Statistical Science
2004, Vol. 19, No. 4, 712–719
DOI 10.1214/088342304000000611

Nonparametric Approaches to the Analysis of Crossover Studies

Mary E. Putt and Vernon M. Chinchilli

Abstract. We illustrate nonparametric, and particularly rank-based analyses of crossover studies, designs in which each subject receives more than one treatment over time. Principles involved in using the Wilcoxon rank sum test in the simple two-period, two-treatment crossover are described through theory and example. We then extend the ideas to two-treatment designs with more than two periods and to three-treatment, three-period designs. When more than one nonparametric approach is available, we consider the issue of statistical power in choosing an appropriate test.

Key words and phrases: Crossover, clinical trials, nonparametrics, rank-based statistics.

1. INTRODUCTION

Crossover trials are repeated measures studies where subjects typically receive more than one treatment over time (Vonesh and Chinchilli, 1997; Senn, 2002). For example, in the 2×2 or $AB : BA$ design, subjects receive either treatment A followed by B or B followed by A. Crossover trials are efficient since estimated treatment effects are based, either wholly or in large part, on within-subject contrasts. This eliminates, or reduces, the contribution of the between-subject component of the variance to the estimated treatment effect. Crossover trials are of interest when either financial resources or subject availability limits study size (e.g., Lagakos, 2003). With small samples, nonparametric approaches are appealing in principle because of the difficulties in verifying normality and because large-sample properties of parametric statistics may not hold. Nonparametric approaches are also appropriate in larger crossover studies if the data appear nonnormal.

Mary E. Putt is Assistant Professor, Department of Biostatistics and Clinical Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: mputt@cceb.upenn.edu). Vernon M. Chinchilli is Professor, Department of Health Evaluation Sciences, Hershey Medical School, Pennsylvania State University, Hershey, Pennsylvania 17033, USA (e-mail: vchinchi@psu.edu).

The issue of carryover, the lingering effect of a treatment from one period into the subsequent period(s), often dominates consideration of crossover trials (e.g., Senn, 2002; Freeman, 1989). In practice, washout periods are sometimes used to reduce or eliminate potential carryover effects. By lengthening the amount of time that a subject is on study, a washout potentially causes subject attrition and missing data (Correa and Bellavance, 2001). While some investigators have developed tests for the presence of carryover, as well as other nuisance parameters, this article focuses on nonparametric tests for the treatment effect, and largely ignores carryover effects. This emphasis is based on our experience with clinical investigations where the analytic focus is rarely on nuisance parameters such as carryover. We consider carryover in the study design so as to eliminate or minimize its impact on the study conclusions (Chinchilli and Esinhart, 1996; Putt and Ravina, 2002).

Rank-based tests use linear combinations of ranked outcomes, and are suitable for continuous data; if ties occur, the variance is adjusted (Koch, 1972; Hollander and Wolfe, 1999). Under the null hypothesis, the distribution of the rank-based statistics is distribution-free or nonparametric (Hettmansperger, 1991). Under the alternative hypothesis, power depends on the distribution. Asymptotic theory as well as simulation studies suggest little loss in efficiency of the rank-based tests when the data are normal and possible gains when the distribution departs from normality (Hettmansperger, 1991; Öhrvik, 1998; Correa and Bellavance, 2001).

Here we illustrate basic theory and practical application of several rank-based and permutation methods in crossover studies. This review updates Tudor and Koch (1994) who discuss nonparametric approaches to ordinal and censored data, topics not covered here. Section 2 presents a nonparametric approach to the simplest crossover, the 2×2 design and extends the principles to two-treatment designs with more than two periods. Section 3 describes three-period designs. The examples we will present were analyzed using R v. 1.70 (*www.r-project.org*). Datasets, programs and comments on the software are available at *http://www.cceb.upenn.edu/main/people/putt.html.*

2. THE 2 × 2 DESIGN

The 2×2 crossover design contains two treatment sequences and two periods (Table 1). Subjects randomized to sequence 1 (AB) receive treatment A in the first period followed by B in the second period; in sequence 2 (BA), subjects receive treatment B in the first period followed by A in the second period. An optional washout period may be used.

Example. We consider a clinical trial that compares nasal steroids (treatment A) and placebo (treatment B) on a measure of daytime sleepiness in patients who suffer from allergic rhinitis. The data in Table 2 are from the 2×2 portion of a study design which combines a 2×2 crossover and a parallel repeated measures study. Patients (five per sequence) were randomized to either treatment A or B and received self-administered treatments twice daily for 8 weeks (Craig et al., 1998; Putt and Chinchilli, 2000). At week 8 each patient crossed over to the other treatment (without washout). Each patient maintained a daily diary rating several aspects of daytime fatigue, including improved daytime sleepiness (IDS) on a scale of 0 (worst) to 4 (best). The average IDS over the final week in each 8-week treatment period was used in the analysis. The data for one patient with missing IDS for the second period were omitted. In this study, we expect limited or null carryover because of the lengthy period between measurements periods. Note that although the daily records were ordinal, we anticipated that averaging over a week would yield data on a continuous scale.

