BIOMETRICS 44, 211–218
March 1988

Using the Maximum Test Statistic in the Two-Period Crossover Clinical Trial

Andrew R. Willan*

National Cancer Institute of Canada, Clinical Trials Group,

and

Department of Community Health and Epidemiology,
Queen's University, Kingston, Ontario K7L 3N6, Canada

SUMMARY

In a two-period crossover trial where residual carryover is suspected, it is often advised that first-period data only be used in an analysis appropriate for a parallel design. However, it has been shown (Willan and Pater, 1986, *Biometrics* **42,** 593–599) that the crossover analysis is more powerful than the parallel analysis if the residual carryover, expressed as a proportion of treatment effect, is less than $2 - \sqrt{2(1 - \rho)}$, where ρ is the intrasubject correlation coefficient. Choosing between the analyses based on the empirical evaluation of this condition is equivalent to choosing the analysis with the larger corresponding test statistic. Approximate nominal significance levels are presented that maintain the desired level when basing the analysis on the maximum test statistic. Furthermore, the power and precision of the analysis based on the maximum test statistic are compared to the crossover and parallel analyses.

1. Introduction

Many authors (Cox, 1958; Grizzle, 1965; Hills and Armitage, 1979; Brown, 1980; Kershner and Federer, 1981; Laska, Meisner, and Kushner, 1983; Patel, 1983; Louis et al., 1984) discuss the design and analysis of the two-period crossover trial and address the issue of carryover effects. The general conclusion of this work is that the presence of carryover effect invalidates the use of the crossover design, and that unless carryover effects are negligible, a parallel design should be employed, or, if a crossover design has been used, that the analysis should be based only on the first-period data. Willan and Pater (1986) take issue with this advice and, using power and precision arguments, conclude that the amount of carryover effect required to make the parallel design preferable is substantial, and in most cases, unlikely to exist.

It is proposed in this paper that, under the assumption that no treatment effect implies no carryover, the test of treatment effect in a two-period crossover clinical trial be based on either the data from both periods in an analysis appropriate for a crossover design or on the data from the first period only in an analysis appropriate for a parallel design, whichever produces the larger test statistic. The model is outlined in Section 2, followed by a description of the analyses in Section 3. In Section 4 nominal significance levels for basing the test of treatment effect on the maximum test statistic are introduced. The maximum test statistic is evaluated with respect to power and precision in Sections 5 and 6, respectively.

---

* *Current mailing address:* Department of Population Medicine, Ontario Veterinary College, University of Guelph, Guelph, Ontario, N1G 2W1, Canada.

Key words: Carryover; Crossover trials; Maximum test statistic; Power; Precision.

2. The Model

Suppose there are two treatment sequences: sequence 1 is treatment 1 during period 1, followed by treatment 2 during period 2; and sequence 2 is treatment 2 during period 1, followed by treatment 1 during period 2. Let sequences and periods be indexed by i and k, respectively. Suppose there are n_i, indexed by j, subjects randomized to sequence i. Let Y_{ijk} be the observed outcome on the jth subject in the ith sequence during period k.

Following Grizzle (1965) and Brown (1980), we assume the following model:

$$Y_{ijk} = \mu + \pi_k + \phi_v + (k - 1)\lambda_i + \xi_{ij} + \varepsilon_{ijk},$$

where

μ = overall mean;

π_k = the effect of the kth period, $\pi_1 + \pi_2 = 0$;

ϕ_v = the effect of the vth treatment, $v = i*k(\text{mod } 3)$, $\phi_1 + \phi_2 = 0$;

λ_i = the carryover effect of treatment i from period 1 to period 2;

ξ_{ij} = the effect of the jth subject in the ith sequence;

ε_{ijk} = the within-subject deviation for the kth period.

We assume μ, π_k, ϕ_v, and λ_i to be fixed, and ξ_{ij} and ε_{ijk} to be normally distributed and mutually independent with mean zero and variances σ_ξ^2 and σ_ε^2, respectively. Consequently,

$$\text{covariance}(Y_{ijk}, Y_{ijk'}) = \begin{cases} \sigma_\xi^2 + \sigma_\varepsilon^2 & \text{if } k = k' \\ \sigma_\xi^2 & \text{if } k \neq k' \end{cases}$$

and

$$\text{correlation}(Y_{ij1}, Y_{ij2}) = \sigma_\xi^2/(\sigma_\xi^2 + \sigma_\varepsilon^2) = \rho.$$

Finally, let $n = n_1 + n_2$, $\phi = \phi_2 - \phi_1$, and $\lambda = \lambda_2 - \lambda_1$. Residual carryover is said to exist if $\lambda \neq 0$.

