

11-1-2007

## Tests for 2 x 2 Tables in Clinical Trials

Vic Hasselblad  
*Duke University*

Yulia Lokhnygina  
*Duke University*

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

---

### Recommended Citation

Hasselblad, Vic and Lokhnygina, Yulia (2007) "Tests for 2 x 2 Tables in Clinical Trials," *Journal of Modern Applied Statistical Methods*: Vol. 6 : Iss. 2 , Article 10.  
DOI: 10.22237/jmasm/1193890140

---

# Tests for 2 x 2 Tables in Clinical Trials

## **Cover Page Footnote**

The authors wish to thank Paula Smith for manuscript editing and formatting assistance

---

## Tests for $2 \times 2$ Tables in Clinical Trials

Vic Hasselblad      Yuliya Lokhnygina  
Duke University

---

Five standard tests are compared: chi-squared, Fisher's exact, Yates' correction, Fisher's exact mid-p, and Barnard's. Yates' is always inferior to Fisher's exact. Fisher's exact is so conservative that one should look for alternatives. For certain sample sizes, Fisher's mid-p or Barnard's test maintain the nominal alpha and have superior power.

Key words: Power, sample size, dichotomous endpoint, alpha level

---

### Introduction

The literature on tests for  $2 \times 2$  tables is extremely vast and controversial. However, the issues can be focused somewhat when considering the use of these tests for clinical trials. In this situation, the trials have two arms and the sample size of each arm is fixed. Tests are almost always made at the 0.05 nominal alpha level. There is no requirement that the tests be computationally simple, but only that they are available in standard commercial statistical software. The following two examples illustrate many of the issues of interest.

Cotter et al (2000) conducted a small, randomized pilot study (15 patients per treatment arm) comparing N<sup>o</sup>-nitro-L-arginine

methyl ester (L-NAME) to placebo in patients with cardiogenic shock. Mortality results are given in Table 1.

Cotter et al. reported a p-value of 0.028 (no test specified), a value which is consistent with the standard chi-square test. However, if Fisher's exact test had been used in the standard manner, the p-value would have been 0.0656. If Fisher's mid-p or Barnard's test had been used, then the p-value would have been 0.0374 or 0.0352, respectively. The results of this trial, along with other preliminary data, were suggestive of an effect, and so a second study, SHOCK II (Dzavik et al., submitted), was conducted. Ironically, the SHOCK II Trial showed no evidence of a treatment effect, but there were significant differences between the SHOCK II Trial and the Cotter Trial.

The second example is taken from the A to Z Trial (Blazing et al., 2004). This trial compared enoxaparin with un-fractionated heparin for the treatment of 3905 patients with acute coronary syndrome (ACS). Based on other studies, there was a concern that enoxaparin might lead to an increase in the number of bleeding events. Given in Table 2 are the counts of patients with TIMI major bleeding events by treatment arm.

Note that the bleeding rates are quite low in both arms (less than one percent). The Statistical Analysis Plan specified that "Statistical comparison will be conducted using Fisher's exact test ..."

---

Vic Hasselblad earned a Ph. D. in Biostatistics in 1967 from the University of California at Los Angeles. An expert in the area of synthesis of evidence (meta-analysis), he coauthored a book and more than 20 peer-reviewed articles. E-mail: victor.hasselblad@duke.edu. Yuliya Lokhnygina is Assistant Professor of Biostatistics at Duke University. Her research interests include statistical methods in clinical trials, survival analysis, causal inference in observational studies and dynamic (adaptive) treatment strategies. E-mail: yuliya.lokhnygina@duke.edu

Table 1. Deaths by Treatment Arm in the L-NAME Trial

	L-NAME	No L-NAME	Total
Died	4	10	14
Survived	11	5	16
Total	15	15	30

Table 2. Bleeding Events by Treatment Arm for the A to Z Trial

	Enoxaparin	Un-fractionated Heparin	Total
Bleed	18	8	26
No bleed	1922	1957	3879
Total	1940	1965	3905

In this case, Fisher's exact test gives a p-value between 0.0285 and 0.0501. The problem with Fisher's exact test is that it is .0393 or 0.0352, respectively. It is clear that summarizing the results of the above table as non-significant would not accurately describe the information.

These two examples point out some of the difficulties in choosing a statistical test in the simplest of trials, namely the two-arm dichotomous trials. There are several possible tests that can be used and they have different implications for both the nominal alpha level as well as the power. We will restrict our consideration to those tests available in commercial software packages such as SAS<sup>®</sup>

(SAS Institute, 1999) or StatXact (StatXact with Cytel Studio, 2005).

#### Methodology

Assume a study where the number of positives and negatives are measured for a control group and a treated group, and that the results are summarized in a standard 2 x 2 contingency table where  $A$ ,  $B$ ,  $C$ , and  $D$  are the observed counts. Let  $T = A + B + C + D$ . The rate in the treated group,  $p_1$ , is estimated by  $A / N_1$  and the rate in the control group,  $p_2$ , is estimated by  $C / N_2$ . The null hypothesis is that  $p_1 = p_2$  and the usual alternative hypothesis is

	Treated	Control	Total
Positive	$A$	$C$	$S_1$
Negative	$B$	$D$	$S_2$
Total	$N_1$	$N_2$	$T$

that  $p_1 \neq p_2$ . The object is to find a test statistic which is a function of  $A$ ,  $B$ ,  $C$ , and  $D$  such that the value of the test is very different when  $p_1 = p_2$  as compared with when  $p_1 \neq p_2$ . There are several test statistics which could be used to test the null hypothesis, and the properties of five such tests will be investigated: 1) the uncorrected chi-squared test, 2) Fisher's exact test, 3) Yates' correction to the chi-squared test, 4) Fisher's exact mid-p test, and 5) Barnard's test.

#### Uncorrected chi-squared test

The standard uncorrected chi-squared statistic (Pearson, 1900) is:

$$CS = \frac{T(AD - BC)^2}{N_1 N_2 S_1 S_2}. \quad (1)$$

For an intended  $\alpha$ -level of 0.05, the test rejects the null hypothesis whenever  $CS > 3.8415$  and accepts otherwise. The power of the test is the probability that  $CS > 3.8415$  given particular values of  $p_1$ ,  $p_2$ ,  $N_1$ , and  $N_2$ .