TABLE 1
Treatment assignments in the 2×2 *crossover*

Sequence	**Period**		
	1	Washout	2
1 (AB)	A	Optional	B
2 (BA)	B	Optional	A

Statistical model and approach. Let Y_{ijkl} be the outcome for the jth subject ($j = 1, \ldots, n_i$) from the ith sequence ($i = 1, 2$) on the kth treatment ($k = A, B$) in the lth period ($l = 1, 2$). Then

$$Y_{ijkl} = \mu_k + \pi_l + \lambda_{k'_{l-1}} + \varepsilon_{ijkl}, \tag{1}$$

where μ_k is the mean effect for the kth treatment, π_l is the mean added effect of the lth period, $\lambda_{k'_{l-1}}$ is the mean added carryover of the k'th treatment administered in the $(l-1)$st period into the lth period ($\lambda_{k'_0} = 0$) and ε_{ijkl} is a random error term. Subjects are independent with $E(\varepsilon_{ijkl}) = 0$, $\mathrm{Var}(\varepsilon_{ijkl}) = \sigma^2$ and $\mathrm{Cov}(\varepsilon_{ijkl}, \varepsilon_{ijk'l'}) = \rho\sigma^2$, where ρ is the correlation coefficient.

For the 2×2 study, Table 3 shows the expectation $E(\cdot)$ of the Y_{ijkl} for each sequence/period combination and the contrast for each subject, Y^*_{ij}, between outcomes for the first and second periods, that is,

$$Y^*_{ij} = Y_{ijk1} - Y_{ijk'2}.$$

Each Y^*_{ij} has variance $2\sigma^2(1-\rho)$. For treatment difference $\mu_{\mathrm{D}}^{(AB)} = \mu_A - \mu_B$ the null hypothesis is,

$$H_0 : \mu_{\mathrm{D}}^{(AB)} - \tfrac{1}{2}(\lambda_A - \lambda_B) = 0. \tag{2}$$

Under H_0, $E(Y^*_{ij})$ is identical for each subject. If carryover from the two treatments is identical, then H_0 tests whether the treatment means are identical. Moreover, if equality of treatments implies equality of carryover (i.e., $\mu_{\mathrm{D}}^{(AB)} = 0$ implies $\lambda_A = \lambda_B$), then the test is valid under the null hypothesis. Our alternative is

$$H_1 : \mu_{\mathrm{D}}^{(AB)} - \tfrac{1}{2}(\lambda_A - \lambda_B) \neq 0.$$

Under H_1, carryover effects in the same direction as the treatment effect (e.g., if $\mu_A > \mu_B$, then $\lambda_A > \lambda_B$) reduce the test's power (Öhrvik, 1998; Putt and Ravina, 2002).

For a nonparametric analysis the null hypothesis may be stated in terms of equality of cumulative distribution functions, $F_{Y^*_{ij}}(\cdot)$'s, of the Y^*_{ij}, that is,

$$H_0 : F_{Y^*_{1j}}(\cdot) = F_{Y^*_{2j}}(\cdot),$$

and this leads to a more general interpretation of the location parameters from (1) and (2). For example, we may consider differences in median rather than mean treatment effects.

To construct the test, we pool Y^*_{ij}'s from both sequences, assign a rank to each observation, $D_{ij}(Y^*_{ij})$,

TABLE 2
Mean weekly IDS score by period and difference between periods for individual patients on the AB and BA sequences

	***AB* sequence**			***BA* sequence**		
Patient	**Period 1**	**Period 2**	**Difference**	**Period 1**	**Period 2**	**Difference**
1	2.00	1.29	0.71	4.00	4.00	0.00
2	1.83	0	1.83	0	0.86	−0.86
3	0	0	0	0	0	0
4	0.89	NA	NA	1.14	2.14	−1.00
5	3.00	3.00	0	0	2.29	−2.29

NOTE: Data from Putt and Chinchilli (2000). Reprinted with permission of the *Journal of the American Statistical Association*.

and sum the ranks for the (arbitrary) first sequence, that is, $R_1 = \sum_{j=1}^{n_1} D_{1j}(Y_{1j}^*)$ (Koch, 1972; Tudor and Koch, 1994). In the absence of tied observations, the Wilcoxon rank sum (W^*) statistic is

$$W^* = \frac{\sqrt{12}(R_1 - n_1(n_1+n_2+1)/2)}{\sqrt{n_1 \cdot n_2(n_1+n_2+1)}}.$$

