advocates, this approach is seriously flawed as we can never satisfactorily eliminate possible biases due to other factors that may have changed over time. Pocock (1977) showed that in 19 cases where the same therapy was used in two consecutive trials of cancer chemotherapy in the same institution there were large changes in the observed death rates, ranging from −46% to +24%. While some of the variation was probably due to small sample sizes, four of the differences were statistically significant at the 2% level. Sacks *et al.* (1983) compared trials of the same therapies in which randomized or historical controls were used, and found a consistent tendency for historically controlled trials to yield more optimistic results than randomized trials. The use of historical controls can only be justified in tightly controlled situations of relatively rare conditions, such as in evaluating therapies for advanced cancer.

The balance of opinion has now swung so far towards randomized trials that the results of non-randomized trials may cause major controversy. A recent example was the study of the possible benefit of vitamin supplementation at the time of conception in women at high risk of having a baby with a neural tube defect (NTD) (Smithells *et al.*, 1980). They found that the vitamin group subsequently had fewer NTD babies than the placebo control group, but because the study was not randomized the findings are not widely accepted, and the Medical Research Council is now running a large randomized trial to try to get a proper answer to the question.

15.2.5 Alternative designs

The simplest design for a clinical trial is called the **parallel group** design, in which two different groups of patients are studied concurrently. This is the design that has been implicit in this chapter so far. The most common alternative is the **crossover design**, which is described below together with some other less common designs that are worth knowing about.

(a) Crossover design

A crossover trial is one in which the same group of patients are given both (or all) treatments of interest in sequence. Here randomization is used to determine the order in which the treatments are received. The crossover design has some attractive features, in particular that the treatment comparison is 'within-subject' rather than 'between-subject', and that the sample size needed is smaller. There are some important disadvantages, however, which I shall describe in relation to a two-period crossover trial:

1. Patients may drop out after the first treatment, and so not receive the second treatment. Withdrawal may be related to side-effects. Treatment periods should be fairly short to minimize the risk of drop-out for other reasons.

2. There may be a **carry-over** of treatment effect from one period to the next, so that the results obtained during the second treatment are affected by what happened in the first period. In other words, the observed difference between the treatments will depend upon the order in which they were received. In the presence of such a **treatment-period interaction** the data for the second period may have to be discarded, severely weakening the power of the trial.
3. There may be some systematic difference between the two periods of the trial. For example, the observations in the second period may be somewhat lower than those in the first period, regardless of treatment. A small **period effect** is not too serious, as it applies equally to both treatments.
4. Crossover studies cannot be used for conditions which can be cured, and are most suitable when the effect of the treatment can be assessed quickly.

It is desirable to establish in advance that there will not be any carry-over treatment effect, but the information may be unavailable. A **wash-out** period is sometimes introduced between the treatment periods to try to eliminate carry-over effects. Because of the problems described, crossover studies are probably overused. Further discussion is given by Woods *et al*. (1989).

The analysis of crossover trials is explained and illustrated in section 15.4.10.

(b) Within group (paired) comparisons

Another type of within group design is when alternative treatments are investigated in the same subjects *at the same time*. It can be used for treatments that can be given independently to matching parts of the anatomy, such as limbs or eyes. The matched design has all the advantages of the crossover design, but none of the disadvantages, so is a very powerful design. Unfortunately, there are few circumstances in which it can be used.

The nearest equivalent to the paired within subject design is the matched pairs design, where pairs of subjects are matched for, say, age, sex and certain prognostic factors, and the two treatments are then allocated to the pair of subjects at random. This design can only be used easily when there is a pool of subjects that can be entered into the trial, in order to be able to find matched pairs. Where there are known important prognostic variables the design removes much of the between subject variation, and ensures that the subjects receiving each treatment have very similar characteristics.

(c) Sequential designs

Another type of design is the **sequential trial**, in which parallel groups are

studied, but the trial continue
seen or it is unlikely that any
of sequential trials is that they
there is a large difference in the

In sequential trials the data
become available. Their use is
outcome is known relatively qui
section 15.2.6), and possibly als

A useful variation on this pri
the data are analysed after eac
four or five times in all. This
(regarding length) but also enal
treatment difference is seen.