3. The Analyses

The parallel analysis is based on the test statistic

$$f_{\text{par}} = n_1 n_2 \hat{\phi}_{\text{par}}^2/[n(\hat{\sigma}_\xi^2 + \hat{\sigma}_\varepsilon^2)]$$

which when compared to $F_{1,n-2}(\alpha)$ provides a one-sided, level $\alpha/2$ test or a two-sided, level α test of treatment effect, where

$$\hat{\phi}_{\text{par}} = Y_{2.1} - Y_{1.1},$$

$$Y_{i.k} = \sum_{j=1}^{n_i} Y_{ijk}/n_i,$$

and

$$\hat{\sigma}_\xi^2 + \hat{\sigma}_\varepsilon^2 = \left[\sum_{i=1}^{2} \sum_{j=1}^{n_i} (Y_{ij1} - Y_{i.1})^2\right] \Big/ (n - 2).$$

The crossover analysis is based on the test statistic

$$f_{\text{co}} = 2n_1 n_2 \hat{\phi}_{\text{co}}^2/(n\hat{\sigma}_\varepsilon^2)$$

which when compared to $F_{1,n-2}(\alpha)$ provides a one-sided, level $\alpha/2$ test or a two-sided, level α test of treatment effect, where

$$\hat{\phi}_{\text{co}} = [Y_{1.2} - Y_{1.1} + Y_{2.1} - Y_{2.2}]/2$$

and

$$2\hat{\sigma}_\varepsilon^2 = \left\{ \sum_{i=1}^{2} \sum_{j=1}^{n_i} [Y_{ij2} - Y_{ij1} - (Y_{i.2} - Y_{i.1})]^2 \right\} \Big/ (n-2).$$

It should be stressed that if residual carryover persists in the absence of a treatment effect, f_{co} does not have a noncentral F distribution under the hypothesis, H: $\phi = 0$, of no treatment effect, and the level of the test based on f_{co} is invalid. However, in many trials residual carryover results not because a first-period treatment remains active in the second period but because a treatment effect produces expectations for the second period that differ between sequence groups. In these trials, and in the following arguments, residual carryover cannot exist in the absence of a treatment effect, and the hypothesis H implies that $\lambda = 0$. Discussion of the relaxation of this implication is found in Section 7.

Willan and Pater (1986) show that f_{co} provides the more powerful test if and only if

$$\lambda/\phi < 2 - \sqrt{2(1-\rho)}. \tag{3.1}$$

Hence, even in the presence of substantial residual carryover, the analysis of the data from both periods can provide a more powerful test of treatment effect. Substituting into (3.1) the estimates

$$\hat{\lambda} = (Y_{2.1} + Y_{2.2} - Y_{1.2} - Y_{1.1}),$$

$$\hat{\phi}_{\text{par}}, \quad \text{and} \quad \hat{\rho} = \hat{\sigma}_\xi^2/(\hat{\sigma}_\xi^2 + \hat{\sigma}_\varepsilon^2)$$

yields the condition $f_{\text{co}} > f_{\text{par}}$. Consequently, choosing between f_{co} and f_{par}, based on the empirical evaluation of (3.1), is equivalent to choosing the larger. Therefore, it appears that basing the test of treatment effect on $f_{\max} = \max\{f_{\text{par}}, f_{\text{co}}\}$ is a reasonable approach. Since σ_ξ^2 and $\sigma_\xi^2 + \sigma_\varepsilon^2$ must be estimated, the appropriate critical value for $f_{\max}$ could be calculated from the multivariate t distribution. In the following section, however, use is made of the corresponding z-statistics to determine nominal significance levels for $f_{\max}$.