If observations are tied, the average of the potential ranks of the tied observations is used in place of the ranks, and the variance is adjusted (Hollander and Wolfe, 1999). Under the null hypothesis in (2), W^* is asymptotically distributed $(0, 1)$, that is, normal with mean 0 and variance 1. More formally, if as $n_1 + n_2 \to \infty$, $n_1/(n_1+n_2) \to \delta$ $(0 < \delta < 1)$, then $W^* \xrightarrow{D} N(0, 1)$ (Hettmansperger, 1991), where the notation $\xrightarrow{D}$ indicates convergence in distribution. With small sample sizes, exact p values are computed using permutation. Under (2), sequence assignments are exchangeable among subjects. To compute the permutation distribution, consider all $(n_1 + n_2)!/n_1!n_2!$ assignments of subjects to the two sequences. We recompute R_1 for each assignment; for a two-sided test, the exact p value is twice the proportion of permuted R_1 that are as large as or larger than the observed R_1 (Hollander and Wolfe, 1999; Good, 2000).

Finally, Hodges–Lehmann (HL) provides a robust estimate of the treatment effect. Note that the difference of each pair of Y_{ij}^*'s from the two sequences is an unbiased estimate of $2(\mu_A - \mu_B) - (\lambda_A - \lambda_B)$. The HL estimate is one half the median of all pairwise differences of the Y_{ij}^*'s, that is,

$$\frac{1}{2} \cdot \left[\text{Median}_{j_1 \le j_2}(Y_{1j_1}^* - Y_{2j_2}^*)\right], \tag{3}$$

where j_1 and j_2 index the subjects in sequences 1 and 2, respectively. Similarly, an exact confidence interval can be obtained from the quantiles of the pairwise differences (Hettmansperger, 1991).

TABLE 3
Expectations for 2×2 *crossover by period*

Sequence	**Period 1**	**Period 2**	**Contrast** (Y_{ij}^*)
(1) $A : B$	$\mu_A + \pi_1$	$\mu_B + \pi_2 + \lambda_A$	$\mu_{\mathrm{D}}^{(AB)} + (\pi_1 - \pi_2) - \lambda_A$
(2) $B : A$	$\mu_B + \pi_1$	$\mu_A + \pi_2 + \lambda_B$	$-\mu_{\mathrm{D}}^{(AB)} + (\pi_1 - \pi_2) - \lambda_B$

Results for the example. Four patients had no difference in IDS between periods, while the remaining five had better IDS on steroids compared to placebo (Table 2). Accounting for ties, the observed W^* yields a two-sided asymptotic p value of 0.0407 and an exact p value of 0.0950. If the data contained no ties, the smallest possible p value for a two-sided test of these data would be 0.0159 (2 out of 126 possible permutations). However, the ranks of the four patients who show no difference between periods are set at the midrank of the four observations (5.5 here), yielding 16 unique values of the test statistic. The exact p value was the smallest value possible for the observed data. Here, the power of the permutation test was restricted by the limited number of unique observations. Lastly the HL estimate indicated that steroids improved daytime sleepiness by 0.6425 units with a 95% exact confidence interval of (0.0, 1.794).

Comparison with standard approach. For the 2×2 crossover, the power of W^* for data that are not normal may exceed that of the standard approach, the two-sample t test (T^*). Asymptotic theory suggests the efficiency of W^* relative to T^* in the vicinity of the null hypothesis is 96% when the data are normally distributed (Hettmansperger, 1991). $T^*$is robust in the

TABLE 4
Components of the $AAB : BBA$ and $AABB : BBAA$ designs needed for a nonparametric analysis

Sequence	Contrast (Y^*_{ijk})	Expectation
$AAB : BBA$		
(1) AAB	$\frac{1}{2}\sum_{l=1}^{2} Y_{ijkl} - Y_{ijk3}$	$\mu_{\mathrm{D}}^{(AB)} + \frac{1}{2}(\pi_1 + \pi_2) - \pi_3 - \lambda_A$
(2) BBA	Same as sequence 1	$-\mu_{\mathrm{D}}^{(AB)} + \frac{1}{2}(\pi_1 + \pi_2) - \pi_3 - \lambda_B$
$AABB : BBAA$		
(1) $AABB$	$\frac{1}{2}(\sum_{l=1}^{2} Y_{ijkl} - \sum_{l=3}^{4} Y_{ijkl})$	$\mu_{\mathrm{D}}^{(AB)} + \frac{1}{2}(\sum_{l=1}^{2}\pi_l - \sum_{l=3}^{4}\pi_l) - \frac{1}{2}\lambda_A$
(2) $BBAA$	Same as sequence 1	$-\mu_{\mathrm{D}}^{(AB)} + \frac{1}{2}(\sum_{l=1}^{2}\pi_l - \sum_{l=3}^{4}\pi_l) - \frac{1}{2}\lambda_B$

sense that its level is conservative when the data are nonnormal (e.g., Everitt, 1979). However, W^* may be more efficient than T^* under H_1. For example, in a mixture of normals with identical locations but different variances (i.e., a contaminated normal), efficiency was 20% higher for W^* even when the contamination was only 5%. Note, however, that if a permutation test is used, the level and power of W^* and T^* are identical (Tudor and Koch, 1994).

