Impact of Stratification Imbalance on Probability of Type I Error

Kazem KAZEMPOUR

The impact of ignoring the stratification effect on the probability of a Type I error is investigated. The evaluation is in a clinical setting where the treatments may have different response rates among the strata. Deviation from the nominal probability of a Type I error, α, depends on the stratification imbalance and the heterogeneity in the response rates; it appears that the latter has a larger impact. The probability of a Type I error is depicted for cases in which the heterogeneity in the response rate is present but there is no stratification imbalance. Three-dimensional graphs are used to demonstrate the simultaneous impact of heterogeneity in response rates and of stratification imbalance.

KEY WORDS: Analysis of covariance; Baseline adjustment; Pretest–posttest.

1. INTRODUCTION

Safety and efficacy of a treatment are the most important issues in any clinical trial. Often, other factors such as gender, stage of disease, etc., are taken into account as part of the trial design. If these factors are not incorporated in the protocol at the design stage, their impact on the Type I error rate may mask the findings of the trial by causing a false positive rate different than the nominal α. This stems from the fact that treatments may not have the same response rate in different patient groups such as those formed by gender, stage of disease, race, etc. For example, in an AIDS clinical trial, the rates of adverse events of an antiretroviral treatment, AZT, are different among the HIV-infected individuals who have reached the AIDS stage compared to those who are in the AIDS-related complex stage (McLeod and Hammer 1992). Or in a hypertension study it is known that the effect of a beta blocker, a treatment to reduce blood pressure, is different in elderly compared to young patients (Neutel, Smith, Lefkowitz, Kazempour, and Weber 1993).

In clinical trials, stratification and randomization are used to distribute the controlled and uncontrolled factors evenly among the treatment groups. It is common to apply a statistical test (although there is some argument against its use) to compare the distribution of the important characteristics of different treatment groups before the initiation of treatment, i.e., baseline. The baseline characteristics can be discrete, e.g., race, gender, or continuous, e.g., weight, blood pressure. Those characteristics that are discrete may be used as stratification factors; those that are continuous may be used as covariates or pretest values. Brogan and Kunter (1980) compared different methods for "Pretest–Posttest" for case in which the baseline variables are continuous variables. They recommended that investigators choose the method of analysis based on the objective(s) of the trial; their preferred method was the repeated-measure/split-plot analysis over the t test. Laird (1983) proposed an alternative method for analyzing randomized studies with covariates and demonstrated that the results are identical to the ordinary analysis of covariance when the pretest is used as the covariate. For discussion opposing the pretesting strategy, see, for example, Permutt (1990), Senn (1989), and Altman (1985). Permutt (1990) showed that pretesting for imbalance of baseline values affects the final probability of Type I error and results in a significance level lower than the nominal level. Altman (1985) suggested that comparability of prognostic factors should be evaluated partly on the basis of clinical knowledge. He showed that a nonsignificant imbalance in an important covariate may exert an impact on the final result. Senn (1989) showed the effect of baseline imbalance on the Type I error rates. For continuous variable, Senn has proposed a procedure to assess the effect of baseline covariates on the probability of Type I error.

In this article we look at discrete characteristics (stratification), and we use a method similar to Senn's (1989) to investigate the impact of ignoring the stratification factor on the probability of a Type I error. In the next section we describe the problem, our approach, and the notation. The impact of different response rates in different strata and the stratification imbalance are explained in Section 3. The discussion and conclusion are presented in Section 4.

2. DERIVATION AND NOTATION

In a clinical trial that compares the effect of two treatments, assume that there are two strata, and n subjects are assigned to each treatment group. Also assume that for the binary response of interest, the response rates may not be the same in both strata within each treatment group. The following 2×2 tables display the true response rates, the sample allocations, and the observed response frequencies. The proportion of patients from the ith treatment group that are in the first stratum is denoted by f_i; $i = 1, 2$. This quantity is assumed fixed in all derivations. The main objective is to compare the response rates of two treatment groups. (See the top of p. 171.)