#### Fisher's exact test

In 1925, Fisher (1925) gave an exact test which requires a bit more effort to compute. The test is based on the hyper geometric distribution. Assume that the four marginal totals,  $N_1$ ,  $N_2$ ,  $S_1$ , and  $S_2$ , are fixed. Under the null hypothesis, the probability that  $A = i$  for  $i = 0, 1, \dots, \min(N_1, S_1)$  is:

$$\text{Prob}(A|N_1, N_2, S_1, S_2) = \frac{\binom{S_1}{A} \binom{S_2}{N_1 - A}}{\binom{T}{N_1}}. \quad (2)$$

The (two sided) probability of an observed or more extreme than observed result is given by

$$\sum_{\text{Prob}(i|N_1, N_2, S_1, S_2) < \text{Prob}(A|N_1, N_2, S_1, S_2)} \text{Prob}(i|N_1, N_2, S_1, S_2) + \text{Prob}(A|N_1, N_2, S_1, S_2) \quad (3)$$

For example, the values for Cotter et al (2000) are  $0.0092 + 0.0564 = 0.0656$ . The two values, 0.0092 and 0.0656, are the only two reasonable values for the size of Fisher's exact test in this particular case (see Kendall and Stuart, Vol. 2, pp. 553, 1961). A non-randomized test cannot be constructed at any arbitrary level. But by convention, the largest value, 0.0656, is often taken as the p-value from the test. This value is often described as conservative, but it is only conservative if the object is to reject the null hypothesis. Thus, the null hypothesis would not be rejected at the 0.05 level using this test in this particular manner. The test could be made exact by choosing a random number between the values of 0.0092 and 0.0656 as the p-value. However, using randomization as part of the hypothesis testing procedure has never been accepted in clinical literature. This example demonstrates that using a conservative test is not necessarily a conservative strategy when the endpoint in question is a safety endpoint.

#### Yates' corrected chi-square test

The third test is Yates' (1934) correction to the Pearson chi-squared statistic:

$$CSC = \frac{T(|AD - BC| - T/2)^2}{NN_2S_1S_2} \tag{4}$$

This correction is designed to make the chi-squared statistic give a p-value which is often very close to the p-values calculated from Fisher's exact test.

Fisher's mid-p test

The fourth test is a modification of Fisher's exact test, known as Fisher's mid-p value, as defined by Lancaster (1961). The calculations are made exactly as those done for Fisher's exact test, except that the probability of a result more extreme is averaged with the probability of a result as extreme or more so. In the Cotter et al. (2000) example, this would be  $(0.0092 + 0.0656)/2 = 0.0374$ . StatXact (2005) and LogXact (LogXact with Cytel Studio, 2005) report mid-p values as part of their output.

Barnard's test

Barnard (1947) proposed an unconditional exact test based on a minimax elimination of the nuisance parameter. The reference set was defined to be the set of all 2 x 2 tables with fixed row margins and all possible column margins. Because the reference set for Barnard's test does not fix the column margins, the distribution of the test statistic is less discrete than would be obtained by permuting the conditional reference set in which both margins are fixed. However, Barnard was not satisfied with his test, and disavowed it two years later (Barnard, 1949). There is an interesting discussion by Barnard of the reasons for his disavowal in Yates (1984, with discussion). Barnard invoked Fisher's principle of ancillarity (see Fisher, 1973, Chapter IV), whereby inference should be based on hypothetical repetitions of the original experiment, fixing those aspects of the experiment that are unrelated to the hypothesis under test. Little (1989) gives a clear discussion of this topic. In two more recent publications, Barnard (1989, 1990) provided additional arguments against the

test. However, Little (1989) showed that the row totals are not ancillary statistics.

If the true value of p was known under the null hypothesis ( $p_1 = p_2 = p$ ), then the probability of any possible outcome could be calculated, e.g. the probability of  $x_1$  events in the first arm (of size  $N_1$ ), and  $x_2$  events in the second arm (of size  $N_2$ ):

$$\Pr(x_1, x_2) = \binom{N_1}{x_1} p^{x_1} (1-p)^{N_1-x_1} \binom{N_2}{x_2} p^{x_2} (1-p)^{N_2-x_2} \tag{5}$$

Next, order the outcomes. One possible ordering would be to use the D statistic:

$$D = \frac{\frac{x_2}{N_2} - \frac{x_1}{N_1}}{\sqrt{\left(\frac{x_1 + x_2}{N_1 + N_2}\right) \left(\frac{N_1 + N_2 - x_1 - x_2}{N_1 + N_2}\right) \left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \tag{6}$$

Using this ordering, the probabilities can be found of all tables at least as extreme, or more so, than the observed table for a given p. The sum of all these probabilities is the p-value associated with the specified p. Calculate this p-value for all possible specified p's and take their maximum. This is Barnard's p-value. A plot of the extreme values as a function of p for the Cotter et al (2000) example is in Figure 1.

Note that the statistic reaches a maximum of 0.0352, and this is Barnard's p-value for the Cotter et al study (2000). Barnard's test is actually guaranteed to be conservative for certain specific sample sizes. The reason that the test is not always conservative is that it uses a normal approximation to order the outcomes.

Power Formulas

The formula for the probability of rejection for any test of equality of proportions is given by:

