizing the subject to the two groups as a basis for inference.

Cox (1958) and Cochran and Cox (1957) describe the crossover for the case in which there is no residual effect of treatment (i.e. λ_v can be assumed to be zero). In this case the analysis is straightforward. In particular, with the arbitrary constraint that $\phi_A + \phi_B = 0$, the treatment difference, $\delta = \phi_B - \phi_A$, can be estimated by

$$\hat{\delta}_{CO} = \tfrac{1}{2}\{(\bar{Y}_{1.2} - \bar{Y}_{1.1}) + (\bar{Y}_{2.1} - \bar{Y}_{2.2})\}. \tag{2}$$

From the model (1) it can be seen that $\hat{\delta}_{CO}$ is unbiased, with variance

$$\text{var}\,(\hat{\delta}_{CO}) = \frac{\sigma_\varepsilon^2}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \tag{3}$$

Since twice σ_ε^2 can be estimated from subject differences in the two periods, by pooling the variances for the two groups, with $(n_1 - 1) + (n_2 - 1)$ total degrees of freedom, inferences concerning δ are straightforward.

Note that if the second period of the crossover experiment is omitted, one has the usual completely randomized experiment. Chassan (1970) has noted the efficiency of the crossover design relative to the completely randomized design. If one uses n subjects in each group for a crossover (CO) and m subjects in each group for a completely randomized experiment (CR) with no crossover, the variance of the estimator, $\hat{\delta}_{CR}$, from the completely randomized experiment will be

$$\text{var}\,(\hat{\delta}_{CR}) = (\sigma_\xi^2 + \sigma_\varepsilon^2)(2/m) \tag{4}$$

and the ratio of the variances (3) and (4) from the two experiments is

$$\frac{\text{var}\,(\hat{\delta}_{CO})}{\text{var}\,(\hat{\delta}_{CR})} = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\xi^2}\frac{m}{2n}. \tag{5}$$

Note that $\sigma_\xi^2/(\sigma_\varepsilon^2 + \sigma_\xi^2)$ is the correlation between pairs of measurements in Periods 1 and 2, taken on randomly selected subjects. If we take $m = 2n$ in order to equate the numbers of *measurements* taken in the two designs, then the ratio of variances is simply the complement of this intrasubject correlation, i.e. $\sigma_\varepsilon^2/(\sigma_\varepsilon^2 + \sigma_\xi^2)$.

Chassan (1970) also noted that the completely randomized experiment could incorporate baseline measurements. No crossover would be involved but differencing within subjects is introduced. The sequence of treatments in the two periods might be *CA* and *CB* for Groups I and II (*C* denoting a placebo treatment, no treatment or perhaps one of the two study treatments). For n subjects in each group the model now yields an estimate of $\delta = \phi_B - \phi_A$, call it $\hat{\delta}_{CR/BL}$, with variance

$$\text{var}\,(\hat{\delta}_{CR/BL}) = 4\sigma_\varepsilon^2/n, \tag{6}$$

four times as large as that of the crossover estimate regardless of the value of σ_ξ^2. Use of a strong baseline covariate in a completely randomized design would achieve much the same order of efficiency.

Cox (1958) emphasized in his very clear discussion that the crossover experiment is but

a simple example of a split-plot design, when the assumption of no residual effect ($\lambda_v = 0$ in our notation) is valid.

A number of investigators, e.g. Williams (1949), Patterson (1951) and Cochran and Cox (1957), have proposed design and estimation procedures for estimating the carryover or residual effects in a crossover design. However, additional combinations of treatment sequences must be used, e.g. *AA*, *AB*, *BA* and *BB* in the two-treatment, two-period crossover. This more complex design defeats the object of economy that is a major appeal of the crossover to both clinical investigators and statisticians.

Grizzle (1965) pointed out that the simple crossover design, with randomization, would indeed yield an estimate of the residual effect, and thus would permit testing the presence of the residual effect and proceeding with the orthodox estimation method if the hypothesis of no residual effect were confirmed.