Let R_i be the true response rate of the ith treatment group; then

$$R_1 = f_1\theta + (1 - f_1)\phi$$
$$R_2 = f_2(\theta - \delta) + (1 - f_2)(\phi - \delta).$$

Then

$$\widehat{R}_i = (X_{1i} + X_{2i})/n$$

Kazem Kazempour is Mathematical Statistician, Division of Biometrics, U.S. Food and Drug Administration, Rockville, MD 20857. The views expressed here are those of the author and not of the U.S. Food and Drug Administration. The author thanks Dr. L. Kammerman, Dr. P. Flyer, and an anonymous referee for helpful comments that substantially improved the quality of the manuscript.

True Response Rates

Stratum	Treatment 1	Treatment 2
1	θ	$\theta - \delta$
2	ϕ	$\phi - \delta$

Sample Allocation Frequencies

Stratum	Treatment 1	Treatment 2
1	nf_1	nf_2
2	$n(1-f_1)$	$n(1-f_2)$

Observed Response Frequencies

Stratum	Treatment 1	Treatment 2
1	X_{11}	X_{12}
2	X_{21}	X_{22}

is an unbiased estimator of the ith treatment's response rate, and

$$\hat{D} = \hat{R}_1 - \hat{R}_2$$

is an unbiased estimator of their difference, $R_1 - R_2$. The first two moments of this difference are

$$E(\hat{D}) = R_1 - R_2 = (\theta - \phi)(f_1 - f_2) + \delta,$$

$$V(\hat{D}) = V(\hat{R}_1) + V(\hat{R}_2)$$
$$= \frac{1}{n}[(f_1 + f_2)\theta(1-\theta) + (2 - f_1 - f_2)\phi(1-\phi) + 2\delta(f_2\theta + (1-f_2)\phi) - (\delta + \delta^2)].$$

Under the null hypothesis that these two treatments have the same effect; that is, H_0: $\delta = 0$, the second moment of $\hat{D}$ is obtained by setting $\delta = 0$ in the above variance. We refer to this as the true variance and denote it by V_0.

If the stratification effect is ignored, the variance of the overall difference in rates, $\hat{D}$ given $\delta = 0$ is calculated as

$$V_c = \frac{2}{n}\bar{p}(1-\bar{p})$$

where

$$\bar{p} = \frac{f_1 + f_2}{2}\theta + \left(1 - \frac{f_1 + f_2}{2}\right)\phi.$$

Note that in practice, θ and ϕ are not known, and they are replaced with proper estimates based on the observed responses. In this paper we are using the true values of θ and ϕ. To evaluate the probability of false rejection, the normal approximation to the binomial distribution is employed. Using this approach in a one-sided test, the hypothesis of no difference in the response rates between treatment groups is rejected if

$$\hat{D} \geq Z_\alpha \sqrt{V}, \qquad (1)$$

where Z_α is the $(1 - \alpha)$ percentile from a standard normal table; V is the variance that is either the calculated variance, V_C, or the true variance, V_0; the effect of this difference is studied in this article. The impact of using the calculated variance rather than the true variance on the probability of false rejection depends on the stratification imbalance and the response rates in different strata.

The Normal approximation to the conditional probability of falsely rejecting the null hypothesis in the case that $f_1 = f_2 = f$ is

$$P(\hat{D} > Z_\alpha \sqrt{V_C}; f) = 1 - \Phi(Z_\alpha \sqrt{V_C/V_0}), \qquad (2)$$

where

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx.$$

For the case where f_1 and f_2 are not equal, the calculated probability of a Type I error is

$$P(\hat{D} > Z_\alpha \sqrt{V_C}; f_1, f_2)$$
$$= 1 - \Phi(Z_\alpha \sqrt{V_C/V_0} - (f_1 - f_2)(\theta - \phi)/\sqrt{V_0}). \qquad (3)$$

It is obvious from Equations (2) and (3) that the probability of false rejection is affected by the response rates as well as the proportions, f_i; these impacts are depicted in Figures 1 and 2, respectively. In these figures the values for ϕ are fixed, and the values for θ range from 0 to ϕ; the values for f_1 are also fixed, and the values of f_2 range from 0 to f_1. The nominal probability of Type I error α is set at .05. In the next section the impact of different values of these parameters is evaluated.

