Crossover Trials with a Binary Response: A Powerful Method Despite the Carryover Effect

Ton J. M. Cleophas, MD, PhD, FICA, FACA, and Henk J. J. van Lier, MSc, PhD

The two-period crossover trial has the evident advantage that, by use of within-patient comparisons, the usual large between-patient variability is not used as a measure to compare treatments. A prerequisite, however, is that the order of the treatments does not substantially influence the outcome of the treatment. Crossover studies with a binary response (such as yes/no or present/absent), although widely used for initial screening of new compounds, have not previously been studied for such order effects. This study uses a mathematical model based on standard statistical tests to study to what extent such order effects, here identical to carryover effects, may reduce the power of detecting a treatment effect. It is concluded that, despite large carryover effects, the crossover study with a binary response remains a powerful method and that testing for carryover effects makes sense only if the null hypothesis of no treatment effect cannot be rejected.

The crossover design is widely used in clinical research, especially in instances where a limited number of patients is available for a study. The main advantage of within-patient comparisons in comparison to between-patient comparisons is that between-subject variability is not used in the comparisons. A prerequisite, however, for this type of trial is that the order of the treatments does not influence the outcome of the treatment. If the effect of the treatment administered in the first period carries over into the second period, then it may influence the measured response in the second period. This essentially means that only symptomatic treatments qualify for crossover comparisons and curative treatments do not.

Symptomatic treatments frequently have small curative effects, however, such as wound healing by vasodilators or, more recently, cardiac remodeling by afterload reduction. The treatment group that is treated with the effective compound first and with the less effective compound or placebo second is frequently biased by carryover effect from the first period to the second, whereas the alternative group that is treated in the reverse order is not so. For example, of 73 reports of crossover trials published in 1989–1990, only 6 reported the data of the separate periods. In 5 of them (83%) this very type of carryover effect was demonstrable. Such a mechanism may cause a severe underestimation of the treatment results, and this possibility should, therefore, be assessed in the analysis.

Most of the reports on the subject of order effects so far have addressed crossover studies with a quantitative rather than binary response. Although Hills and Armitage11 mentioned the tests of Gart12 and Prescott13 for crossover trials with a binary response in their overview of methods in crossover clinical trials and Fidler14 presented a model, little attention has been paid to these kinds of trials. A binary response is different from a quantitative response in that it generally does not answer what exactly can be expected in an individual. Rather it addresses whether or not a particular result has a predictive value, which of two treatments is better, or whether there is a treatment effect in the data. One might contend, therefore, that some undervaluation of a difference in binary data is not that important as long as it does not cause a type II error by in-
CROSSOVER TRIALS WITH BINARY RESPONSE

TABLE I

Example of a Crossover Design with a Binary Response

<table>
<thead>
<tr>
<th>Group</th>
<th>Period 1</th>
<th>Period 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Probability of Treatment Success</td>
<td>Probability of Treatment Success</td>
</tr>
<tr>
<td>I</td>
<td>A</td>
<td>pA</td>
</tr>
<tr>
<td>II</td>
<td>B</td>
<td>pb</td>
</tr>
</tbody>
</table>

*If in Group II treatment B has a carryover effect on the outcome of treatment A, pb changes to pc. If pb = pc, carryover effect is maximal.*

indicating no difference were there one. The main purpose of our analysis was to examine whether in crossover trials with a binary response a significant carryover effect does leave enough power in the data to demonstrate a treatment effect.

**ASESSMENT OF CARRYOVER AND TREATMENT EFFECT**

In a crossover trial with two treatments and two treatment periods, patients are assigned in a randomized fashion to two symmetric groups that are given treatments A and B in a different order (Table I). If groups are symmetric and the results are not influenced by the order of the treatments, the probabilities of treatment success in groups I and 2 should be virtually the same in each period for each treatment: pA being the probability of treatment success from treatment A and pb the probability of treatment success from treatment B (Table I). Results from the group that is treated with the less effective treatment or placebo after the more effective are in danger of being biased by carryover effect from the first period to the second.

Suppose treatment A is far less effective than treatment B (Table I). Then, if in group 2 treatment B has a carryover effect on the outcome of treatment A, the probability of treatment success changes from pb to pc. To detect a carryover effect we compare the outcomes of treatment A in group 1 to the outcomes of treatment A in group 2: pA versus pc, an unpaired comparison. The amount of carryover effect in group 2 is considered to be the difference between pc and pA. Carryover effect in group 1 (ineffective treatment period before effective treatment) is assumed to be negligible. Time effect is assumed to be negligible as well, because we study stable disease only.