Two-treatment multiperiod designs. Two-treatment designs with more than two periods, for example, the three-period $AAB : BBA$ or four-period $AABB : BBAA$, generally have higher efficiency than the 2×2 design, although the potential for missing data increases with the length of time each subject is on study (Carriere, 1994). Nonparametric analyses use the basic approach in Section 2 with components in Table 4. For Table 4, we assumed null carryover from a treatment into itself (e.g., carryover from the first to the second period receiving A). The null hypotheses for two-period and three-period designs are identical [equation 2]. For the four-period design, the null hypotheses is $H_0 : \mu_A - \mu_B - \frac{1}{4}(\lambda_A - \lambda_B) = 0$.

3. THREE-TREATMENT DESIGN

In this section we illustrate an "aligned" rank-based test for a three-treatment, three-period design. The principles extend to designs with more treatments. With more than two treatments, designs based on Williams squares are recommended when period and carryover effects are possible (Bellavance and Tardiff, 1995; Öhrvik, 1998). Table 5 illustrates this type of design for the sequences $ABC : CAB : BCA : ACB : BAC : CBA$. The model from equation (1) generalizes with Y_{ijkl} the outcome for the jth subject ($j = 1, \ldots, n_i$) from the ith sequence ($i = 1, 2, 3, 4, 5, 6$) on the kth treatment ($k = A, B, C$) in the lth period ($l = 1, 2, 3$). Within each Williams square, sequences are chosen such that every treatment occurs in every period and precedes every other treatment twice. Our analysis initially assumes "complete" Williams squares, that is, equal numbers of subjects per sequence ($n_i = n$ for $i = 1, \ldots, 6$).

Aligned observations Y^*_{ijk} are based on subtracting an estimate of the period effect from each Y_{ijkl}, that is,

$$Y^*_{ijk} = Y_{ijkl} - \bar{Y}_{\cdots l},$$

where $\bar{Y}_{\cdots l}$ is a function of the observations in the lth period. For example, $\bar{Y}_{\cdots l}$ may be the mean, $\frac{1}{6n} \cdot \sum_{i=1}^{6}\sum_{j=1}^{n} Y_{ijkl}$, the median for the lth period or the HL estimate, that is, $\mathrm{Median}_{i \le i', j \le j', l = l'} \frac{1}{2}(Y_{ijkl} + Y_{i'j'k'l'})$. For data with a symmetric distribution, the estimates have the same expectation, that is,

$$(4) \quad E(\bar{Y}_{\cdots l}) = \begin{cases} \pi_l + \bar{\mu}_{\cdot}, & \text{for } l = 1, \\ \pi_l + \bar{\mu}_{\cdot} + \bar{\lambda}_{\cdot}, & \text{for } l > 1, \end{cases}$$

where $\bar{\mu}_{\cdot} = \frac{1}{3}\sum_{k=1}^{3}\mu_k$ and $\bar{\lambda}_{\cdot} = \frac{1}{3}(\lambda_A + \lambda_B + \lambda_C)$. Note that carryover is not altered by the period in which it occurs, or the treatment that occurs in the period receiving the carryover; for example, carryover from A into B is the same as carryover from A into C. Next generate the within-subject contrasts, $D_{ijkk'}, k \neq k'$, for the three possible treatment pairs $kk' \in \mathcal{K}$ for $\mathcal{K} = \{AB, AC, BC\}$, that is,

$$D_{ijkk'} = Y^*_{ijk} - Y^*_{ijk'}.$$

Expectations of these contrasts appear in Table 5 for the case of equal number of observations per sequence. Here we see that while alignment removes period effects, carryover effects remain.

Now consider the null hypothesis

$$(5) \quad H_0 : \mu_A - \tfrac{1}{3}\lambda_A = \mu_B - \tfrac{1}{3}\lambda_B = \mu_C - \tfrac{1}{3}\lambda_C$$

versus the alternative

$$H_1 : \mu_k - \tfrac{1}{3}\lambda_k \neq \mu_{k'} - \tfrac{1}{3}\lambda_{k'} \quad \text{for some } k \neq k'.$$

TABLE 5
Expectations of within-subject contrasts for complete Williams squares. Mean treatment differences are denoted $\mu_{\mathrm{D}}^{(kk')}$ where k and k' index pairs of treatments