3. IMPACT OF THE PARAMETERS

The Type I error rate is affected by θ, ϕ, f_1, and f_2. To evaluate this effect, we examine the following three scenarios:

1) $\theta/\phi = 1$
2) $\theta/\phi < 1$, and $f_2/f_1 = 1$
3) $\theta/\phi < 1$, and $f_2/f_1 < 1$.

Under the first scenario there is no impact on the probability of Type I error. This is expected because the response rates in both strata are the same. If we assume $f_1 = f_2 = f$ and $\theta/\phi < 1$, the second scenario, then the Type I error rate is affected. The size of this effect depends on the values of f, the number of patients in each cell relative to the total number of patients in each treatment group, and the size of ϕ.

Figure 1 depicts these effects for four different sets of values of f and ϕ. Note that there is no imbalance between the treatment groups, $f_1 = f_2$; the only difference is the response rates between the strata within each treatment group. The values of ϕ are displayed on the top of the left corner of the boxes. The values for θ are chosen as fractions of ϕ. The Y axes are the probability of Type I error conditioned on the proportions, f. The X axes are $1 - (\theta/\phi)$. When $X = 0$, the response rates in both strata are the same; that is, $\theta = \phi$. Therefore, the calculated Type I error rates are the same as the nominal values. The Type I error rates obtained from Equation (2) are, in general, less than the nominal values, so the tests are conservative.

The line parallel to the X axis in Figure 1a is for $f = 1$ (i.e., no stratification). This line displays that the nominal α is the same as the calculated probability of Type I error for $\phi = .9$. The same results are displayed for other values of ϕ in Figure 1b–d. When f is not 1, the difference between the nominal α (.05) and the calculated probability of false rejection is more pronounced in Figure 1a than