The procedure for estimating the residual effect is simple. In the model above, let $\lambda_A + \lambda_B = 0$. This means that the *average* residual (i.e. carryover) effect is absorbed by the parameters for period effect, and $\gamma = \lambda_B - \lambda_A$ is the *difference* in residual effects of the two treatments. An unbiased estimate of γ, denoted by $\hat{\gamma}_{CO}$, is

$$\hat{\gamma}_{CO} = (\bar{Y}_{2.1} + \bar{Y}_{2.2}) - (\bar{Y}_{1.1} + \bar{Y}_{1.2}) \tag{7}$$

$$= (\lambda_B - \lambda_A) + 2(\bar{\xi}_{2.} - \bar{\xi}_{1.}) + (\bar{\varepsilon}_{2.1} + \bar{\varepsilon}_{2.2} - \bar{\varepsilon}_{1.1} - \bar{\varepsilon}_{1.2}). \tag{8}$$

Note that $\hat{\gamma}_{CO}$ is simply the difference between the mean responses for Group I and Group II. The variance of the estimate, of course, contains the between-subject component of variance:

$$\text{var}\,(\hat{\gamma}_{CO}) = (4\sigma_\xi^2 + 2\sigma_\varepsilon^2)\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \tag{9}$$

The sum of the two observations for any individual (one observation per period) will have a variance of $4\sigma_\xi^2 + 2\sigma_\varepsilon^2$. Thus, the variance of these sums for each group can be pooled to yield an estimate of $4\sigma_\xi^2 + 2\sigma_\varepsilon^2$, with $n_1 + n_2 - 2$ degrees of freedom, and thus inferences concerning γ are straightforward.

If γ, the difference in residual effects, is not zero then $\hat{\delta}_{CO}$ will not be an unbiased estimate of the main treatment difference, δ. In this case (indeed, regardless of the value of γ) an unbiased estimate of δ can be obtained by using only the data from the first period,

$$\hat{\delta}_{COI} = \bar{Y}_{2.1} - \bar{Y}_{1.1}, \tag{10}$$

but this estimator makes no use of the crossover nature of the experiment and leads to an estimator with both between-subject and within-subject components of variance:

$$\text{var}\,(\hat{\delta}_{COI}) = (\sigma_\xi^2 + \sigma_\varepsilon^2)\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \tag{11}$$

Cox (1958) and Grizzle (1965) advise that the crossover be used only when it can be assumed that there are no residual effects, but Grizzle goes on to advise that in doubtful cases one could do the crossover, test whether $\gamma = 0$ at a significance level of 10% and, if the hypothesis is not rejected, use the orthodox estimate $\hat{\delta}_{CO}$ rather than resorting to $\hat{\delta}_{COI}$.

Despite these discussions in the literature, the Bureau of Drugs of the Food and Drug Administration remained puzzled concerning the validity of certain clinical crossover experiments and asked an advisory committee for further guidance. In 1977 Dr. R. T. O'Neill distributed a report of that committee at a meeting of the Pharmaceutical

Manufacturers' Association Firm Statisticians' Meeting. In this report the Biometrics and Epidemiology Methodology Advisory Committee (BEMAC) held that 'In most cases, the completely randomized (or randomized block) experiment will be the design of choice because it furnishes unbiased estimates of treatment effects without appeal to any modeling assumptions save those associated with the randomization procedure itself'. Thus, BEMAC was very negative about the usefulness of the crossover experiment for clinical studies.

The purpose of the present paper is to examine more closely the cost savings that can be realized by the crossover experiment and to study the Grizzle suggestion as a means of saving the crossover experiment from the BEMAC criticism. During the preparation of this paper Hills and Armitage (1979) published a very comprehensive discussion of the two-period crossover. They came to much the same view as that suggested below.

In the following sections of this paper we first look at the savings achievable with the crossover design when the assumption of no residual effects is valid. Next we examine the strategy, suggested by Grizzle, of testing for residual effects and proceeding conditionally on the results. We than present an example. In a final section we summarize and suggest further work.