Sequence	*A* vs. *B*	*A* vs. *C*	*B* vs. *C*
ABC	$\mu_{\mathrm{D}}^{(AB)} + (\bar{\lambda} - \lambda_A)$	$\mu_{\mathrm{D}}^{(AC)} + (\bar{\lambda} - \lambda_B)$	$\mu_{\mathrm{D}}^{(BC)} + (\lambda_A - \lambda_B)$
BCA	$\mu_{\mathrm{D}}^{(AB)} - (\bar{\lambda} - \lambda_C)$	$\mu_{\mathrm{D}}^{(AC)} + (\lambda_C - \lambda_B)$	$\mu_{\mathrm{D}}^{(BC)} + (\bar{\lambda} - \lambda_B)$
CAB	$\mu_{\mathrm{D}}^{(AB)} + (\lambda_C - \lambda_A)$	$\mu_{\mathrm{D}}^{(AC)} - (\bar{\lambda} - \lambda_C)$	$\mu_{\mathrm{D}}^{(BC)} - (\bar{\lambda} - \lambda_A)$
CBA	$\mu_{\mathrm{D}}^{(AB)} + (\lambda_B - \lambda_C)$	$\mu_{\mathrm{D}}^{(AC)} - (\bar{\lambda} - \lambda_B)$	$\mu_{\mathrm{D}}^{(BC)} - (\bar{\lambda} - \lambda_C)$
ACB	$\mu_{\mathrm{D}}^{(AB)} + (\bar{\lambda} - \lambda_C)$	$\mu_{\mathrm{D}}^{(AC)} + (\bar{\lambda} - \lambda_A)$	$\mu_{\mathrm{D}}^{(BC)} + (\lambda_C - \lambda_A)$
BAC	$\mu_{\mathrm{D}}^{(AB)} - (\bar{\lambda} - \lambda_B)$	$\mu_{\mathrm{D}}^{(AC)} + (\lambda_B - \lambda_A)$	$\mu_{\mathrm{D}}^{(BC)} + (\bar{\lambda} - \lambda_A)$

Under the null hypothesis in (5) the distribution of contrasts for each of the three sets of treatment pairs, and hence the distribution of the $D_{ijkk'}$ considered as a whole, is centered around zero. The null hypothesis of equality of treatment and carryover effects,

$$H_0 : \mu_A = \mu_B = \mu_C \quad \text{and} \quad \lambda_A = \lambda_B = \lambda_C,$$

is a special case of (5), and one that may be more intuitive to consider.

To construct the test statistic we pool the $N = 3\sum_{i=1}^{6} n_i = 18n$ contrasts and assign a rank $R(D'_{ijkk'})$ to each of the pooled observations in the sample. The sample is split into two groups that correspond to $D_{ijkk'}$ positive or negative and, for each of the positive and negative groups, the sum of the ranks is computed within each treatment pair, that is,

$$R^{+}_{kk'} = \sum_{i=1}^{6} \sum_{j=1}^{n_i} I(D_{ijkk'} > 0) R(D_{ijkk'})$$

and

$$R^{-}_{kk'} = \sum_{i=1}^{6} \sum_{j=1}^{n_i} I(D_{ijkk'} \le 0) R(D_{ijkk'}),$$

where $I(\cdot)$ is the indicator function. Under the null hypothesis in (5), the $R(D_{ijkk'})$ are symmetrically distributed around zero. The test statistic, referred to subsequently as Öhrvik's Q, is

$$(6) \quad Q = \frac{12}{3(N+1)(2N+1)} \sum_{kk' \in \mathcal{K}} \frac{(R^{+}_{kk'} - R^{-}_{kk'})^2}{N_{kk'}},$$

where $N_{kk'} = \sum_{i=1}^{6} n_i = 6n$ and as before $\mathcal{K} = \{AB, AC, BC\}$. Under H_0, $E(Q) = 2$ and, asymptotically, if $\lim_{N\to\infty} N_{kk'}/N$ exists and is positive, Q has a chi-squared distribution with 2 degrees of freedom. Alternatively, the exact p value is determined by randomly assigning subjects to sequences, recomputing Q and comparing the observed Q to its permutation distribution. If Q suggests a treatment effect, then the test of interest becomes

$$(7) \qquad H_0 : \mu_{\mathrm{D}}^{(kk')} - \tfrac{1}{3}(\lambda_k - \lambda_{k'}) = 0$$

and a test for the individual treatment contrasts uses

$$(8) \quad Q_{kk'} = \frac{(R^{+}_{kk'} - R^{-}_{kk'})}{N_{kk'}} \Bigg/ \sqrt{\frac{(N+1)(2N+1)}{6}},$$

which has a limiting Normal$(0, 1)$ distribution under the null hypothesis in (7).