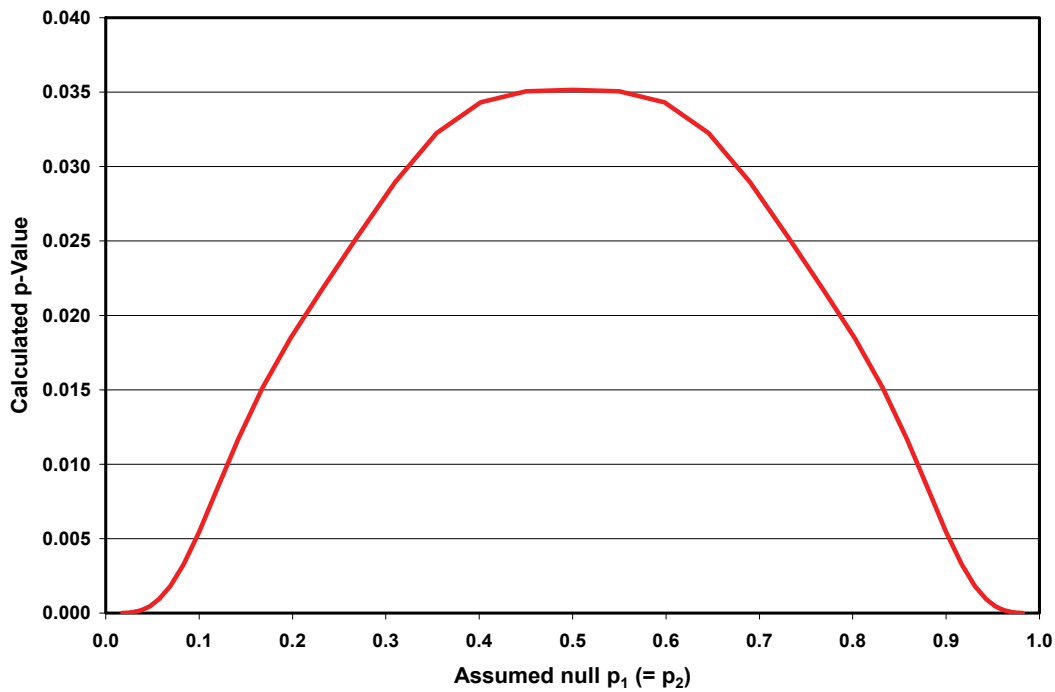


Figure 1. Calculation of Barnard’s Statistic for Specified Null Probabilities ( $p_1$ )

$$\Pr[reject]=$$

$$\sum_{i=0}^{N_1} \sum_{j=0}^{N_2} \delta_{ij} \binom{N_1}{i} \binom{N_2}{j} p_1^i (1-p_1)^{N_1-i} p_2^j (1-p_2)^{N_2-j} \quad (7)$$

where  $N_1$  and  $N_2$  are the sample sizes of the two arms respectively, where  $p_1$  and  $p_2$  are the true event rates in each arm, and where  $\delta_{ij}$  is one if the test statistic based on  $i, N_1, j, N_2$  is statistically significant, and zero otherwise.

This formula can be used to determine either the nominal alpha level for a given test (by assuming that  $p_1$  equals  $p_2$ ) or to determine the power (by not assuming equality). The formula is an exact one – no simulations are necessary. All results presented in the next section are exact calculations.

### Results

The actual alpha-levels are calculated for all five tests assuming that the intended alpha-level was 0.05 and  $N_1 = N_2 = 25, N_1 = N_2 = 50, N_1 = N_2 =$

100, and  $N_1 = 25, N_2 = 50$ . The calculations were made for the entire range of  $p_1$  (with  $p_2 = p_1$ ) and these are shown in Figures 2, 3, 4 and 5.

Note that the actual alpha-levels for the standard chi-square, Fisher’s mid-p, and Barnard’s tests are reasonably close to the intended alpha-level for  $0.2 < p_1 < 0.8$ . The maximum actual alpha-level for any test never exceeds .065 for any  $p_1$ . Note also that Fisher’s exact test has very low alpha-levels. The maximum alpha-level for Fisher’s exact test for  $N_1 = N_2 = 25$  is 0.0328. Yates’ correction to the standard chi-square test yields alpha levels as low as or lower than Fisher’s exact test. Fisher’s mid-p test falls below the nominal alpha level of 0.05 everywhere, but is uniformly larger than either Fisher’s exact or Yates’ correction. Barnard’s test is as large, or larger, than Fisher’s mid-p, but it does exceed 0.05 for event rates between 0.107 and 0.172 and between 0.828 and 0.893.