4. Nominal Significance Levels for $f_{\max}$

To determine the nominal significance level, α', to apply to $f_{\max}$ to maintain the significance levels at α, the covariance between the test statistics, f_{par} and f_{co}, must be determined. To ease computation, use was made of the corresponding z-statistics defined as:

$$z_{\text{par}} = \hat{\phi}_{\text{par}} \sqrt{n_1 n_2/[n(\sigma_\xi^2 + \sigma_\varepsilon^2)]};$$

$$z_{\text{co}} = \hat{\phi}_{\text{co}} \sqrt{2 n_1 n_2/(n\sigma_\varepsilon^2)}.$$

The covariance between z_{par} and z_{co} is $\sqrt{(1-\rho)/2}$ and therefore $(z_{\text{par}}, z_{\text{co}})$ has, under H, a bivariate normal distribution with mean (0, 0) and covariance matrix with ones as diagonal elements and $\sqrt{(1-\rho)/2}$ as off-diagonal elements. Hence, an approximation for α', which is a function of ρ and hereafter written as $\alpha'(\rho)$, can be defined as

$$\Pr(z_{\text{par}} \geqslant Z_{\alpha'(\rho)} \text{ or } z_{\text{co}} \geqslant Z_{\alpha'(\rho)} \mid H) = \alpha,$$

and can be determined by numerical integration, where a standard normal deviate has probability α of exceeding Z_α. The significance levels $\alpha'(\rho)$ were determined for each combination of α = .05, .025, and .01 and ρ = 0(.1)1 and are presented in

Table 1
Significance levels; desired level = .05

$\hat{\rho}$	Nominal levels for f_{max}	Simulated true levels $n = 20$	$n = 50$	$n = 100$	$n = 200$
.0	.03037	.04968	.04912	.04983	.05022
.1	.02974				
.2	.02917	.05080	.05098	.05037	.05065
.3	.02864				
.4	.02814	.05132	.05097	.05154	.05010
.5	.02766				
.6	.02721	.05091	.05037	.05038	.05061
.7	.02679				
.8	.02637	.05067	.05057	.04895	.05040
.9	.02594				
1.0	.02532				

Table 2
Significance levels; desired level = .025

$\hat{\rho}$	Nominal levels for f_{max}	Simulated true levels $n = 20$	$n = 50$	$n = 100$	$n = 200$
.0	.01469	.02551	.02502	.02496	.02491
.1	.01441				
.2	.01414	.02559	.02480	.02563	.02488
.3	.01390				
.4	.01368	.02504	.02540	.02555	.02496
.5	.01348				
.6	.01329	.02533	.02446	.02505	.02592
.7	.01311				
.8	.01295	.02550	.02526	.02496	.02481
.9	.01279				
1.0	.01258				

Table 3
Significance levels; desired level = .01

$\hat{\rho}$	Nominal levels for f_{max}	Simulated true levels $n = 20$	$n = 50$	$n = 100$	$n = 200$
.0	.00569	.01031	.01021	.01011	.01055
.1	.00558				
.2	.00549	.01057	.01002	.01033	.00985
.3	.00541				
.4	.00533	.00969	.01057	.01036	.01018
.5	.00527				
.6	.00521	.01056	.01007	.01016	.01060
.7	.00515				
.8	.00511	.01043	.01048	.01009	.00952
.9	.00506				
1.0	.00501				

Tables 1, 2, and 3 for the three levels of α, respectively. Two-sided nominal significance levels are provided by $\alpha'(\rho)/2$.

These nominal levels were validated using simulation methods. For each combination of $\rho = 0(.2).8$ and $n = 20$, 50, 100, and 200, 100,000 sets of data were generated. Use was made of the IMSL subroutine GGNML. For each set of data $\hat{\rho}$ and $f_{\max}$ were determined. The simulated "true" significance level was determined by the proportion of data sets in which the value of $\alpha'(\rho)$ exceeded the standard significance level associated with $f_{\max}$, based on an F distribution with 1 and $n - 2$ degrees of freedom, and where $|\rho - \hat{\rho}|$ was minimized for ρ in $\{0, .1, .2, \ldots, 1\}$. The true simulated levels are also presented in Tables 1, 2, and 3. In each case the simulated levels appear sufficiently close to the desired levels.