While Q is most intuitive when observations are aligned to remove period effects, the test remains valid if the observations are not aligned, as long as the Williams squares are complete and carryover is equal among all treatments. Here, period effects balance; for example, within the A versus B contrast, $(\bar{\lambda}_{\cdot} - \lambda_A)$ is balanced by $-(\bar{\lambda}_{\cdot} + \lambda_c)$ as long as $\lambda_A = \lambda_C$.

Example. Milk production was compared in 18 cows randomized to three diets (A, roughage; B, limited grain; C, full grain) (Bellavance and Tardiff, 1995). During the experiment, milk production decreased by roughly one-third from around 1650 units (pounds per 6 weeks) in period 1 to just under 1200 units in period 3, with all three location estimates yielding similar results (Table 6). The left-hand column of Figure 1 shows contrasts for observations without alignment (top) and for those aligned with the HL estimate of the period effect (lower). For each contrast, aligning the observations reduces the spread around the median. For these data, alignment reduced the number of positive observations and decreased the amount of overlap among the ranks for the positive and negative observations (middle and right columns of Figure 1). The plot suggests that the test statistics should have

TABLE 6
Mean milk production by period (pounds per 6 weeks) and Öhrvik's Q for different methods of aligning the observations

Alignment method	Period			Q (p value)
	1	2	3	
None	NA	NA	NA	7.56 (0.023)
Mean	1655	1435	1180	26.04 (< 0.0001)
Median	1648	1332	1192	24.03 (< 0.0001)
Hodges–Lehmann	1648	1416	1182	26.44 (< 0.0001)

greater power when the observations are aligned. Table 6 confirms that Q is largest when the statistic is aligned using the mean or the HL estimate, slightly smaller for the median and dramatically smaller for the statistic based on unaligned ranks. Note that these data display substantial differences in outcome by period. The performance of the aligned and unaligned statistics may be more similar when period effects are small.

Since the overall test statistic was (highly) significant, we examined individual contrasts (Table 7). Tests based on different methods of alignment gave similar results with p values that are much smaller than the test using the unaligned observations. Finally, the HL estimates of differences in milk production using the observations aligned with HL were -165 (AB), -273 (AC) and -107 (BC) pounds per 6 week interval.

Alternative nonparametric approaches. Alternatively, observations may be doubly aligned within each Williams square (Bellavance and Tardiff, 1995). Unlike Öhrvik's Q, where the rankings are made for observations pooled across blocks, the Bellavance and Tardiff (BT) approach uses rankings made within individual blocks and subsequently pooled across blocks to form the test statistic. For the milk production data, the asymptotic p value of the test for the hypothesis in (5) is 0.065, suggesting that for this example, BT has substantially lower power than Öhrvik's Q [equation (6)].

Senn's (2002) approach in the three-treatment, three-period design pools tests for differences between pairs of treatments. Sequences are paired by matching the periods in which the treatment pairs occur; for example, for the AB contrast we form three sequence pairs (ABC with BAC), (ACB, BCA) and (CAB, CBA), or strata, considering only periods that contain the A and B treatments. This essentially leaves a series of 2×2 crossovers. Within each stratum, we construct W^* as described in Section 2 and pool results. Under the assumption of null carryover, this test is valid. However, for our example the p value for the test of the AB contrast was 0.0943, suggesting that the method may sometimes have substantially lower power than Öhrvik's Q.

Comparison with standard approaches. The standard alternative is an analysis of variance decomposition for the parameters of interest (Bellavance and