In the right circumstances seq
should be used more frequently.

(d) Factorial designs

One further type of design is
treatments, say A and B, are sin
with a control. Patients are di
control treatment, A only, B on
the investigation of the interacti
factorial design is rarely used in
describes some examples of its us

(e) Adaptive designs

Ethical considerations have led s
in which the proportion of subje
shes as the trial proceeds. In oth
some extent on the outcome of t
Apart from practical difficulties
results from each patient, it is
any ethical problems. Adaptive d

(f) Zelen's design

Lastly, Zelen (1979) proposed a
seems to avoid problems associat
the subjects are allocated at rand
are treated as if they were not i
new experimental treatment, bu
treatment if they wish. An esse
1979) is that the two groups
regardless of which treatment tho
While this design has some useful
high proportion of those offered

younger and older patients and compare the two P values. This analysis makes comparisons *between* the two groups based on analyses carried out separately *within* each group, and is not a valid method. (A similar situation was described in the previous section.) The correct approach is to compare the difference between the treatments for the two age groups; in other words we look at the *interaction* between age and treatment. The possibility of an interaction can be examined within an appropriate multiple regression model, whether the outcome variable is continuous, binary or survival time. I recommend expert advice for this analysis. (See also Pocock, 1983, p. 213.) Note that this analysis is more like that from an observational study, and so we cannot infer causality from any association.

15.4.10 Crossover trials

Crossover trials were described in section 15.2.5. The analysis of a crossover trial will be illustrated using data from a trial comparing nicardipine, a calcium-channel blocker, and placebo in the treatment of Raynaud's phenomenon (Kahan *et al.*, 1987). The data, representing the number of attacks in two weeks, are shown in Table 15.5 separately for the groups having nicardipine followed by placebo and *vice versa*.

The analysis is simplified by calculating for each subject the difference (d_i) and average (a_i) of the observations in the two periods, and averaging these for each group as shown in Table 15.5. It is incorrect to ignore the design of the study and just perform a simple comparison of treatments. Before comparing the treatments there are two other tests that should be carried out. The correct analysis consists of three two sample t tests or Mann-Whitney tests; t tests are used here. (For categorical data we use χ^2 tests.)

The possibility of a period effect is tested by a two sample t test to compare the differences between the periods in the two groups of patients. If there was no general tendency for patients to do better in one of the periods we would expect the mean differences between the periods in the two groups to be of the same size but having opposite signs. The test for a period effect is thus a two sample t test comparing $\bar{d}_1$ with $-\bar{d}_2$.

We investigate the possibility of a treatment–period interaction by noticing that in the absence of an interaction a patient's average response to the two treatments would be the same regardless of the order in which they were received. The test for interaction is thus a two sample t test comparing $\bar{a}_1$ with $\bar{a}_2$.

If there is no period effect and no treatment-period interaction the analysis of a crossover trial is simple. However, it is important to investigate possible problems before carrying out the treatment comparison. Both a marked period effect and a treatment-period interaction are worrying because they mean that the observed magnitude of the treatment

Table 15.5 Results from a randomized double-blind crossover trial comparing nicardipine (N) and placebo (P) in patients with Raynaud's phenomenon (Kahan *et al.*, 1987). The data are the number of attacks in two weeks. There was a one-week wash-out period between the two treatment periods

Group A: Nicardipine followed by placebo ($n = 10$)

	Period 1 Nicardipine	Period 2 Placebo	(1) − (2)	$\frac{(1)+(2)}{2}$	P − N
	16	12	4	14	−4
	26	19	7	22.5	−7
	8	20	−12	14	12
	37	44	−7	40.5	7
	9	25	−16	17	16
	41	36	5	38.5	−5
	52	36	16	44	−16
	10	11	−1	10.5	1
	11	20	−9	15.5	9
	30	27	3	28.5	−3
Mean	24.0	25.0	−1.0 ($\bar{d}_1$)	24.5 ($\bar{a}_1$)	1.0
SD	15.61	10.84	9.87	12.50	9.87

Group B: Placebo followed by nicardipine ($n = 10$)