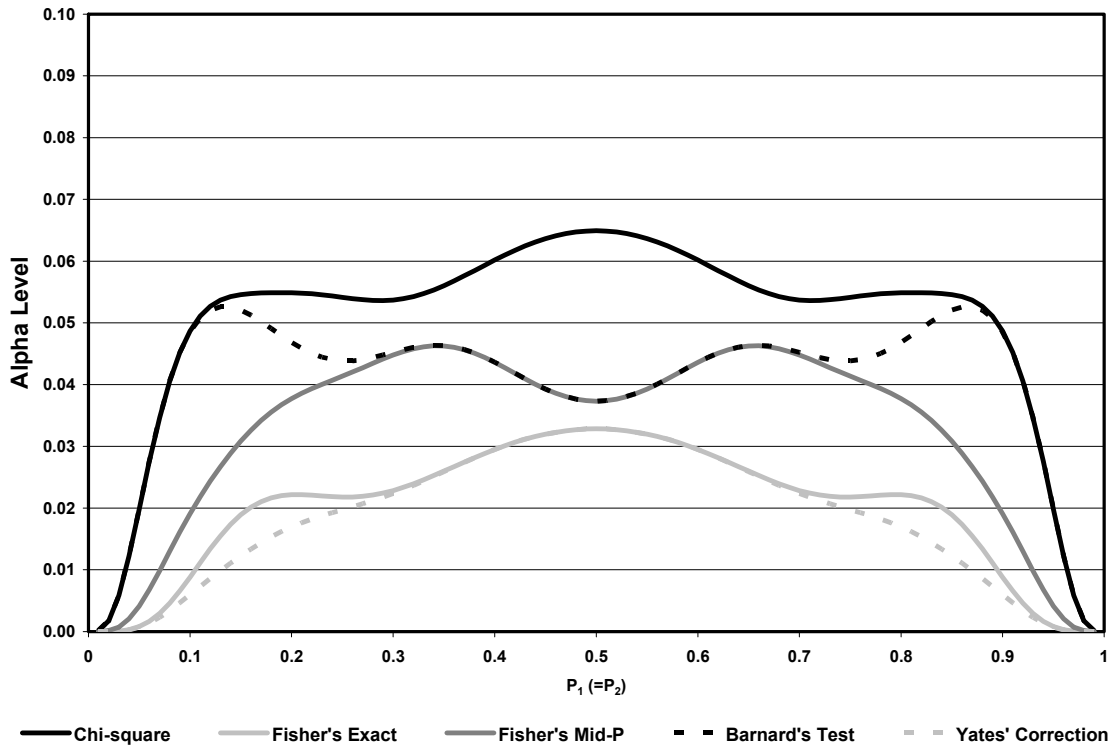


Figure 2. Alpha Levels for Two Arm Dichotomous Tests for  $N_1 = N_2 = 25$

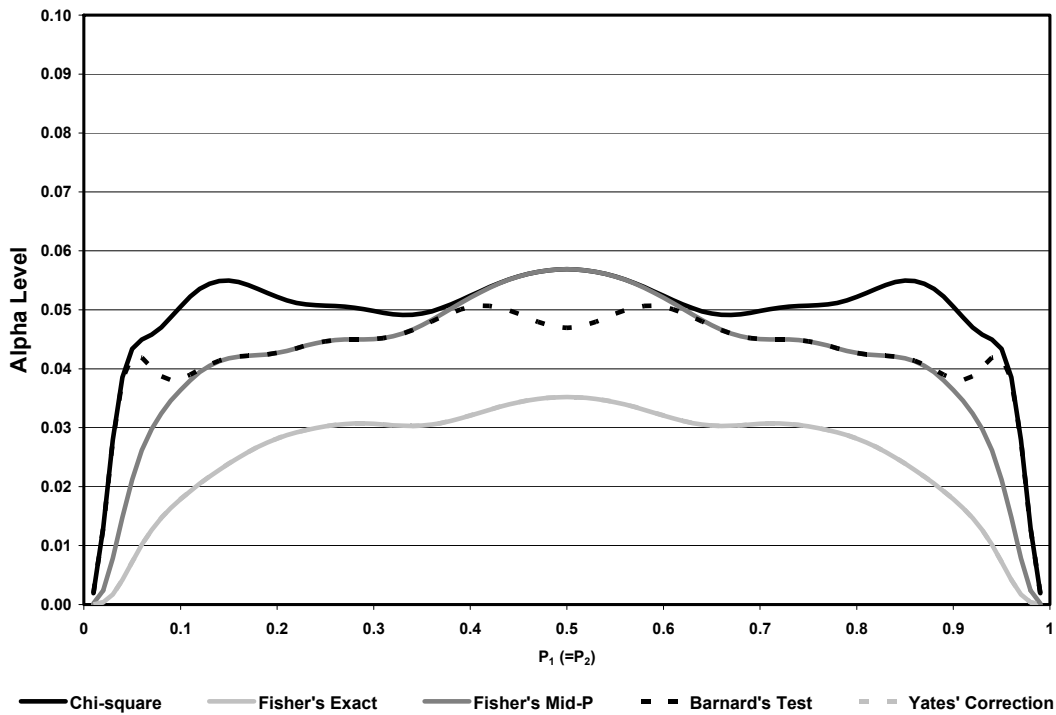


Figure 3. Alpha Levels for Two Arm Dichotomous Tests for  $N_1 = N_2 = 50$



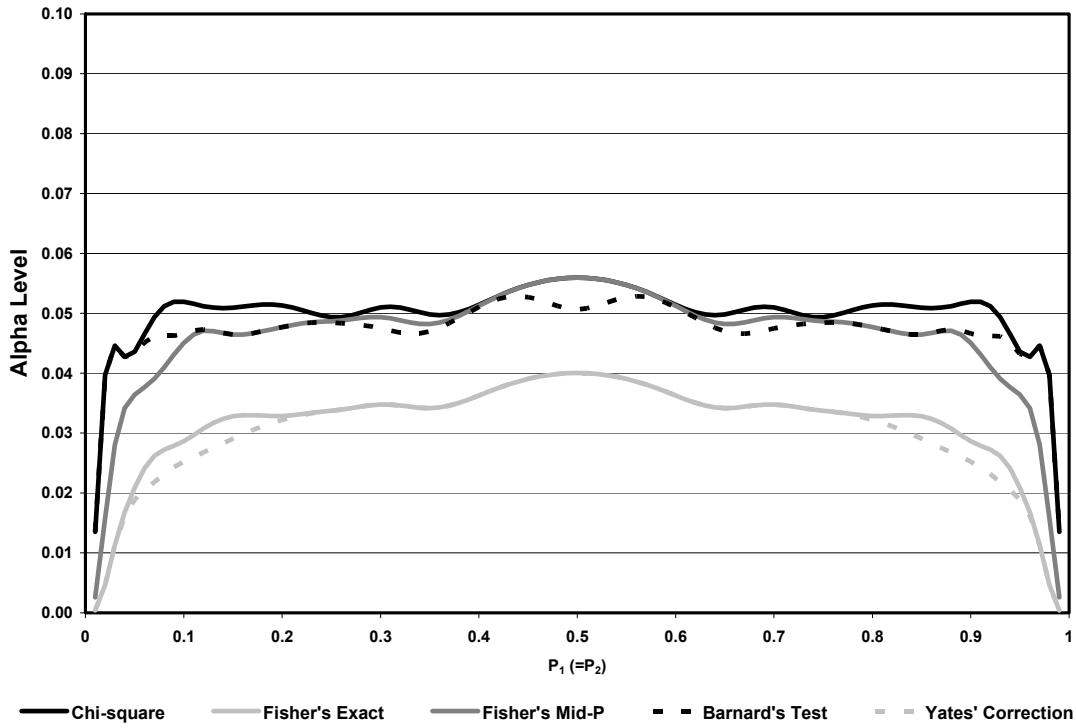


Figure 4. Alpha Levels for Two Arm Dichotomous Tests for  $N_1 = 100, N_2 = 100$

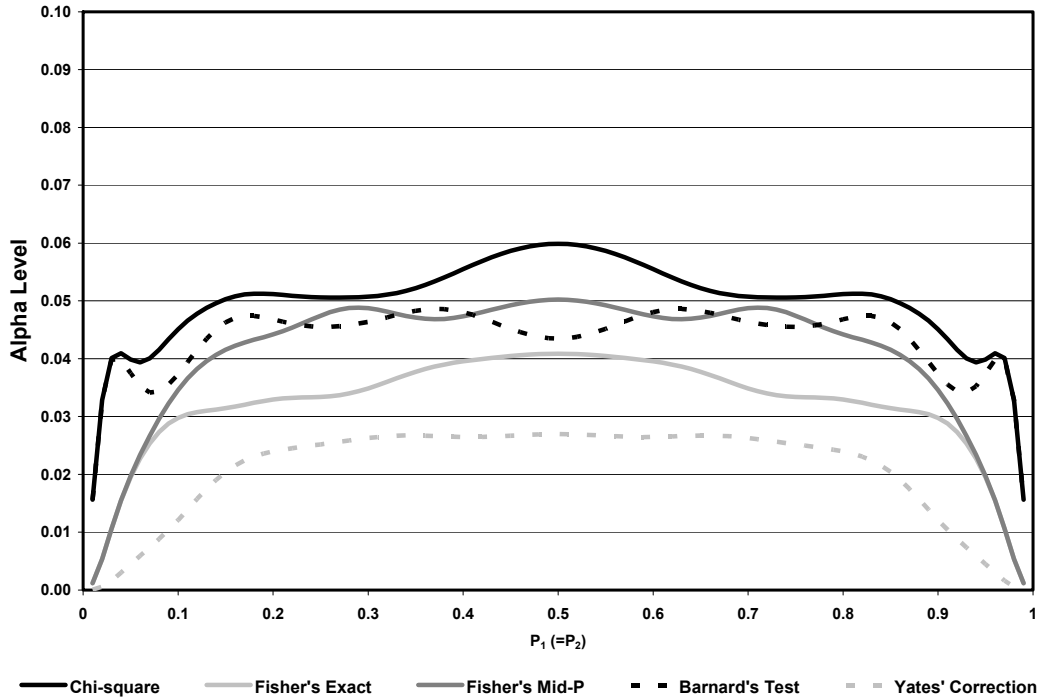


Figure 5. Alpha Levels for Two Arm Dichotomous Tests for  $N_1 = 25, N_2 = 50$

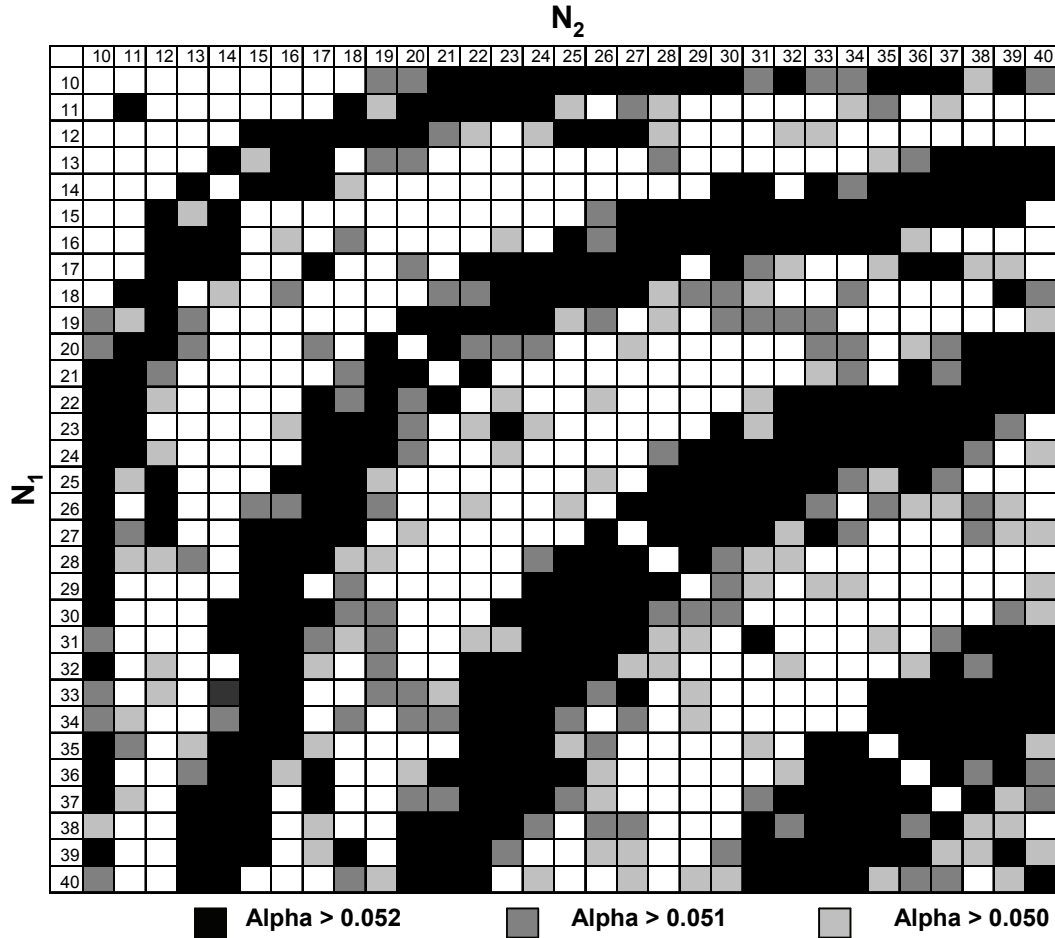


Figure 6. Diagram of Maximum Alpha Levels for Fisher's Mid-P Test with Nominal Alpha Level of 0.05.

For sample sizes of 50 per arm, the actual alpha-levels for the standard chi-square, Fisher's mid-p, and Barnard's tests approach the nominal alpha-level for  $0.2 < p_i < 0.8$ . The maximum actual alpha-level for any test never exceeds .057 for any  $p_i$ . Fisher's exact test still has very low alpha-levels, falling below 0.035 everywhere. Fisher's mid-p test remains below the nominal alpha level of 0.05 for event rates below 0.3, but does reach a maximum of 0.057. Barnard's test never exceeds 0.0507, and is generally closer to 0.05 than any of the other tests.

For sample sizes of 100 per arm, the actual alpha-levels for the standard chi-square, Fisher's mid-p, and Barnard's tests approach the nominal alpha-level for  $0.1 < p_i < 0.9$ . The maximum actual alpha-level for any test never exceeds .056 for any  $p_i$ . Fisher's exact test is increased, but still falls below 0.040 everywhere. Fisher's mid-p test falls below the nominal alpha level of 0.05 for event rates below 0.3, but does reach a maximum of 0.056. Barnard's test never exceeds 0.053, and is generally closer to 0.05 than any of the other tests.