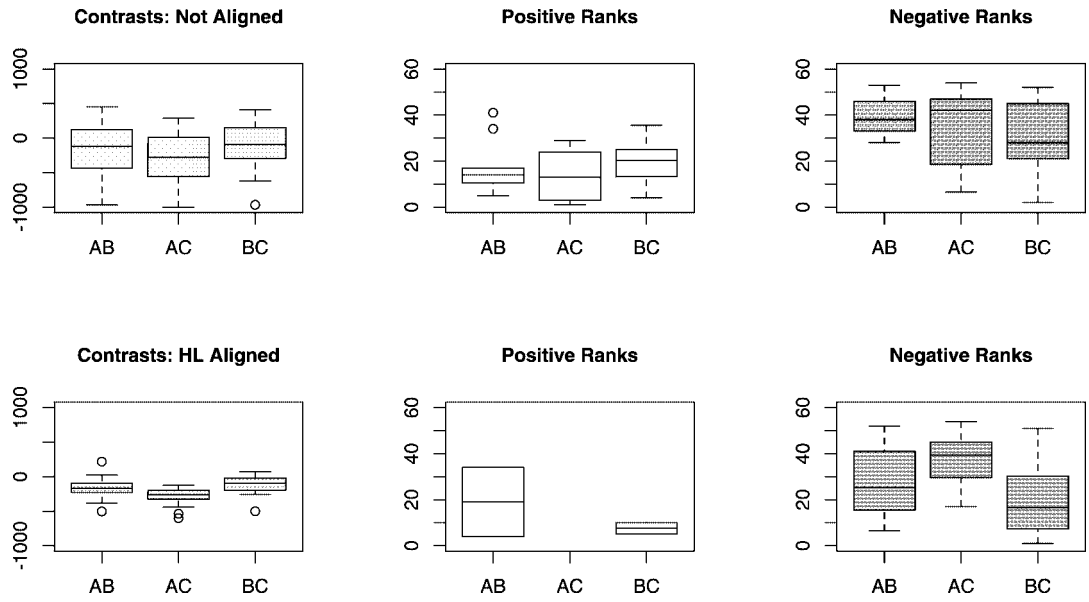

FIG. 1. *Boxplots showing effect of not aligning* (top row) *and aligning using HL* (bottom row): *observations on the contrasts* (left-hand side panel), *and the ranks of the positive* (middle panel) *and the negative* (right-hand side panel) *contrasts*.

TABLE 7
Test statistics with p values in parentheses for testing individual contrasts

Alignment method	$Q_{kk'}$ (p value)		
	AB	AC	BC
None	1.49 (0.1349)	2.81 (0.0099)	1.08 (0.2812)
Median	2.93 (0.0033)	4.76 (<0.0001)	2.19 (0.0286)
Mean	2.99 (0.0029)	5.05 (<0.0001)	2.16 (0.0306)
HL	2.99 (0.0022)	5.08 (<0.0001)	2.20 (0.0273)

Tardiff, 1995). This approach is based on variance ratios that have an F distribution if the data are normally distributed and the covariance structure of the repeated measures within individuals is exchangeable. Under more general covariance structures a modified F test (mF) is needed to maintain valid Type I error rates (Bellavance, Tardif and Stephens, 1996; Correa and Bellavance, 2001).

Under a normal shift model, the relative efficiency of Öhrvik's Q to its corresponding parametric test is asymptotically equivalent to that of the Wilcoxon signed-rank test to the paired t test (Öhrvik, 1998). The loss in efficiency of Öhrvik's Q relative to the parametric test when the data are normal is thus minor. Correa and Bellavance (2001) carried out simulation studies of Öhrvik's Q, mF and BT under three covariance structures using multivariate normal and gamma distributions. The covariance matrices included an exchangeable structure, as originally specified by Öhrvik (1998), as well as sphericity and an unstructured form. Under multivariate normality and for carryover effects equal to 50% of treatment effects, Öhrvik's Q had valid Type I error rates and power that was similar to or higher than mF. Under the multivariate gamma, both Öhrvik's Q and mF were somewhat anticonservative (empirical Type I error rates of up to 6.9% for Öhrvik's Q and up to 8.3% for mF for a nominal Type I error rate of 5%). However, Öhrvik's Q had substantially higher power than mF. We expect that if the permutation distribution were used, the nominal and empirical Type I error rates would be more similar for Öhrvik's Q.

Of all three tests, BT was the only one to maintain strictly valid Type I error rates under both the multivariate normal and the multivariate gamma distributions. Unlike Öhrvik's Q, the power of this test is not altered by carryover. However, for reasonable carryover levels (e.g., 50% of the treatment effect), BT had substantially lower power than Öhrvik's Q (e.g., 76% vs. 32%). Because of this we hesitate to recommend BT, despite its performance under the null hypothesis.

Comments. In agricultural or laboratory studies, Williams square designs are feasible to complete. In clinical research, patient-related issues (e.g., recruitment, attrition, early-stopping rules) make it difficult to achieve complete Williams squares. If the design is imbalanced in the sense that some sequences have more patients than others or individuals have missing data, then the procedures described above are not strictly appropriate. For example, $\bar{Y}_{\cdots l}$ is not necessarily unbiased for the period effects shown in (4). An alternative in this case is to use a U-statistic such as

$$\bar{Y}_{\cdots l} = \frac{1}{\prod_{i=1}^{6} n_i} \cdot \sum_{j_1=1}^{n_1} \cdots \sum_{j_6=1}^{n_6} \left(Y_{1j_1kl} + Y_{2j_2kl} + \cdots + Y_{6j_6kl}\right)$$