5. Power

Numerical integration was used to evaluate the probability

$$\Pr(z_{\text{par}} \geqslant Z_{\alpha'(\rho)} \quad \text{or} \quad z_{\text{co}} \geqslant Z_{\alpha'(\rho)} \mid \phi = \phi_0)$$

to compare the power of the level .05 test based on $f_{\max}$, as described above, to the power of the tests based on f_{par} and f_{co}. Comparisons were performed for each combination of $\rho = 0(.2).8$ and $\lambda/\phi = 0(.2)1, 1.5, 2$. The value of ϕ_0 was set so that the power of the level .05 test based on f_{co}, for $\lambda/\phi = 0$, was 95%. By doing so the relative efficiencies are independent of n, and n was chosen to be 100 arbitrarily. The power comparisons for $\rho = .4$ are demonstrated in Figure 1.

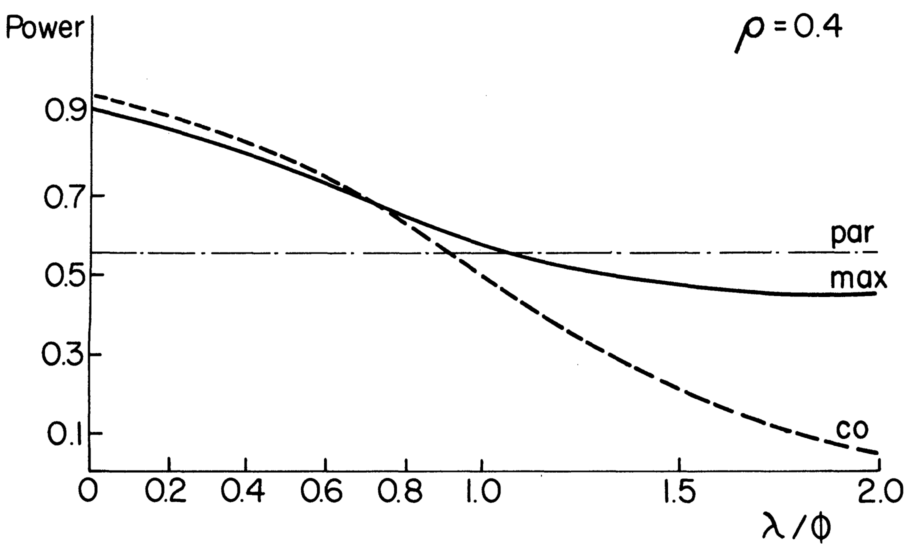

Figure 1. Power by degree of residual carryover for the parallel analysis (par), the crossover analysis (co), and the analysis based on $f_{\max}$ (max).

Over the range $\lambda/\phi < 2 - \sqrt{2(1 - \rho)}$, $f_{\max}$ compares favourably with f_{co} and outperforms f_{par}. Over the range $\lambda/\phi > 2 - \sqrt{2(1 - \rho)}$, $f_{\max}$ compares reasonably well with f_{par} and clearly outperforms f_{co}. In the range around $\lambda/\phi = 2 - \sqrt{2(1 - \rho)}$, $f_{\max}$ outperforms both f_{par} and f_{co}. The relative efficiencies of $f_{\max}$ as compared to f_{par} and f_{co} are displayed in Table 4. These relative efficiencies illustrate that $f_{\max}$ maintains most of the powers advantage of f_{co} for small amounts of residual carryover, and yet protects against the existence of large amounts of residual carryover.

Table 4
$RE_{co} = 100 * Power_{max}/Power_{co}$ *and* $RE_{par} = 100 * Power_{max}/Power_{par}$

	ρ									
	0		.2		.4		.6		.8	
λ/ϕ	RE_{co}	RE_{par}	RE_{co}	RE_{par}	RE_{co}	RE_{par}	RE_{co}	RE_{par}	RE_{co}	RE_{par}
0	98	124	98	139	97	164	97	214	97	336
.2	98	118	97	132	96	155	96	201	95	315
.4	100	111	98	123	96	143	94	184	93	285
.6	105	104	101	113	98	130	94	163	91	249
.8	117	98	110	104	104	116	97	141	91	209
1.0	141	94	129	96	117	104	105	121	93	170
1.5	329	90	284	87	235	86	184	87	132	100
2.0	1,347	90	1,149	86	919	82	664	77	399	73