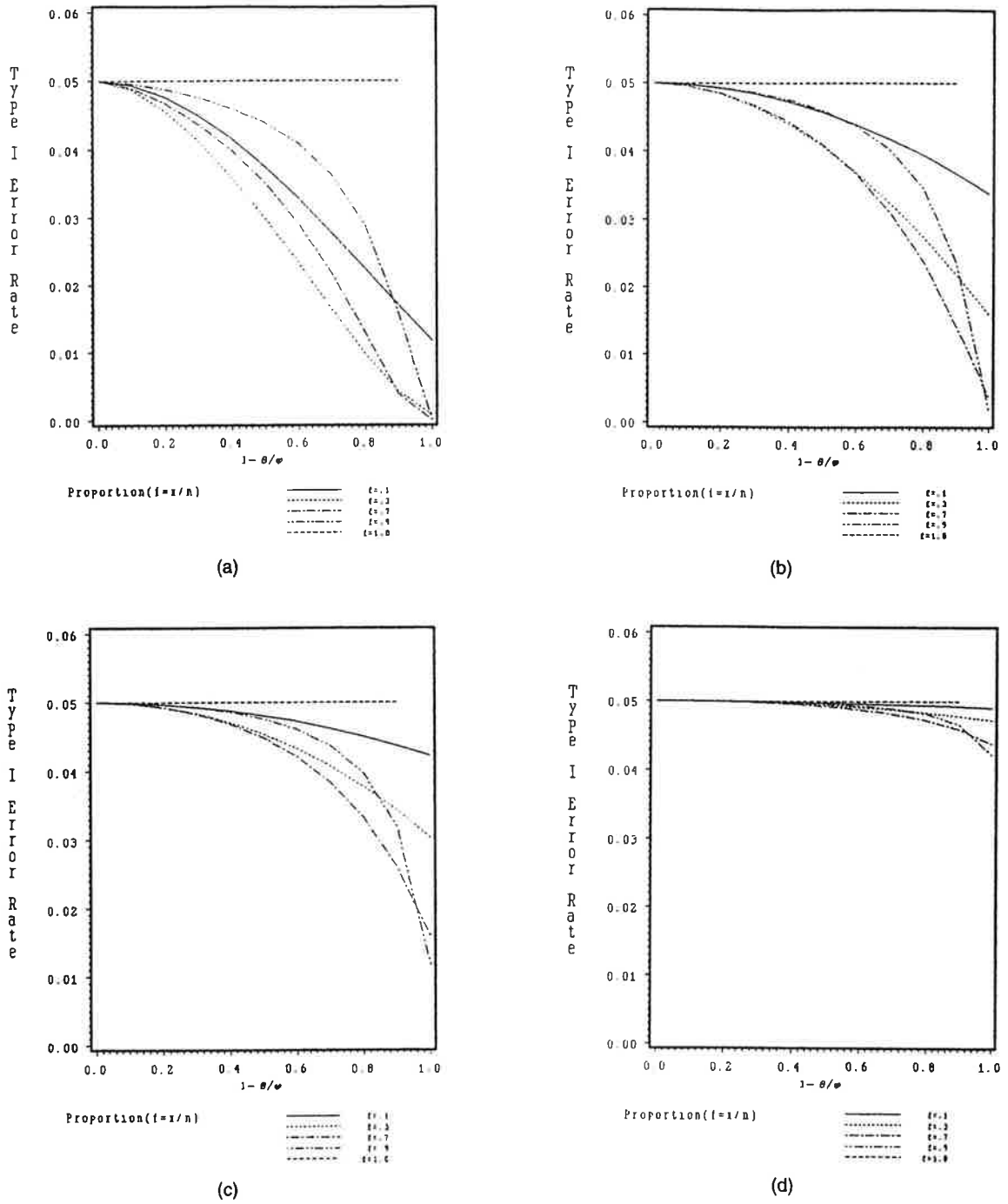

Figure 1. Depicting the Impact of Heterogeneity in Response Rates in Strata on Type I Error Rate. Four different fixed values for ϕ, response rate, are used. The x axis is $1 - \theta/\phi$. The proportion of participants in strata are equal, $f = f_1 = f_2$; in each box five different values of f are displayed. (a) The response rate ϕ is .9, and the θ ranges from 0 to .9. (b) The response rate in one stratum is .7, and for the other stratum ranges from 0 to .7. (c) The response rate in one stratum is .5, and for the other stratum ranges from 0 to .5. (d) The response rate in one stratum is .1, and for the other stratum ranges from 0 to .1.

Figure 1b–d. This is because the largest difference between ϕ and θ can be observed in Figure 1a. In Figure 1d the largest value that θ or ϕ can have is .1. Therefore $\phi - \theta$ cannot be greater than .1; this small difference has little effect on the Type I error rate.

Figure 2 demonstrates the third scenario. If both the response rates and the proportion rates are different, the statistical tests are the most affected. In this scenario Type I error rate is affected from two sources: (1) $E(\widehat{D} \mid H_0) \neq 0$, and (2) $V_C \neq V_0$. In this article we used the unbiased estimator $\widehat{D}$ to demonstrate the impact of stratification imbalance; this estimator is not an unbiased estimator of zero under the null hypothesis for this scenario. One might use a different estimator that may not be unbiased, but is an unbiased estimator of zero under the null hypothesis. This will be addressed in the next section.

The effects of simultaneous changes in ϕ and f_1 are depicted in four three-dimensional graphs for four sets of (ϕ, f_1). The Z axes in Figure 2a–d are the conditional probability of Type I error. In Figure 2 the X axes are the ratios of the proportions, f_2/f_1. The values of f_1 are displayed on the top of the left corner of the boxes. The values of f_2 are calculated as fractions of f_1. The Y axes in these figures are the ratios of the response rates, θ/ϕ. The