(or a more robust generalized L statistic) to align the observations (Putt and Chinchilli, 2000). These $\bar{Y}_{\cdots l}$ have expectation shown in (4), and if carryover is assumed null, then the tests in (6) and (8) are valid. However, in the presence of carryover effects Öhrvik's Q (6) and $Q_{kk'}$ (8) test hypotheses that are somewhat different from those shown in (5) and (7). For example, let $n_{\min} = \min_i n_i$ and $n_i^* = n_i - n_{\min}$ and supose that $\mu_{\mathrm{D}}^{(AB)}$ is of interest. The distribution of the sample of D_{ijAB}'s has expectation

$$\begin{aligned} E^{AB} &= \frac{1}{\sum_{i=1}^{s} n_i} \sum_{i=1}^{6} \sum_{j=1}^{n_i} E(D_{ijAB}) \\ &= \mu_{\mathrm{D}}^{(AB)} - \frac{2n_{\min}(\lambda_A - \lambda_B)}{(6n_{\min} + \sum_{i=1}^{s} n_i^*)} \\ &\quad + \frac{(\bar{\lambda}k_1 - \lambda_A k_2 + \lambda_B k_3 + \lambda_C k_4)}{(6n_{\min} + \sum_{i=1}^{s} n_i^*)} \end{aligned}$$

for $k_1 = n_1^* - n_2^* + n_5^* - n_6^*$, $k_2 = n_1^* + n_3^*$, $k_3 = n_4^* + n_6^*$, $k_4 = n_2^* - n_4^* + n_3^*$. We would not necessarily expect the distribution of signed ranks to be centered around zero, even if $\mu_{\mathrm{D}}^{(AB)} = \lambda_A - \lambda_B = 0$. However, if the sample size is large, and the degree of imbalance is small, for example, $\sum_{i=1}^{6} n_i^* \ll 6n_{\min}$, the distribution will be asymptotically centered around zero.

Senn (2002) argues that the simple carryover model used here is unrealistic in clinical studies and, in particular, that the treatment in the period in which the carryover occurs should be considered in the model. For example, consider a trial testing a placebo and two

different doses of an active compound. Carryover from the high-dose period is likely to be larger in the placebo period than in the low-dose period. To our knowledege, the impact of this type of carryover has not been examined for the test described here.

ACKNOWLEDGMENTS

We thank the reviewers for their thoughtful comments.

REFERENCES

BELLAVANCE, F. and TARDIF, S. (1995). A nonparametric approach to the analysis of three-treatment three-period crossover designs. *Biometrika* **82** 865–875.

BELLAVANCE, F., TARDIF, S. and STEPHENS, M. A. (1996). Tests for the analysis of variance of crossover designs with correlated errors. *Biometrics* **52** 607–612.

CARRIERE, K. C. (1994). Crossover designs for clinical trials. *Statistics in Medicine* **13** 1063–1069.

CHINCHILLI, V. M. and ESINHART, J. D. (1996). Design and analysis of intrasubject variability in cross-over experiments. *Statistics in Medicine* **15** 1619–1634.

CORREA, J. A. and BELLAVANCE, F. (2001). Power comparison of robust approximate and nonparametric tests for the analysis of cross-over trials. *Statistics in Medicine* **20** 1185–1196.

CRAIG, T. J., TEETS, S., LEHMAN, E. B., CHINCHILLI,V. M. and ZWILLICH, C. (1998). Nasal congestion secondary to allergic rhinitis as a cause of sleep disturbance and daytime fatigue and the response to topical nasal corticosteroids. *J. Allergy and Clinical Immunology* **101** 633–637.

EVERITT, B. S. (1979). A Monte Carlo investigation of the robustness of Hotelling's one- and two-sample T^2 tests. *J. Amer. Statist. Assoc.* **74** 48–51.

FREEMAN, P. R. (1989). The performance of the two-stage analysis of two-treatment, two-period cross-over trials. *Statistics in Medicine* **8** 1421–1432.

GOOD, P. I. (2000). *Permutation Tests*: *A Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd ed. Springer, New York.

HETTMANSPERGER, T. (1991). *Statistical Inference Based on Ranks*. Krieger, Malabar, FL.

HOLLANDER, M. and WOLFE, D. A. (1999). *Nonparametric Statistical Methods*, 2nd ed. Wiley, New York.

KOCH, G. (1972). The use of nonparametric methods in the statistical analysis of the two-period change-over design. *Biometrics* **28** 577–584.

LAGAKOS, S. (2003). Clinical trials and rare diseases. *New England J. Medicine* **348** 2455–2456.

ÖHRVIK, J. (1998). Nonparametric methods in crossover trials. *Biometrical J.* **40** 771–789.

PUTT, M. E. and CHINCHILLI, V. M. (2000). A robust analysis of crossover designs using multisample generalized L-statistics. *J. Amer. Statist. Assoc.* **95** 1256–1262.

PUTT, M. E. and RAVINA, B. (2002). Randomized, placebo-controlled, parallel group versus crossover study designs for the study of dementia in Parkinson's disease. *Controlled Clinical Trials* **23** 111–126.

SENN, S. (2002). *Cross-over Trials in Clinical Research*, 2nd ed. Wiley, New York.

TUDOR, G. and KOCH, G. G. (1994). Review of nonparametric methods for the analysis of crossover studies. *Statistical Methods in Medical Research* **3** 345–381.

VONESH, E. F. and CHINCHILLI, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. Dekker, New York.