It thus seems that neither a test for carryover effect in group 1 nor a test for time effects needs to be included in our assessment. Treatment effect is assessed by taking the two groups together, after which all outcomes of treatment A are compared with those of treatment B in a paired comparison. The assumption that carryover effect in group 1 is negligible implies that the test for carryover effect uses only half of the available data and might therefore be expected to be less sensitive. Sensitivity not only depends on sample size, however, but also on the size of differences and their variances.

**STATISTICAL MODEL FOR TESTING TREATMENT AND CARRYOVER EFFECTS**

We assume a unidirectional assessment where p is between 0.0 (no more symptoms) and 1.0 (100% of patients remain symptomatic despite treatment). When carryover effect is in the data, pA for group 2 becomes pc (Table I). The difference between pc and pA is considered to be the amount of carryover effect in the data. Fisher's exact test is used to determine whether pc is significantly different from pA. Using the program of Bavry,15 the values for pc that should yield a significant carryover effect in 80% of the trials (i.e., power = 80%) are determined. A number of patients between 10 and 25 is chosen for both groups, because many crossover trials have 20 to 50 patients. The values for pc are then used to determine whether enough power is left in the data to demonstrate a significant treatment effect in crossover trials with significant carryover effect and a binary response.

**TABLE II**

Power to Demonstrate a Treatment Effect in Spite of the Presence of a Significant Carryover Effect

<table>
<thead>
<tr>
<th>Total No. of Patients</th>
<th>2 x 10</th>
<th>2 x 15</th>
<th>2 x 20</th>
<th>2 x 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>pA 0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Values are power (%) of McNemar test for treatment effect (α = 0.05, p = 0) [pc value just yielding a significant test for carryover effect (α = 0.05, power = 80%)].
To test the treatment effect, all data for treatment A are taken together and compared with data for treatment B. The power of this test depends not only on the probabilities $p_A$ and $p_B$, but also on the correlation between the treatment responses. This correlation is expressed as $\rho = p_{A/B} - p_A$, where $p_{A/B}$ is the probability of treatment success with treatment A, given that treatment B was successful. When $\rho = 0$, treatments A and B act independently. When $p_B$ equals $p_C$, this would mean that the carryover effect in group 2 is not only significant but also maximal, given the amount of treatment effect. Considering this situation of maximal carryover effect, we calculated the power of detecting treatment effects. The power of the McNemar test, with $p_B$ being equal to $p_C$ and various values for $\rho$, was calculated according to Bavry.\textsuperscript{15}

\textbf{RESULTS}

\textbf{Calculation of $p_C$ Values Yielding a Significant Result for Carryover Effect}

For various numbers of patients and various values of $p_A$ (the probability of success with treatment A in period 1; Table I), the $p_C$ values (the probability of success with treatment A in period 2) are calculated so that a power of 80\% will yield a significant test for carryover effect ($p_A$ versus $p_C$; $\alpha = 0.05$). Table II shows that carryover effects (difference between $p_A$ and $p_C$) as large as 0.60, 0.50, 0.40, and 0.35 are required for a significant test. For $\alpha = 0.01$, these values are approximately 0.70, 0.60, 0.50, and 0.45. Using these $p_C$ values, we then calculated the probability of detecting a treatment effect (i.e., power of testing treatment effect). We report minimal values of power only, i.e., the situation where $p_B = p_C$. Whenever $p_B < p_C$, we would have an even better power for testing treatment effect.

\textbf{Power of Paired Comparison for Treatment Effect}

When the result of treatment B ($p_B$) is taken equal to the maximal values of $p_C$ and treatments A and B act independently ($\rho = 0$), the probability or power of detecting a treatment effect in the crossover situation when $n$ is between 20 and 50 is always more than 94\% (Table II). Usually, however, treatments A and B do not act independently. With a negative correlation between the two treatment modalities power is lost, and with a positive correlation it is augmented. Table III shows power values adjusted for different levels of $\rho$. With negative levels of $\rho$ and 20 patients, the power for detecting a treatment difference is not less than 74\%, which is approximately as large as that chosen for the test on carryover effect (80\%). When more patients are admitted to the trial this value will be \~90\%.