3. Cost Efficiency of the Crossover

In order to appreciate the economics of the crossover relative to the completely randomized design, with or without baseline measurements, the cost must be separated into two components. Let S_0 be the cost of obtaining a new patient (by seeking referrals and screening, then testing the candidate for eligibility and obtaining the patient's agreement to participate). Let S_1 be the cost of treating and measuring the patient *in a given period.*

Thus, in a crossover experiment the total cost, S_{CO}, for a study with n patients in each group would be

$$S_{CO} = 2nS_0 + 4nS_1. \tag{12}$$

The cost of the completely randomized experiment with m patients per group, denoted by S_{CR}, would be

$$S_{CR} = 2mS_0 + 2mS_1. \tag{13}$$

If the sample sizes, n and m, for the crossover and the completely randomized experiments respectively, are selected so as to make the estimates of treatment effect equally precise, using expression (5) the following relationship between required sample sizes is obtained:

$$n = \frac{\sigma_\varepsilon^2}{\sigma_\xi^2 + \sigma_\varepsilon^2} \cdot \frac{m}{2}. \tag{14}$$

Let R denote the relative costs of these two experiments, i.e. $R = S_{CO}/S_{CR}$. Then, from (12), (13) and (14), we have

$$R = \frac{1}{1 + \sigma_\xi^2/\sigma_\varepsilon^2} \frac{(1 + 2S_1/S_0)}{2(1 + S_1/S_0)}. \tag{15}$$

Thus, the cost of the crossover experiment relative to that of the completely randomized experiment depends on two ratios: the ratio of between-subject variation to within-subject variation ($\sigma_\xi^2/\sigma_\varepsilon^2$) and the ratio of the cost of treating a patient for one period relative to the cost of recruiting the patient into the study (S_1/S_0).

Of course, large recruiting costs (S_1/S_0 small) and large variation between subjects (large $\sigma_\xi^2/\sigma_\varepsilon^2$) argue against the completely randomized experiment and favor the crossover experiment. Note that

$$\lim_{S_1/S_0 \to 0} R = \frac{1}{2} \cdot \frac{1}{1+\sigma_\xi^2/\sigma_\varepsilon^2}, \tag{16}$$

$$\lim_{\sigma_\xi^2/\sigma_\varepsilon^2 \to \infty} R = 0. \tag{17}$$

On the other hand, if recruiting costs are relatively small and variation between subjects is relatively small, the cost efficiency of the crossover experiment relative to the completely randomized experiment can approximate to unity. In fact, the cost saving in using the crossover design as contrasted with the completely randomized experiment in a typical clinical trial situation can be surprisingly small. Table 2 shows the cost efficiency, R, for various combinations of S_1/S_0 and $\sigma_\xi^2/\sigma_\varepsilon^2$. For example, it can be seen from Table 2 that if the variance between subjects is equal to the variance within subjects, and if the cost of treatment is four times the cost of recruitment, the cost of the crossover is 45% of the cost of the comparable completely randomized experiment. Savings displayed in the table range from 13% ($R = 0.87$) to 95% ($R = 0.05$), the latter in the extreme case of between-subject variance ten times the within-subject variance and costs of recruiting ten times the cost of treatment.

This method of cost comparison may favor unfairly the crossover experiment in several regards. First, the cost of recruiting, S_0, is taken to be the same for the crossover and for the completely randomized experiment. In fact, the requirement of two treatment periods per patient for the crossover may make recruitment for the crossover more difficult and thus increase the costs of recruitment for the crossover. (However, in some applications, candidates may be *positively* influenced by the assurance that they will be allowed both treatments.) Second, there may be more dropouts from the crossover. This will necessitate recruitment of more patients to compensate for anticipated losses in cases and the losses themselves will cause problems in analysis of the results. These considerations will be specific to each study and no effort has been made to incorporate them into quantitative comparisons of the costs of the two designs.

Table 2

Cost efficiency for the crossover experiment relative to the completely randomized experiment for various ratios of cost for treatment to cost for recruitment, and for various ratios of the component of variance between to within patients

S_1/S_0^*	Ratio of between-subject to within-subject variance, $\sigma_\xi^2/\sigma_\varepsilon^2$						
	1/10	1/4	1/2	1	2	4	10
1/10	0.50†	0.44	0.36	0.27	0.18	0.11	0.05
1/2	0.61	0.53	0.44	0.33	0.22	0.13	0.06
1	0.68	0.60	0.50	0.38	0.25	0.15	0.07
2	0.76	0.67	0.56	0.42	0.28	0.17	0.08
4	0.82	0.72	0.60	0.45	0.30	0.18	0.09
10	0.87	0.76	0.64	0.48	0.32	0.19	0.09

* S_1 = cost of treatment per patient per period; S_0 = cost of recruitment per patient.

† Cost of the crossover divided by the cost of a randomized experiment with the same precision. See (15).