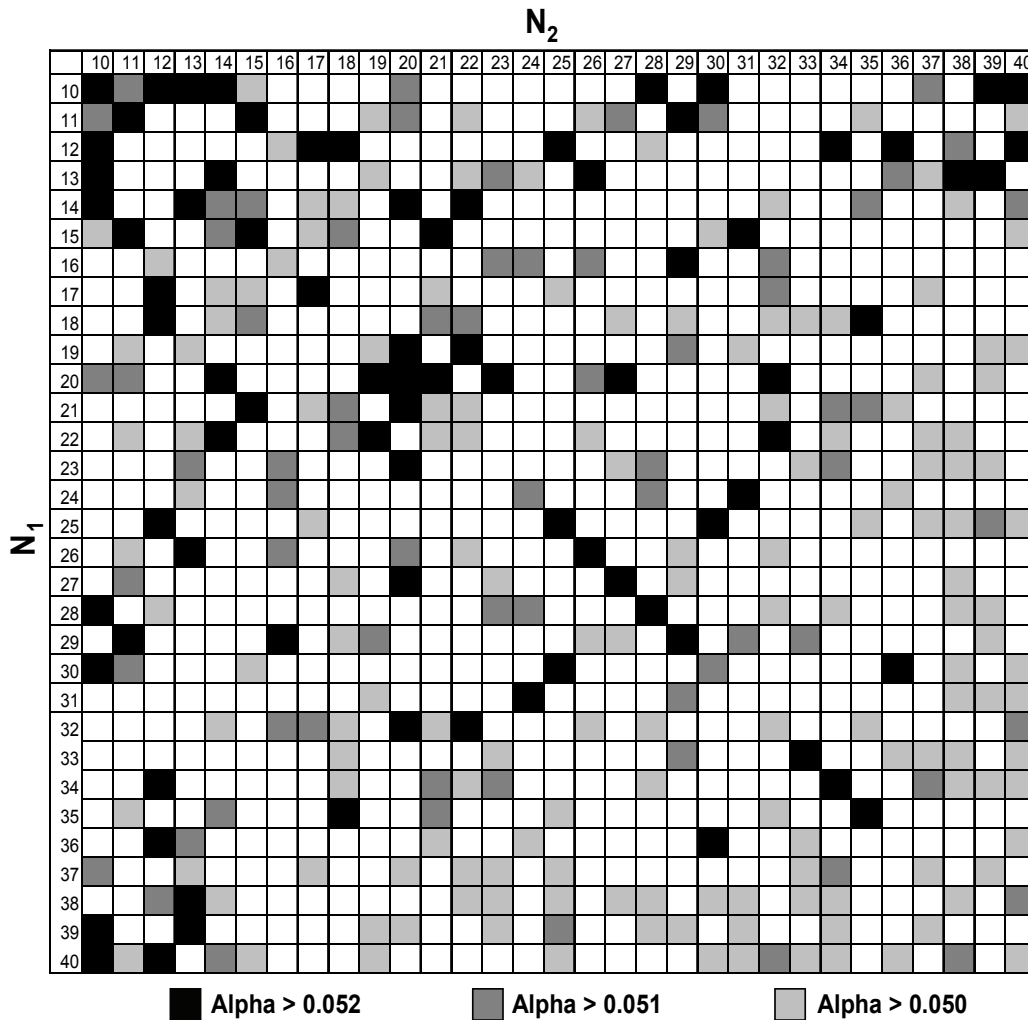


Figure 7. Diagram of Maximum Alpha Levels for Barnard's Test with Nominal Alpha Level of 0.05.

Fisher's exact test had alpha levels a bit closer to that of the other tests, but Yates' correction had very low alpha levels, achieving a maximum of 0.0270.

For unequal samples of 25 and 50 per arm, the results were somewhat similar to the previous results. Barnard's test had a maximum alpha level of 0.0484 and Fisher's mid-p test had a maximum alpha level of 0.0503. However, the chi-square test had a maximum of 0.0599.

The results from Figures 2 – 5 are consistent with the results presented by Hasselblad and Allen (2003). Their results

suggested that an expected number of events of approximately 40 is required to insure that the actual alpha level for the chi-square test is between 0.049 and 0.051 when the intended alpha level is 0.05.

Fisher's mid-p and Barnard's tests are examined in greater detail. Specifically, the interest is to determine if those tests were conservative for all values of  $p_1$  (with  $p_2 = p_1$ ) for specific values of  $N_1$  and  $N_2$ . The results for Fisher's mid-p for  $N_1 = 10, \dots, 40$  and  $N_2 = 10, \dots, 40$  are in Figure 6. Those squares which are white correspond to an actual alpha level less