	Period 1 Placebo	Period 2 Nicardipine	(1) − (2)	$\frac{(1)+(2)}{2}$	P − N
	18	12	6	15	6
	12	4	8	8	8
	46	37	9	41.5	9
	51	58	−7	54.5	−7
	28	2	26	15	26
	29	18	11	23.5	11
	51	44	7	47.5	7
	46	14	32	30	32
	18	30	−12	24	−12
	44	4	40	24	40
Mean	34.3	22.3	12.0 ($\bar{d}_2$)	28.3 ($\bar{a}_2$)	12.0
SD	14.99	19.14	16.34	15.12	16.34

effect depends on the order in wh
is a more serious problem becau
treatment effect. (See also section

We can test the treatment effec
20 within subject differences betw
crossover groups may not be the
average effect in the two periods.
sample t test to compare $\bar{d}_1$ and $\bar{a}$

For this example the period eff
$t = 1.82$ and $t = 0.613$ respectivel
$P = 0.09$ and $P = 0.55$. As neither
evaluate the treatment effect usin
$t = 2.154$ on 18 degrees of freedo
two weeks on nicardipine was on
on placebo, with a 95% confiden
statistically significant at the 5%
nicardipine is uncertain, reflecting

A problem with the analysis of
for a possible treatment–period in
power. The above analysis is a g
that patients in group 1 did nea
nicardipine, suggesting a long-last
Patients in group 2 showed a big
placebo to nicardipine. This appa
significant. The data from perio
benefit of nicardipine might well
overall results of the trial.

In contrast Ueshima *et al.*
($0.05 < P < 0.10$) treatment–perioc
tigate the possible effect on bloc
They discarded the data from the s

A graphical approach is to pl
between the two periods against
different symbols to identify the
Vertical separation of the two g
between the treatments. If there
should be no horizontal differenc
the two groups should lie symmetr
Figure 15.3(a). Figure 15.3(b) sho
indicating both horizontal and vert
in line with the results already pres

A comparison of baseline readir
show whether the washout period

double-blind crossover trial comparing its with Raynaud's phenomenon (Kahan *et* tacks in two weeks. There was a one-week nt periods

o ($n = 10$)

(1) − (2)	$\frac{(1)+(2)}{2}$	P − N
4	14	−4
7	22.5	−7
−12	14	12
−7	40.5	7
−16	17	16
5	38.5	−5
16	44	−16
−1	10.5	1
−9	15.5	9
3	28.5	−3
−1.0 ($\bar{d}_1$)	24.5 ($\bar{a}_1$)	1.0
9.87	12.50	9.87

($n = 10$)

l) − (2)	$\frac{(1)+(2)}{2}$	P − N
6	15	6
8	8	8
9	41.5	9
−7	54.5	−7
26	15	26
11	23.5	11
7	47.5	7
32	30	32
12	24	−12
40	24	40
12.0 ($\bar{d}_2$)	28.3 ($\bar{a}_2$)	12.0
16.34	15.12	16.34

effect depends on the order in which the treatments were given. The latter is a more serious problem because it leads to a biased estimate of the treatment effect. (See also section 15.2.5.)

We can test the treatment effect by performing a one sample t test on all 20 within subject differences between the two treatments. Because the two crossover groups may not be the same size it is preferable to consider the average effect in the two periods, which is equivalent to performing a two sample t test to compare $\bar{d}_1$ and $\bar{d}_2$.

For this example the period effect and treatment–period interaction give $t = 1.82$ and $t = 0.613$ respectively, both on 18 degrees of freedom, giving $P = 0.09$ and $P = 0.55$. As neither is statistically significant we can go on to evaluate the treatment effect using a further two sample t test, which gives $t = 2.154$ on 18 degrees of freedom ($P = 0.045$). The number of attacks in two weeks on nicardipine was on average 6.5 fewer than during two weeks on placebo, with a 95% confidence interval from 0.18 to 12.82. Although statistically significant at the 5% level, the magnitude of the effect of nicardipine is uncertain, reflecting the small sample size.