4. Testing the Validity of the Crossover

Suppose that calculations based on the approach of the preceding section suggest that the savings which might be realized by the crossover design are substantial. If the validity of the assumption of no residual or carryover effects seems tenable, the crossover design should be pursued. If the assumption is clearly justified and acceptable from prior information, the crossover design should be used, with sample size determined by use of (3).

Suppose, however, that the crossover design holds great potential for savings but the assumption of no carryover effects is in doubt. One might consider a crossover design with sufficient power to test the assumption of no carryover effects, with the intention of doing a preliminary test, as Grizzle (1965) suggested, and estimating the treatment effects on the basis of first-period results or the results of both periods, depending on the results of the preliminary test. This plan can be shown to require so many subjects that one is better off using the completely randomized design. Consider the following argument.

Suppose one desires a test of the null hypothesis that $\delta = \phi_B - \phi_A = 0$ against the alternative that $\delta = \delta_1 \neq 0$ at the α level of significance with power of $1-\beta$. For the completely randomized experiment, (4) implies that the appropriate sample size (i.e. the number of subjects per group) is approximately

$$n_{CR} = \frac{2(Z_{\alpha/2} + Z_\beta)^2(\sigma_\xi^2 + \sigma_\varepsilon^2)}{\delta_1^2}, \tag{18}$$

where $Z_{\alpha/2}$ and Z_β are upper percentiles of the normal distribution. [This is a good approximation for large degrees of freedom; see, e.g., Snedecor and Cochran (1967) for a discussion of the one-sample analogue.]

If one follows Grizzle's advice, the preliminary test of the hypothesis $\gamma = \lambda_B - \lambda_A$ will be done at the $\alpha' = 10\%$ level of significance. This preliminary test must be powerful enough to detect any alternative that would result in undue bias in the test of $\delta = \phi_B - \phi_A$, and this bias can be seen from expressions (1) and (2) to be equal to $\frac{1}{2}\gamma$. Suppose that we consider a bias of one fourth of the alternative of interest to be sufficiently small. Then, we might require a power of $1-\beta' = 95\%$ at an alternative of $\gamma_1 = \frac{1}{2}\delta_1$ in order to ensure that the bias is less than one fourth of the alternative of interest.

From expression (1) it is clear that an unbiased estimate of γ is obtained from the differences in group means as given by expression (7), with variance given by expression (9).

If n_{CO} is the number of subjects in each crossover group, the desired power will be achieved by n_{CO} approximately equal to

$$n_{CO} = \frac{2(Z_{\alpha'/2} + Z_{\beta'})^2(4\sigma_\xi^2 + 2\sigma_\varepsilon^2)}{(\delta_1/2)^2}, \tag{19}$$

where $Z_{\alpha'/2}$ and $Z_{\beta'}$ are upper percentiles of the normal distribution.

Comparing expressions (18) and (19), it is clear that for $\alpha' = 10\%$ and $\beta' = 5\%$, for any usual choices of α and β and for common values of $\sigma_\xi^2/\sigma_\varepsilon^2$, the number of subjects required by the crossover experiment will be far greater than the number required by the completely randomized experiment. For example, suppose that $\alpha = 5\%$ and $\beta = 5\%$, $\alpha' = 10\%$ and $\beta' = 5\%$, and $\sigma_\xi^2 = \sigma_\varepsilon^2$; we have

$$\frac{n_{CO}}{n_{CR}} = \frac{2(Z_{0.05} + Z_{0.05})^2(6\sigma_\varepsilon^2)}{\delta_1^2/4} \cdot \frac{\delta_1^2}{2(Z_{0.025} + Z_{0.05})^2 2\sigma_\varepsilon^2} = 10. \tag{20}$$

Thus, in order to assure a test with relatively small bias and the same power as the completely randomized experiment, the crossover experiment would require 10 times as many subjects as the completely randomized experiment.

5. Example of a Crossover Analysis

Varma and Chilton (1974) discussed the analysis of a simple crossover dental study comparing a test compound with a placebo with regard to effect on dental hygiene as measured by the decrease in an oral hygiene index. The study was first reported by Zenner, Duany and Chilton (1971). The data were supplied to the present writer by O'Neill and are presented in Table 3. The summary statistics, estimates and standard errors are presented in Table 4.