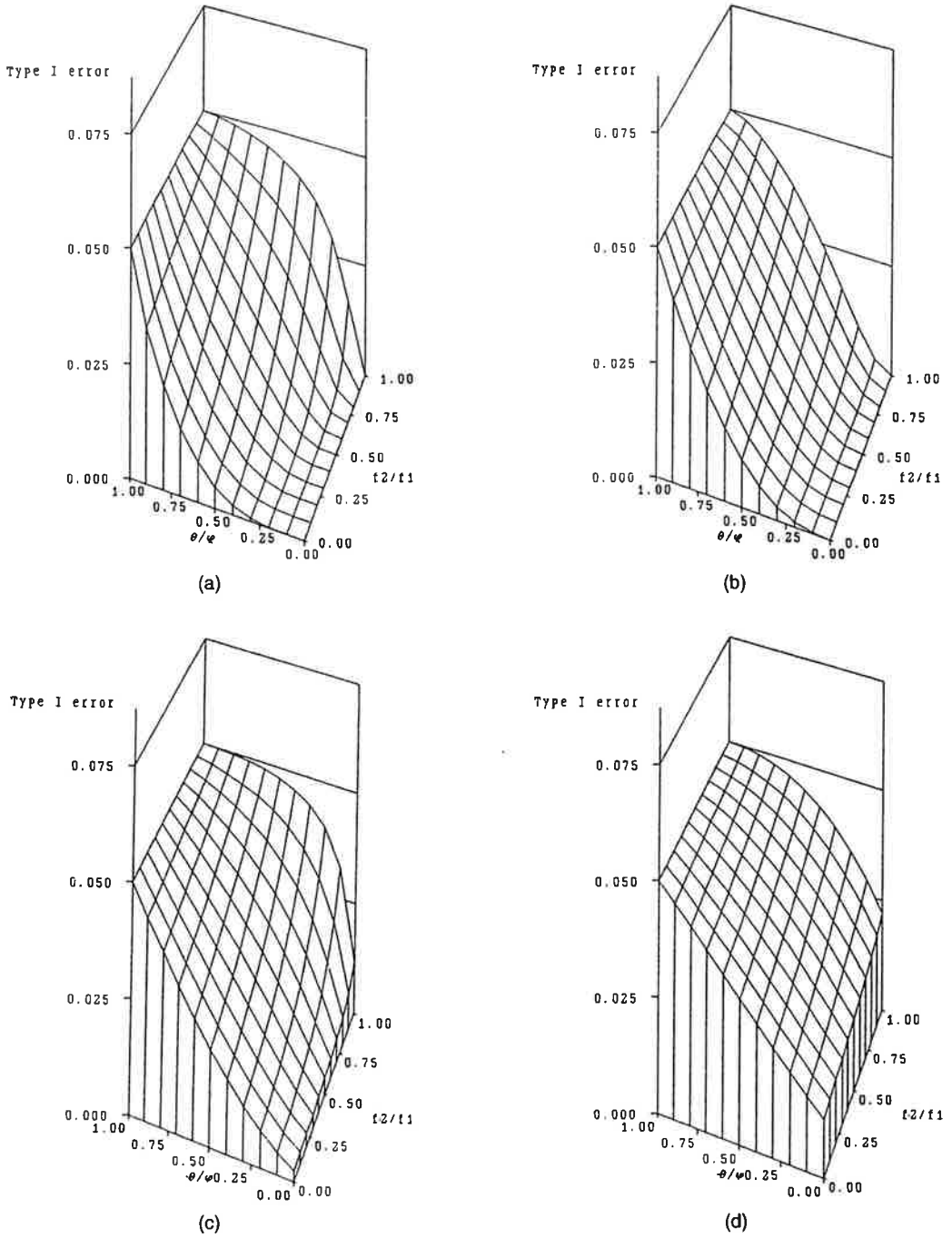

Figure 2. Depicting the Impact of Imbalance in the Stratification Factor and Heterogeneity in the Response Rates. The largest proportion rate, f_1, and the response rate, ϕ, are displayed on the top of the left corner of each box. (a) ϕ is .9, and θ ranges from 0 to .9; f_1 is .9, and f_2 ranges from 0 to .9. (b) ϕ is .5, and θ ranges from 0 to .5; f_1 is .9, and f_2 ranges from 0 to .9. (c) ϕ is .9, and θ ranges from 0 to .9; f_1 is .5, and f_2 ranges from 0 to .5. (d) ϕ is .5, and θ ranges from 0 to .5; f_1 is .9, and f_2 ranges from 0 to .9.