A problem with the analysis of crossover trials is that the important test for a possible treatment–period interaction is noted for its lack of statistical power. The above analysis is a good example, because Table 15.5 shows that patients in group 1 did nearly as well on placebo as they had on nicardipine, suggesting a long-lasting 'carry-over' effect of the active drug. Patients in group 2 showed a big improvement when they changed from placebo to nicardipine. This apparent interaction is not nearly statistically significant. The data from period 1 taken alone suggest that the true benefit of nicardipine might well be rather greater than indicated by the overall results of the trial.

In contrast Ueshima *et al.* (1987) found a marginally significant ($0.05 < P < 0.10$) treatment–period interaction in a crossover trial to investigate the possible effect on blood pressure of reducing alcohol intake. They discarded the data from the second period.

A graphical approach is to produce a scatter plot of the difference between the two periods against the average of the two periods, using different symbols to identify the two groups (Clayton and Hills, 1987). Vertical separation of the two groups is an indication of a difference between the treatments. If there is no treatment–period interaction there should be no horizontal difference between the groups, and the data for the two groups should lie symmetrically either side of the line $y = 0$, as in Figure 15.3(a). Figure 15.3(b) shows such a plot for the nicardipine trial, indicating both horizontal and vertical differences between the two groups, in line with the results already presented.

A comparison of baseline readings taken at the start of each period can show whether the washout period was successful. For example, Table 15.6

shows baseline data from a randomized crossover trial comparing rifampicin with phenobarbitone for treatment of pruritus in biliary cirrhosis. It is clear that patients in the first group had less severe pruritis at the beginning of the second period than at the start of the study. Thus either

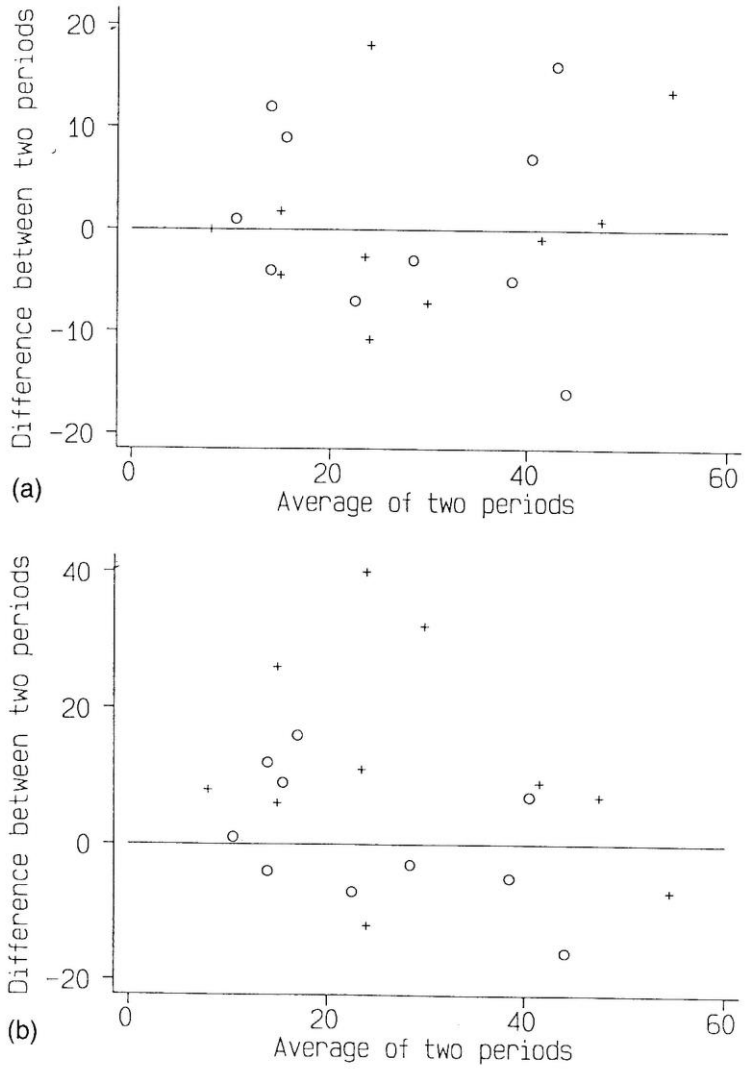


Figure 15.3 (a) Ideal plot of the difference between the responses in the two periods against the average in the two periods showing the symmetry of the responses in the two groups (shown as ○ and +). (b) Plot of difference between periods against average of two periods for patients receiving nicardipine followed by placebo (○) or placebo followed by nicardipine (+) (data from Kahan *et al.*,1987).