Note that the estimate of the treatment effect is +0.7712, more than six standard errors distant from zero and hence very highly significant. But this presupposes that the residual

Table 3
Improvement in hygiene index for a crossover study

Subject	Group I		Group II	
	Period 1 Placebo	Period 2 Test	Period 1 Test	Period 2 Placebo
1	0.83	1.83	1.67	0.33
2	1.00	2.17	2.50	0.50
3	0.67	1.67	1.00	−0.17
4	0.50	1.50	1.67	0.50
5	0.50	2.33	1.83	0.50
6	0.83	1.83	0.50	0.33
7	1.00	0.50	1.33	0.67
8	0.67	0.33	1.33	0.00
9	0.67	0.50	0.50	0.17
10	0.33	0.67	2.17	0.83
11	0.00	0.83	1.67	0.33
12	1.17	1.33	1.50	0.00
13	0.00	0.67	1.33	0.50
14	0.50	1.83	1.50	0.50
15	0.33	1.50	1.33	0.00
16	0.33	1.50	0.67	−0.17
17	0.50	1.17	1.67	0.50
18	1.00	1.67	2.50	0.67
19	0.00	1.33	1.83	0.00
20	0.50	1.50	0.83	0.67
21	−0.50	2.83	2.33	0.17
22	0.17	2.33	1.17	0.50
23	1.00	1.33	1.33	0.00
24	1.00	1.67	1.33	0.83
25	1.33	0.67	0.33	1.33
26	0.33	0.83	2.17	1.17
27	2.00	1.00	1.00	0.33
28	4.00	0.17	0.33	1.00
29	0.83	1.67	1.17	0.17
30	0.50	1.33	0.50	0.50
31	0.50	1.50		
32	0.50	1.67		
33	2.17	1.33		
34	0.67	1.17		

Table 4
Summary statistics and tests of significance for the hygiene data

	Analysis of differences		Analysis of sums	
	Group I	Group II	Group I	Group II
Mean	0.5985	−0.9440	2.1176	1.7883
Variance	1.3268	0.5173	0.6059	0.5333
Number	34	30	34	30
Pooled variance	0.9482		0.5719	
	Est. trtmt effect	+0.7712	Est. residual effect	−0.3294
	Std error	0.1220	Std error	0.1894
	$t = 6.32$		$t = 1.73$	
	df = 62		df = 62	
	$P < 0.001$		$P = 0.087$	

effect is zero. When the residual effect is calculated the estimate is −0.3294 with a standard error of 0.1894. By Grizzle's criterion, the residual effect is just statistically significant ($P = 0.09$) and we might decide to use only the first-period data. The bias in estimating the treatment effect when there is a residual effect, γ, is $-\frac{1}{2}\gamma$; an estimate of treatment effect based on both periods must allow for a possible bias in the range $-\frac{1}{2}(\gamma \pm 2SE)$ or -0.1647 ± 0.1894. Thus, it is clear that the possibility of a residual effect compromises the estimate of the treatment effect and its precision, even though the data do not demonstrate persuasively the existence of a residual effect of important magnitude. If we estimate the treatment effect on the basis of the first-period data alone we have the following results:

	Group I (Placebo)	Group II (Treatment)
Mean	0.7597	1.3663
Variance	0.6000	0.3845
Number	34	30
Pooled variance	0.4992	
Treatment effect	0.6066	
Standard error	0.1770	
t	3.4271	
df	62	
P	0.001	

Thus, the treatment effect based on the first-period data only is 0.6066 ± 0.1770 (est. ± SE), Compared with 0.7712 ± 0.1220 for the crossover; the first-period data provide persuasive evidence of a treatment effect, while the crossover analysis allows the possibility of a large bias due to residual effects. Fortunately, the study was large enough to provide a precise estimate of the treatment effect from the first-period data alone. If the study had been smaller no definitive conclusion would have been possible.

6. Conclusions

The conclusions that might be drawn from this review and discussion simply reinforce the comments of Cox (1958) and Cochran and Cox (1957), namely that the crossover experiment can yield great savings in cost if the assumption of no carryover effect is valid, but the design should not be used if this assumption is in doubt. Grizzle's (1965) test for the validity of the assumption of no residual effect is certainly a valid test but does not have adequate power, so it offers no practical help in testing the assumption using the crossover data themselves. Of course, it is still advisable to do the test when a crossover is analyzed, to obtain whatever information is available on the validity of the assumption.