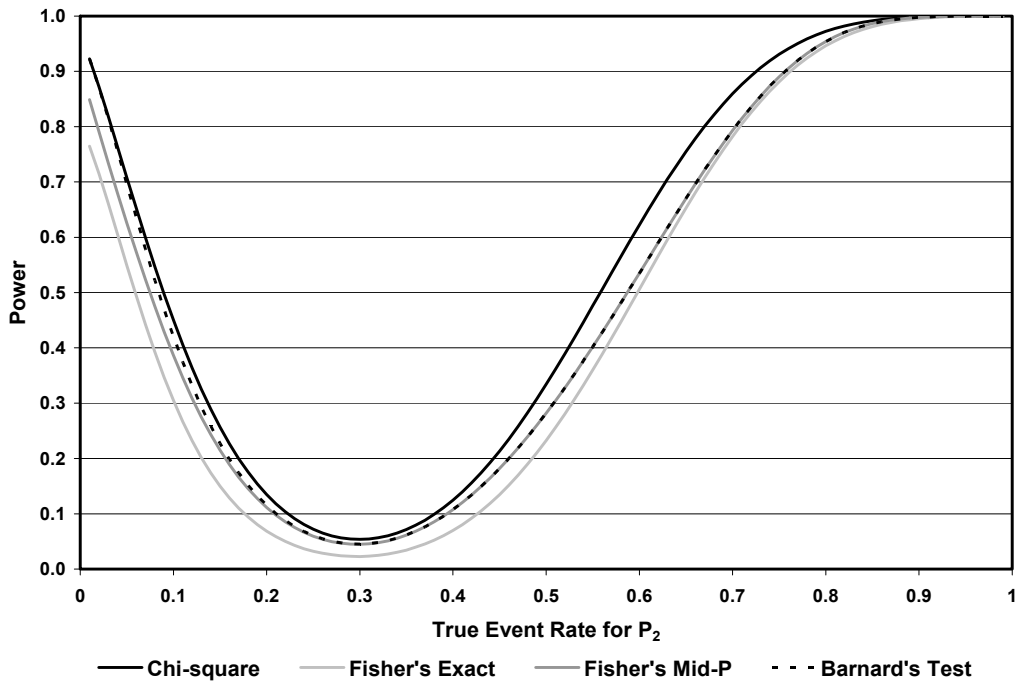


Figure 8. Power of Various Tests for Sample Sizes of 25 Per Arm

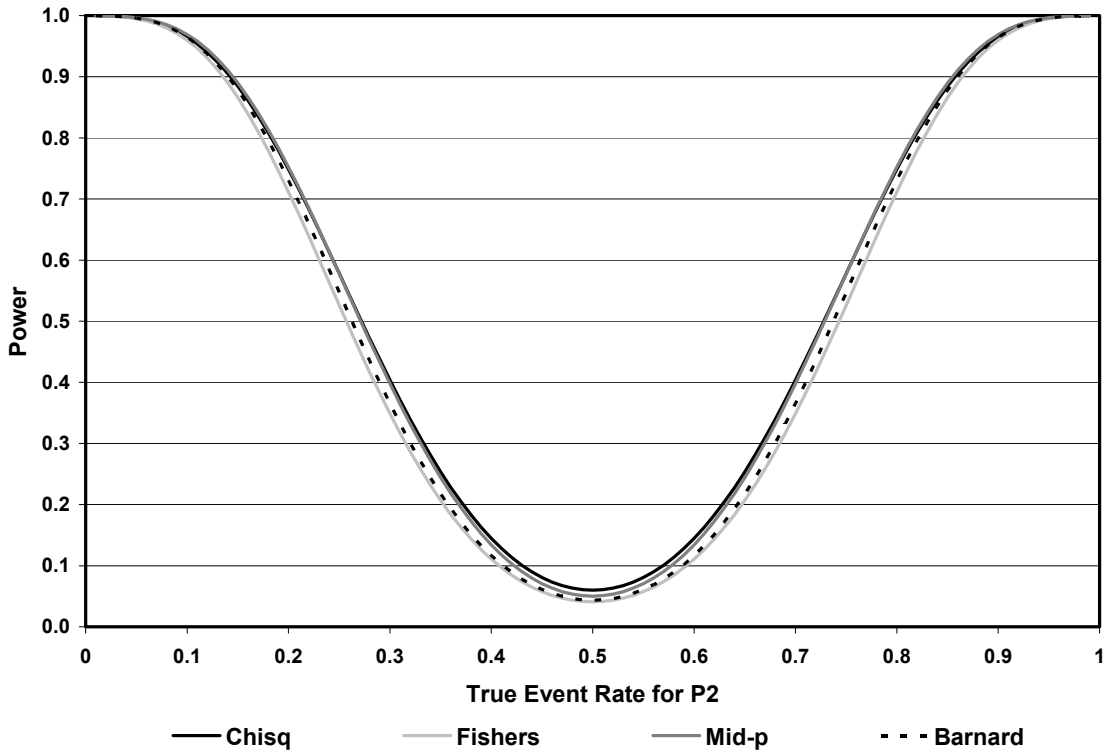


Figure 9. Power of Various Tests for Sample Sizes of 25 in the Control Arm And 50 in the Treated Arm

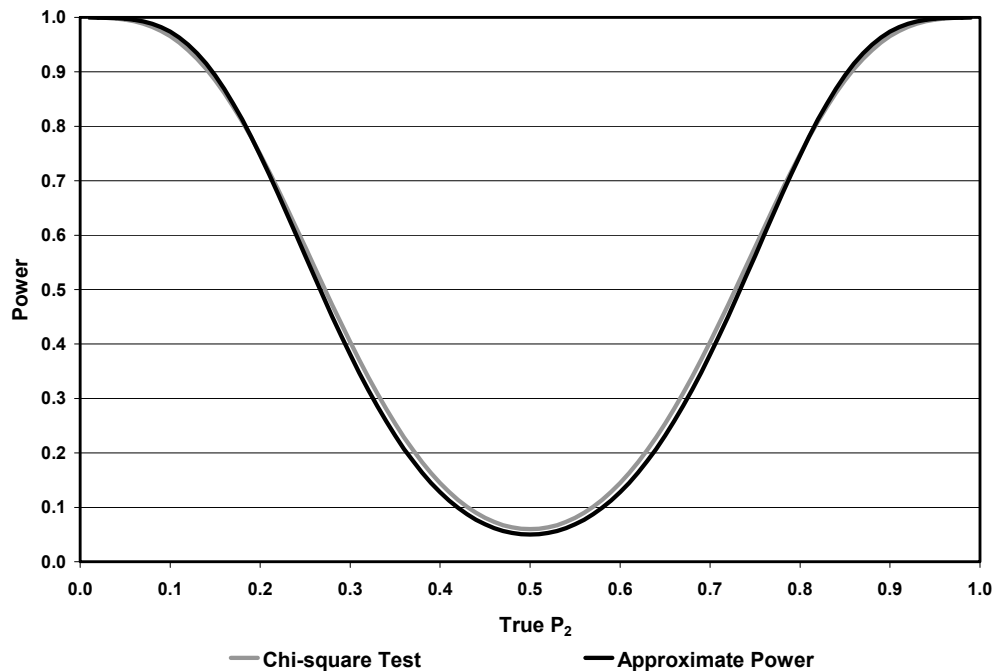


Figure 10. Estimated Power Using Approximation versus Actual Power of the Chi-square Test for Sample Sizes of 25 in the Control Arm and 50 in the Treated Arm

than 0.05 for all  $p$ .

For example, if one sample size is 15 and the other is 15, 16, ..., or 25, then Fisher's mid- $p$  test is conservative. On the other hand, if both sample sizes are 26, then the test may not be conservative, depending on the true null rate. However, for most null rates, the test will still be conservative. Figure 6 only shows the worst possible case. Of the 496 different sample size combinations shown in Figure 6, 40.9 percent had a nominal alpha level less than 0.05.

The results for Barnard's test for  $N_1 = 10, \dots, 40$  and  $N_2 = 10, \dots, 40$  are in Figure 7. For example, if one sample size is 25 and the other is 18, 19, ..., or 24, then Barnard's test is conservative. On the other hand, if both sample sizes are 25, then the test may not be conservative, depending on the true null rate. Of the 496 different sample size combinations shown in Figure 7, 66.5 percent had a nominal alpha level less than 0.05.

The power for four of the tests described previously was calculated for  $N_1 = N_2 = 25$  and  $p_1 = 0.3$  (Yates' test was dropped to make the

graph more readable). The results are in Figure 8.

Note that the power curves behave as expected, that is, they reach a minimum at  $p_1 = p_2 = 0.3$  and then increase rapidly as  $p_2$  moves away from  $p_1$ . The shapes of the power curves are all quite similar. The differences at  $p_1 = p_2 = 0.3$  are exactly the differences in the alpha-levels of the tests. The power curves show one other key point – the tests do not cross each other. That is, if a test has a lower nominal alpha level, then it will have lower power for the alternatives.