6. Precision

Let $\hat{\phi}_{max}$ equal $\hat{\phi}_{co}$ if $f_{max} = f_{co}$ and $\hat{\phi}_{par}$, otherwise. Numerical integration was used to evaluate the mean squared error of $\hat{\phi}_{max}$ (MSE_{max}) defined as

$$E[(\hat{\phi}_{co} - \phi_0)^2\delta(z_{co}, z_{par}) + (\hat{\phi}_{par} - \phi_0)^2\delta(z_{par}, z_{co})],$$

where $\delta(x, y) = 1$ if $x > y$, zero otherwise. Setting $(\sigma_\xi^2 + \sigma_\epsilon^2)$ equal to one, arbitrarily, yields $MSE_{co} = 2(1 - \rho)/n + \lambda^2/4$ and $MSE_{par} = 4/n$. Mean squared errors were calculated for the same values of ρ, n, ϕ, and λ that were used in Section 5 for the power comparisons. The ratios MSE_{co}/MSE_{max} and MSE_{par}/MSE_{max} are presented in Table 5. $\hat{\phi}_{max}$ compares favourably with $\hat{\phi}_{par}$ for all values of ρ and λ/ϕ. $\hat{\phi}_{max}$ compares favourably with $\hat{\phi}_{co}$ except for large values of ρ and small values of λ/ϕ.

Table 5
$RP_{co} = 100 * MSE_{max}/MSE_{co}$ *and* $RP_{par} = 100 * MSE_{max}/MSE_{par}$

	ρ									
	0		.2		.4		.6		.8	
λ/ϕ	RP_{co}	RP_{par}	RP_{co}	RP_{par}	RP_{co}	RP_{par}	RP_{co}	RP_{par}	RP_{co}	RP_{par}
0	88	176	85	213	81	270	75	374	62	619
.2	88	160	84	189	78	235	70	314	54	490
.4	98	137	90	158	82	190	71	246	53	368
.6	117	119	105	133	92	155	77	194	55	279
.8	148	108	127	117	108	131	87	159	60	220
1.0	190	103	159	108	130	117	101	137	67	182
1.5	354	100	285	101	220	103	158	112	96	135
2.0	591	100	473	100	357	101	246	104	138	117

7. Discussion

This paper is presented as an argument in favour of using the maximum test statistic when analysing data from a two-period crossover trial. Nominal significance levels are derived that are valid for trials in which no treatment effect implies no residual carryover. In these trials residual carryover, perhaps best described as a treatment by period interaction, can exist only in the presence of treatment effect. For example, consider an antiemetic trial with previously untreated cancer patients receiving at least two consecutive identical courses of chemotherapy. In general, patients are conditioned by their experience and suffer a greater degree of nausea and vomiting in the second period, resulting in a pure period effect. However, patients receiving the inferior antiemetic in the first period may be conditioned to a greater extent than the patients receiving the superior antiemetic first. The net effect is a psychological residual carryover, resulting in a treatment by period interaction, which will diminish the treatment effect in the second period. An example of such a trial can be found in Willan and Pater (1986). A treatment by period interaction may also exist in the presence of a pure period effect if treatment efficacy is a function of disease severity. Such may be the case if patients are entered during a relatively severe stage of a chronic disease. A treatment effect, if it exists, may be reduced in the second period as disease severity diminishes.

Treatment by period interaction can also result in an exaggerated second period treatment difference. Suppose some of the patients by experiencing a treatment difference can, during the second period, correctly determine in which period they received the superior treatment. In these trials such "unblinding" can easily lead to evaluations of treatment effect during the second period that are exaggerated in favour of the superior arm. These effects can exist only in the presence of treatment effect and, consequently, the nominal levels reported in Section 4 are valid.

However, more caution must be exercised if there exists a physical carryover, resulting because a first-period treatment remains active in the second. In such cases residual carryover can exist in the absence of treatment effect, causing the treatment with the smallest carryover effect to outperform the other in the second period, with the possibility of f_{co}, and thereby f_{max}, achieving significance erroneously. This cannot happen, however, in a placebo-controlled trial or in a trial of a new treatment in which the standard treatment is known, from previous experience, not to carry over to the second period. These trials should be analyzed with one-sided tests in which the direction of the possible erroneous significance is not of interest. In such trials the nominal levels presented in Section 4 are valid. In other trials in which residual carryover may exist in the absence of treatment difference, the use of the second-period data, however analysed, could lead to an uncontrolled Type I error.