Other workers have suggested embellishments of the crossover design, for example inclusion of baseline measurements before the first period or before both periods, and perhaps another measurement after the second period. These designs are worth considering but not for the purpose of economy or precision per dollar expended. All are more expensive than the completely randomized design. Their values lie, not in a simple comparison of the two treatments, but in the information they may yield concerning period effects, washout of residual drug, trends within period and so forth.

Further work on the statistical aspects of the crossover design and its usefulness in clinical research might be expended in several directions. Data might be obtained and evaluated in areas of major biomedical research effort in order to define those areas and experimental conditions in which the assumption of no residual effect might be made safely. Supplementing such results, costs and variance components might be evaluated so as to provide measures of the cost savings realized by the crossover when it can be employed validly. Derivation of the finite permutation model results for the simple crossover would also be of interest. Generalization of cost efficiency studies to crossover designs with more than two treatments would be useful. Finally, development and evaluation of the more complex variations of the crossover, e.g. with baseline measurements or repetition of the same treatment, would be useful. Such work could be of practical interest and use to workers in this field, but the basic results emphasized in the present paper seem inescapable: namely, that in many situations the crossover may not save as much in cost as is believed, and that if the assumption on which the crossover hangs is in doubt, there is no economical way to test the assumption adequately using the data from the crossover itself.

ACKNOWLEDGEMENTS

The author wishes to acknowledge useful suggestions from several persons who read preliminary versions of this paper, notably Rich Simon, Robert O'Neill, Eugene Laska, Rupert Miller, Wilfrid Westlake, D. R. Cox and a referee. Dr Laska allowed a reading of a paper in preparation (with M. Meisner and H. Kushner) proposing designs of the form *AA*, *AB*, *BA*, *BB* as alternatives to the *AB*, *BA* crossover. The present paper was prepared under a Public Health Service Grant and had limited distribution as Technical Report No. 43, Division of Biostatistics, Stanford University.

RÉSUMÉ

On compare la précision statistique et le coût du plan d'expérience avec permutation de traitements sur deux périodes pour les essais cliniques avec d'autres plans simples. On examine la sensibilité du plan avec permutation de traitements au biai dû aux effets de report. On étudie la possibilité d'utiliser les données du plan avec permutation de traitements pour tester l'existence d'effets de report, et on trouve cela peu économique. On présente un exemple numérique.

References

Chassan, J. B. (1970). A note on relative efficiency in clinical trials. *Journal of Clinical Pharmacology* **10,** 359–360.

Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*, 2nd ed. New York: Wiley.

Cox, D. R. (1958). *Planning of Experiments.* New York: Wiley.

Grizzle, J. E. (1965). The two-period change-over design and its use in clinical trials. *Biometrics* **21,** 467–480, and Corrigenda in *Biometrics* **30,** 727 (1974).

Hills, M. and Armitage, P. (1979). The two-period cross-over clinical trial. *British Journal of Clinical Pharmacology* **8,** 7–20.

Kempthorne, O. (1977). Why randomize? *Journal of Statistical Planning and Inference* **1,** 1–25.

Koch, G. G. (1972). The use of non-parametric methods in the statistical analysis of the two-period changeover design. *Biometrics* **28,** 577–584.

McNair, D. M. (1971). Antianxiety drugs and human performance. *Archives of General Psychiatry, Chicago* **29,** 611–617.

Patterson, H. D. (1951). Change-over trials. *Journal of the Royal Statistical Society, Series B* **13,** 256–271.

Snedecor, G. W. and Cochran, W. G. (1967). *Statistical Methods*, 6th ed. Ames: Iowa State University Press.

Varma, A. O. and Chilton, N. W. (1974). Crossover designs involving two treatments. *Journal of Periodontal Research* **9,** Suppl. 14, 160–170.

Williams, E. J. (1949). Experimental designs balanced for the estimation of the residual effects of treatments. *Australian Journal of Scientific Research, Series A* **2,** 149.

Zinner, D. D., Duany, L. F. and Chilton, N. W. (1970). Controlled study of the clinical effectiveness of a new oxygen gel on plaque, oral debris and gingival inflammation. *Pharmacology and Therapeutics in Dentistry* **1,** 7–15.

Received July 1979; *revised October* 1979