The power for four of the tests was also calculated for  $N_1 = 25, N_2 = 50$  and  $p_1 = 0.5$ . The results are in Figure 9.

Figure 9 shows the same general patterns as did Figure 8.

There are approximate formulas for power that are reasonably accurate. One formula given by Fleiss (1981, p. 27) is

$$\beta = \tag{8}$$

$$\Phi \left( \left( c_{\alpha/2} \sqrt{\bar{p}(1-\bar{p}) \left( \frac{1}{N_1} + \frac{1}{N_2} \right)} - (p_2 - p_1) \right) / \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}} \right)$$

where  $\bar{p} = (p_1 + p_2)/2$  and  $\Phi$  is the cumulative normal distribution function. This approximate function is shown in Figure 10, where it is drawn as a function of  $p_2$ . The exact and approximate formulas are reasonably similar, and they get closer as the sample size increases. There are several other formulas that have various correction formulas in order to make the approximation better. There is, however, a limit to the accuracy of these approximations because they are not based on the test statistic itself.

### Conclusion

There are some conclusions which can be made as a result of the calculations presented:

- Even though Fleiss (1981, p. 27) states that “[Yates’] correction should always be used”, the test is always inferior to (its nominal alpha level is less than or equal to) Fisher’s exact test, and for that reason it should not be used.
- Fisher’s exact test is so conservative that one should always look for an alternative even if one requires that the alpha level of the test not exceed the nominal level (by even the smallest amount). For certain sample sizes, either Fisher’s mid-p or Barnard’s test will satisfy the requirement, and those tests have much superior power. For example, knowing that the test is conservative when both arms have 15 observations, the data of Cotter et al. (2000) could have been analyzed using Fisher’s mid-p test.
- For tests of safety, being conservative is not desirable. Because event rates are often very low for safety issues, Fisher’s mid-p test is a very appealing alternative. For example, the maximal nominal alpha level for this test for the

A to Z bleeding data is 0.05007 (assuming that the true event rates are less than 20 percent).

- The chi-square test works adequately for very large sample sizes, but the standard rule of an expected minimum value of 5 (which is commonly used) is not acceptable. Even if the expected number of counts exceeds 40 per cell, the alpha level (for a nominal alpha level of 0.05) is approximately bounded by 0.049 and 0.051. Barnard’s test is certainly an attractive alternative in the moderate sample size situation when the event rates are not especially small.

As mentioned previously, only tests available in widely used commercial software packages were considered. Such restrictions leave out some recently developed unconditional tests for which no commercially developed and tested software is available. An example is a test based on the confidence interval p-value developed by Berger and Boos (1994, 1996). This test can be seen as a modification of Barnard’s test. Although Barnard’s p-value is obtained by maximizing the p-value for given nuisance parameter  $p$  over the unit interval, the p-value of the test by Berger and Boos is obtained as a sum of the supremum of p-values over the  $100(1-\beta)\%$  confidence interval for  $p$  calculated from the data and  $\beta$ . This test can be more powerful than Barnard’s and requires less computational effort.

### Acknowledgements

The authors wish to thank Paula Smith for manuscript editing and formatting assistance.

### References

Barnard, G.A. (1947). Significance tests for 2 x 2 tables. *Biometrika*, 34(1-2), 123-138.  
 Barnard, G.A. (1949). Statistical inference. *Journal of the Royal Statistical Society, Series B*, 11, 115-139.  
 Barnard, G.A. (1989). On alleged gains in power from lower p-values. *Statistics in Medicine* 8, 1469-1477.

Barnard, G.A. (1990). Must clinical trials be large? The interpretation of p-values and the combination of test results. *Statistics in Medicine*, 9, 601-614.

Berger, R.L. & Boos, D.D. (1994). P-values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89, 1012-1016.

Berger, R.L. (1996). More powerful tests from confidence interval p-values. *The American Statistician*, 50, 314-318.

Blazing, M.A., de Lemos, J.A., White, H.D., Fox, K.A.A., Verheugt, F.W.A., Ardissino, D., DiBattiste, P.M., Palmisano, J., Bilheimer, D.W., Snapinn, S.M., Ramsey, K.E., Gardner, L.H., Hasselblad, V., Pfeffer, M.A., Lewis, E.F., Braunwald, E., & Califf, R.M., for the A to Z Investigators (2004). Safety and efficacy of enoxaparin vs unfractionated heparin in patients with non-ST-segment elevation acute coronary syndromes who receive tirofiban and aspirin: a randomized controlled trial. *Journal of the American Medical Association*, 292, 55-64.

Cotter, G., Kaluski, E., Blatt, A., Milovanov, O., Moshkovitz, Y., & Zaidenstein, R. (2000). L-NMMA (a nitric oxide synthase inhibitor) is effective in the treatment of cardiogenic shock. *Circulation*, 101(12), 1358-1361.

Cytel Inc. (2005). *StatXact 7 with cytel studio statistical software for exact nonparametric inference*. Cambridge, MA.

Cytel Inc. (2005). *LogXact 7 with Cytel Studio. Discrete Regression Software Featuring Exact Methods*. Cambridge, MA.

Fisher, R.A. (1973). *Statistical Methods and Scientific Inference, Third Edition*. London: Collier Macmillan Publishers.

Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*. New York: John Wiley & Sons.

Hasselblad, V., & Allen, A.S. (2003). Power calculations for large multi-arm placebo-controlled studies of dichotomous outcomes. *Statistics in Medicine*, 22, 1943-1954.

Kendall, M.G., & Stuart, A. (1961). *The advanced theory of statistics vol. 2*. London: Charles Griffin & Company Limited.

Lancaster, H.O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association*, 56, 223-234.

Little, R.J.A. (1989). Testing the equality of two independent binomial proportions. *The American Statistician*, 43, 283-288.

Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(5), 157-172.

SAS Institute Inc. (1999). *SAS/STAT® user's guide, version 8*. Cary, NC.

Yates, F. (1934). Contingency tables involving small numbers and the  $\chi^2$  test. *Journal of the Royal Statistical Society, (Supp 1)*, 217-235.

Yates, F. (1984). Tests of significance for 2 x 2 contingency tables (with discussion). *Journal of the Royal Statistical Society, Series A*, 147, 426-463.