Table 15.6 Distribution of
(severe) before each peri
(Bachs *et al.*, 1989)

Group 1 (n = 12)
Before rifampicin
Before phenobarbitone*
Group 2 (n = 10)
Before phenobarbitone
Before rifampicin

*One patient dropped out afte

the pruritus had been improved b
trial was inappropriate, or the wa
is advantageous to incorporate ba
makes the analysis more complex.

Crossover trials are particular
withdrawal. If a patient withdra
included in the analysis because
The randomized groups are thus c
especially when these are more c
withdrawals it may be best to disc

In a report of a crossover trial
the trial are documented, with rea
the two randomized groups should
parallel group trials, most publish
this information.

15.5 INTERPRETATION OF

15.5.1 Single trials

In most cases the statistical analysi
with respect to the main outcome
or a Chi squared test. Interpretati
for one difficulty. Inference from
assumption that the trial participar
In most trials, however, particip
inclusion criteria, so extrapolation
not be warranted. For example,
such as beta-blocking drugs, are

Table 15.6 Distribution of pruritus scores, from 0 (mild) to 3 (severe) before each period in a two-period crossover trial (Bachs *et al.*, 1989)

	Pruritus score			
	0	1	2	3
Group 1 (n = 12)				
Before rifampicin	0	2	5	5
Before phenobarbitone*	3	3	1	4
Group 2 (n = 10)				
Before phenobarbitone	0	2	2	6
Before rifampicin	0	2	2	6

*One patient dropped out after period 1.

the pruritus had been improved by the first treatment, so that a crossover trial was inappropriate, or the washout period was too short. In general, it is advantageous to incorporate baseline readings into the analysis, but this makes the analysis more complex.

Crossover trials are particularly vulnerable to the effects of patient withdrawal. If a patient withdraws after the first period they cannot be included in the analysis because they never received the other treatment. The randomized groups are thus compromised when there are withdrawals, especially when these are more common in one group. If there are many withdrawals it may be best to discard the data from the second period.

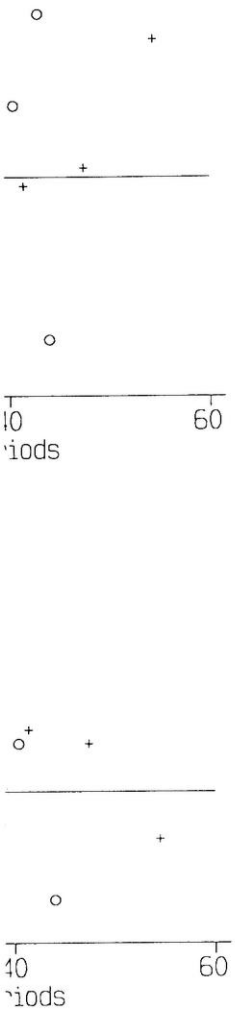
In a report of a crossover trial it is essential that any withdrawals from the trial are documented, with reasons. Also, the baseline characteristics of the two randomized groups should be described. Although this is routine in parallel group trials, most published reports of crossover trials do not give this information.

15.5 INTERPRETATION OF RESULTS

15.5.1 Single trials

In most cases the statistical analysis of a clinical trial will be simple, at least with respect to the main outcome measure, perhaps involving just a t test or a Chi squared test. Interpretation seems straightforward, therefore, but for one difficulty. Inference from a sample to a population relies on the assumption that the trial participants are representative of all such patients. In most trials, however, participants are selected to conform to certain inclusion criteria, so extrapolation of results to other types of patient may not be warranted. For example, most trials of anti-hypertensive agents, such as beta-blocking drugs, are carried out on middle-aged men. Is it

ssover trial comparing rifampi-
ruritus in biliary cirrhosis. It is
d less severe pruritis at the
start of the study. Thus either



tween the responses in the two
ds showing the symmetry of the
+). (b) Plot of difference between
ts receiving nicardipine followed by
+) (data from Kahan *et al.*,1987).