\textbf{EXAMPLES}

Suppose we have a negative crossover in which the probability of treatment success in group 2 $p_C$ (Table IV) may have changed from 0.8 into 0.2 due to carryover effect from the effective treatment B into the second period. Fisher’s exact test for demonstrating

\textbf{TABLE IV

<table>
<thead>
<tr>
<th>Example</th>
<th>Probability of Treatment Success</th>
<th>Probability of Treatment Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period 1</td>
<td></td>
<td>Period 2</td>
</tr>
<tr>
<td>Group</td>
<td>Treatment</td>
<td>Treatment</td>
</tr>
<tr>
<td>I (n = 10)</td>
<td>A $p_A = 0.8$</td>
<td>B $p_B = 0.2$</td>
</tr>
</tbody>
</table>
a carryover effect \( (p_A \text{ versus } p_C) \) is calculated according to

\[
\text{Point probability for carryover effect} = \frac{10! \times 10! \times 10!}{20! \times 2! \times 2! \times 2!} = 0.011
\]

The cumulative tail probability is calculated as

\[
0.011 + 0.003 + 0.007 = 0.021\text{ and is thus significant at a level of } \alpha = 0.021.
\]

If we perform a similar unpaired analysis of the first period to demonstrate a treatment effect, we likewise obtain a significant test at a level of \( \alpha = 0.021 \). Suppose carryover effect would be smaller, e.g., \( p_A = 0.8, p_B = 0.0, p_C = 0.2 \). Then the test for treatment effect would yield an even better result:

\[
\text{Point probability for carryover effect} = \frac{29! \times 10! \times 10!}{20! \times 2! \times 2! \times 10!} = 0.004
\]

The cumulative tail probability is calculated as

\[
0.004 + 0.001 + 0.003 = 0.008.
\]

Therefore, in crossovers with a binary response and a negative result, it does make sense to test for carryover effect by comparing the two periods with the less effective treatment modalities. If a significant test is demonstrated, we obviously will find a significant difference at a similar or even lower level of significance when taking the first period for estimating the difference between treatment A and B. Thus, it would seem appropriate for our purpose to disregard the data of the second period in this particular situation (although the second period might still provide interesting information).

**DISCUSSION**

The power of crossover studies is frequently reduced by carryover effect. This is particularly so when a group that is treated with an effective treatment first is then treated with an ineffective treatment or placebo second. In studies with a quantitative response, this very effect may cause severe underestimation of the treatment effect. Studies with a binary response are different from studies with a quantitative response, however, in that they are mostly designed to determine whether a treatment has any effect rather than the magnitude of that effect. One might contend, therefore, that underestimation in such studies is not important as long as the null hypothesis of no treatment effect doesn’t have to be erroneously accepted. We demonstrate that in crossover trials with a binary response and significant carryover effect, the power of testing the treatment effect remains substantial. This would imply that routinely testing for carryover effects in such studies is not necessary as long as the result of the treatment comparison is positive. When a study is negative it does make sense, however, to test for carryover effect by comparing \( p_A \) and \( p_C \) (Table I).

When \( p_A \) is significantly different from \( p_D \), we assume that there is a carryover effect in group 2. In this situation, a parallel-group analysis of period 1 \( (p_A \text{ versus } p_B) \) can effectively be used to demonstrate a treatment effect. This will provide a significant difference at a similar or even lower level of significance than the test for carryover effect. This occurs because \( p_B \) equals \( p_C \) when carryover effect is maximal. The difference between \( p_B \) and \( p_A \) will therefore be at least as large as the difference between \( p_C \) and \( p_A \), and may be larger. Therefore, no further test for treatment effect seems to be required for our purposes, and it seems appropriate that the results of the second period be disregarded.

Considering that the influence of carryover effects in crossover trials with a binary response may not be a problem, we should shift our standards for choosing this particular trial design and make use of its additional advantages more frequently. For example, this design is particularly powerful for study of rapid relief of symptoms in chronic disease in which the long-term condition of the patient remains fairly stable. This is because between-subject variability is not a factor in a within-subject comparison. Also, we can make use of positive correlations between the treatment modalities tested, because the statistical power of testing treatment comparisons with a positive correlation can be largely enhanced by within-subject comparisons. Further, none of the patients in the trial has to be treated throughout the trial with a less adequate dose or placebo, which is why a crossover design usually raises fewer ethical considerations than does a parallel-group design in which one group is treated with placebo or less adequate dosage throughout the trial. Also, there is the advantage that patients can express their own opinions about which treatment they prefer. This is particularly important with subjective variables such as pain scores. The crossover design also does not require a large study group, because within-subject comparison facilitates the recruitment procedure and reduces costs. Finally, a double-blind design cannot be effectively executed in self-controlled studies without some kind of crossover element.

We therefore conclude that crossover studies with a binary response and positive results do not have to be tested for carryover effects. If such studies have a
negative result, however, testing for carryover effect does make sense. If a carryover effect is demonstrated, the treatment results should be analyzed in the form of a parallel-group study using the data from the first period.

REFERENCES