Response to Letter to the Editor

Permutation Tests Following Restricted Randomization Procedures

In his letter, Kalish [1] discusses the potential differences between a permutation test that is based on the randomization distribution for the randomization actually employed (the *correct* permutation test) and a test based on the randomization distribution for a complete binomial randomization, which Kalish terms the *standard* permutation test. Kalish presents the view that *in practice* correct permutation tests are not necessary for the analysis of clinical trials in order to perform tests of the proper size (type I error probability). His assertion is that standard permutation tests are adequate even when a restricted randomization procedure has actually been employed to assign treatments to patients. To support this view Kalish and Begg [2] had previously conducted a simulation study based upon data from a collection of clinical trials conducted by the Eastern Cooperative Oncology Group (ECOG).

It is perfectly true that in some cases the standard permutation test yields the same result as the restricted randomization permutation test. To put it another way, a permutation test based on the correct permutational distribution is sufficient to yield a test of proper size but is not necessary. Unfortunately, it has not yet been possible to describe universal conditions under which the standard versus the correct permutation tests will yield equivalent results and the conditions under which they will differ. In Matts and Lachin [3], we showed for the permuted-block design that the difference between these test results is directly related to the intrablock correlation coefficient. However, the relationship between these tests for the urn randomization has not been so simply described. In general, therefore, we feel that the only way to know whether or not the correct permutation test makes a difference is to actually compute it, or equivalently for a permuted-block randomization, to compute the intrablock correlation coefficient.

One condition under which it is known that the standard versus the correct permutation tests will yield different results is in the presence of a linear time trend in the patient responses. This is obviously only one of an infinite class of systematic trends that might occur. Based on the results of a collection of ECOG studies, Kalish and Begg [2] concluded that standard permutation tests, and thus ordinary population model tests, are acceptable in practice. This may be true of ECOG cancer trials, or of other cancer trials in general, but certainly is not true of all clinical trials. In fact, in most long-term clinical trials in which patients are recruited over extended periods, perhaps as long as 3 or more years, it is highly likely that major differences will exist between patients recruited early as opposed to late in the study. Therefore, in planning the analysis for any one clinical trial, we do not find it reassuring to know

0197-2456/1990/$3.50

Controlled Clinical Trials 11:150–152 (1990)

that in a relatively small number of trials conducted by the ECOG, the use of a standard permutation test did not make a difference. The only way to find out whether it will make a difference for any one clinical trial is to actually do the analysis using the correct permutation test for that trial.

Kalish [1] also criticizes the time trend example presented by Wei and Lachin [4] because the trend exhibited was quite strong. We agree that this was the case. However, we make no claim that such a linear trend is typical. Rather, the point was to demonstrate one condition under which the standard permutation test yields an incorrect result. Again, the issue is that the standard permutation test *can* yield a result that is different from the correct permutation test, and the only way to be sure of the results in any one particular trial is to perform the correct analysis.

Kalish [1] also suggests that a simple analysis that stratifies by time blocks of entry may be just as effective in providing a test of proper size as does the correct permutation test. This is true if one knows in advance the nature of the trend in responses that could then be accounted for by such stratification. Unfortunately, this can never be known in practice. Also, the efficiency of the time-stratification for any one particular study cannot be known unless one compares the simple stratified analysis with the correct permutational analysis.

In conclusion, therefore, we agree that in some cases a correct permutational analysis may not be necessary or may be supplanted by a simpler time-stratified analysis. Unfortunately, however, one never knows whether these assertions will apply for any one particular trial unless the correct permutational analysis is also conducted.

As a final point, Kalish [1] asserts that the use of the correct permutation test may in fact reduce power in some circumstances. He uses the example of a permuted-block design with B blocks where the intrablock correlation is close to zero. In this case the standard permutation test will yield a test of the proper size, but with a larger degrees of freedom for error in an ordinary ANOVA. The corresponding correct analysis would employ a blocked ANOVA with $B - 1$ degrees of freedom removed from df error. However, the impact on power may be negligible. For example, for a trial with 100 patients and 50 blocks of size 2, a blocked analysis requires a t value of 2.01 at the 5% significance level but an unblocked analysis requires only a slightly smaller value of 1.98. As the number of blocks decreases, the larger value quickly converges to the smaller value.

Further, it has been our experience that the ANOVA is rarely used in the analysis of the results of clinical trials. Rather, the principal methods employed are the χ^2 tests for proportions or for life tables, in which case the critical value of the test is the same for a blocked and an unblocked analysis. In this context, the question then becomes one of comparing the variances for the correct versus the standard permutational test. In some cases the standard permutational variance will be larger than the correct permutational variance, thus yielding a smaller test value and sacrificing power. However, in some rare cases, it may actually be the case that the correct permutational variance is larger than the standard variance, such as where a negative intrablock correlation occurs in a permuted-block design. Although, as described by Matts and Lachin [3], this may be extremely unlikely, it is our view that if

this event were to arise, the correct p value is provided by the correct permutation test rather than the standard test, even though the latter may be "more significant."

We feel that many of these points require further study and are of great practical interest because correct permutation tests are not as readily available or as easy to perform as a t test or a simple χ^2 test. Therefore, we encourage further research to address the issues raised by Professor Kalish. In our view, however, especially in the context of multimillion dollar clinical trials, it makes little sense to simply perform a standard permutation or population model test because of the assertion that such tests *might* yield a test of the proper size. Unless this assertion could be applied to a particular trial with a probability approaching 1, we feel that the correct permutation test should serve as the gold standard.

John M. Lachin, ScD
The George Washington University
Department of Statistics/Computer and Information Systems
The Biostatistics Center
Rockville, Maryland

John P. Matts, PhD
Department of Surgery
University of Minnesota
Minneapolis, Minnesota

L.J. Wei, PhD
Department of Biostatistics
School of Public Health
University of Michigan
Ann Arbor, Michigan

REFERENCES

1. Kalish LA: Permutation tests following restricted randomization procedures [Letter]. Controlled Clin Trials 11:147–149, 1990 [this issue]
2. Kalish LA, Begg CB: The impact of treatment allocation procedures on nominal significance levels and bias. Controlled Clin Trials 8:121–135, 1987
3. Matts JP, Lachin JM: Properties of permuted-block randomization in clinical trials. Controlled Clin Trials 9:327–344, 1988
4. Wei LJ, Lachin JM: Properties of the urn randomization in clinical trials. Controlled Clin Trials 9:345–364, 1988