Investigators may use designs that have more than two sequences and/or periods to provide unbiased estimators of treatment effect in the presence of carryover (Kershner and Federer, 1981; Laska et al., 1983). Among the two-period designs, the design consisting of the four sequences (1, 2), (2, 1), (1, 1), and (2, 2) provides the unbiased estimator with the least variance (Laska et al., 1983). The patients in sequence (i, j) receive treatment i in period 1 and treatment j in period 2. The variance of the estimator of treatment effect for this design is greater than the variance of $\hat{\phi}_{par}$ if $\rho < .5$, and although ρ may sometimes exceed .5, it would appear to be a gamble at best to use the four-sequence design, rather than a parallel design, especially when one considers that twice as many observations are required and that patients will be lost between periods.

A real alternative is provided by the multiperiod designs. For example, the three-period design with the two sequences (1, 2, 2) and (2, 1, 1) provides an unbiased estimator of treatment effect with 25% less variance than that for the standard crossover design (1, 2), (2, 1) with the same number of patients. Whenever feasible, this or other optimal multiperiod designs (see Laska et al., 1983) should be used. However, when only two-period designs are feasible, or when in spite of feasibility, a two-period design has been used and if one is confident that no treatment effect implies no carryover effect, as described in the situations above, the use of the maximum test statistic yields a test with the appropriate level which in most cases will be more powerful and will provide a more precise estimator of treatment effect than the parallel design or the design (1, 2), (2, 1), (1, 1), and (2, 2).

In many situations multiperiod designs are infeasible. For antiemetic trials in cancer chemotherapy, few patients will have three identical courses of chemotherapy. Some will be on alternating regimens and many will have dose changes due to toxicities. Still others will be taken off treatment because of disease progression or death. In other trials it may be impossible to add even 50% to an already lengthy trial. Problems with multiperiod designs may occur if patients after completing the second-period treatment have a strong preference for the treatment they received in a particular period. Such patients may insist on coming off trial to receive the preferred treatment and physicians may well be ethically bound to comply with such requests.

Acknowledgements

This research was supported by a grant from the National Cancer Institute of Canada. The author wishes to thank the referees whose comments have improved the manuscript considerably.

Résumé

Dans un plan d'expérience à deux périodes avec permutation des traitements (cross-over) en présence d'arrière-effets résiduels, il est souvent recommandé de n'utiliser que les données de la première période en les analysant suivant un plan parallèle. Cependant, on a montré (Willan et Pater, 1986, *Biometrics* **42,** 593–599) que l'analyse du plan avec permutations est plus puissante que l'analyse du plan parallèle si les arrière-effets représentent une proportion des effets directs inférieure à $2 - \sqrt{2(1 - \rho)}$ et si ρ est le coefficient de correlation intra-sujet. Faire le choix entre les analyses en se basant sur l'évaluation empirique de cette condition est équivalent à choisir l'analyse avec la plus grande valeur de statistique de test. Des niveaux de signification nominale sont donnés, qui garantissent le niveau désiré quand on se base sur la statistisque maximum. De plus, la puissance et la précision de l'analyse basée sur la statistique maximum est comparée aux analyses des plans avec permutation et des plans parallèles quant à la puissance et à la précision.

References

Brown, B. W. (1980). The crossover experiment for clinical trials. *Biometrics* **36,** 69–79.

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,** 469–480. 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.

Kershner, R. P. and Federer, W. T. (1981). Two-treatment crossover designs for estimating a variety of effects. *Journal of the American Statistical Association* **76,** 612–619.

Laska, E., Meisner, M., and Kushner, H. B. (1983). Optimal crossover designs in the presence of carryover effects. *Biometrics* **39,** 1087–1091.

Louis, T., Lavori, P., Bailar, J., and Polansky, M. (1984). Crossover and self-controlled designs in clinical research. *The New England Journal of Medicine* **31,** 24–31.

Patel, H. I. (1983). Use of baseline measurements in the two-period crossover design. *Communications in Statistics* **12,** 2693–2712.

Willan, A. R. and Pater, J. L. (1986). Carryover and the two-period crossover clinical trial. *Biometrics* **42,** 593–599.

Received June 1986; *revised April and July* 1987.