values for ϕ are displayed next to the f_1 values for each figure, and the values of θ range from 0 to ϕ. When $\theta = \phi$ the nominal α is the same as the calculated probability of false rejection, regardless of the values of f_1 and f_2. When $f_1 = f_2$ and the θ and ϕ are different, the Type I error rate is less than the nominal value. This is what we observed in Figure 1. It is clear from Figure 2 that the Type I error is more sensitive to the dispersion in the response rates than the imbalance of the stratification.

4. DISCUSSION AND CONCLUSION

When stratification is ignored, dispersion of response rates in a heterogenous population will affect the Type I error rate. This effect can be large, and it should be assessed and addressed. The statistical tests will become more conservative if the response rates in different strata are far apart; this is true even if there is no stratification imbalance. For an extreme case where $\phi = 1$ and $\theta = 0$ with $f_1 = f_2 = .5$, the variance of R_1 will be $.5(1 - .5)/n$; however, R_1 is always equal to .5 and the true variance is 0.

From Figure 1 it can be observed that, although the attributes are evenly distributed between the two treatment groups, the tests are affected and they are conservative. This, indirectly, supports the argument that pretesting the baseline characteristics is not the most appropriate method to assess and to remove the effect of the stratification

imbalance. The appropriate approach is to control for important stratification factors in the planning stage, and use the stratification factor in the analysis, even if there is no stratification imbalance.

The cases that we have investigated in this paper are small subsets of all possible cases that one could evaluate. In this article we have used $\hat{D}$, which is an unbiased estimator of $R_1 - R_2$. This is not an unbiased estimator of zero under the null hypothesis in the third scenario. One could use an alternative estimator such as $\hat{D}'$, such that $E(\hat{D}' \mid H_0) = 0$, where

$$\hat{D}' = [(f_1 + f_2)/2](X_{11}/nf_1 - X_{12}/nf_2) + [1 - (f_1 + f_2)/2](X_{21}/n(1 - f_1) - X_{22}/n(1 - f_2)).$$

Other cases of practical interest are: (1) the response rates of the two strata are the same in one treatment group but they are different in the other treatment group, and (2) the response rates of one stratum in both treatment groups are equal but the response rates of the other stratum are different. In this study the normal approximation procedure is employed to calculate the probability of Type I error; the exact binomial distribution is an alternative. The sample sizes in two treatment groups are assumed equal in this article; a difference in the sample sizes in each treatment group might add another layer of complexity. The dispersion in the sample sizes is especially important when the sample sizes are small. This can be evaluated when the exact binomial distribution is used.

[*Received March 1993. Revised October 1994.*]

REFERENCES

Altman, D. G. (1985), "Comparability of Randomized Groups," *The Statistician*, 34, 125–136.

Brogan, D. R., and Kunter, M. H. (1980), "Comparative Analyses of Pretest–Posttest Research Designs," *The American Statistician*, 34, 229–232.

Laird, N. (1983), "Further Comparative Analyses of Pretest–Posttest Research Designs," *The American Statistician*, 37, 329–330.

McLeod, G. X., and Hammer S. M. (1992), "Zidovudine: Five Years Later," *Annals of Internal Medicine*, 117, 484–501.

Neutel, J., Smith, D., Lefkowitz, M., Kazempour, M. K., and Weber, M. (1991), "Whole-Day Blood Pressure Monitoring in Assessing Efficacy of Anti Hypertensive Agents in Population Subgroups," *Clinical Research*, 39(3), 683A, 749A.

Permutt, T. J. (1990), "Testing for Imbalance of Covariates in Controlled Experiments," *Statistics in Medicine*, 9, 1455–1462.

Senn, S. J. (1989), "Covariate Imbalance and Random Allocation in Clinical Trials," *Statistics in Medicine*, 8, 467–475.