Journal of Biopharmaceutical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lbps20

Testing Homogeneity of Stratum Effects in Stratified Paired Binary Data

Yan D. Zhao[a], Dewi Rahardja[b], De-Hui Wang[c] & Haili Shen[d]
[a] Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, USA
[b] US Food and Drug Administration, Silver Spring, Maryland, USA
[c] College of Mathematics, Jilin University, Changchun, Jilin, China
[d] Department of Pain, The Second Hospital of Lanzhou University, Lanzhou, Gansu, China
Accepted author version posted online: 03 Apr 2014.Published online: 15 Apr 2014.

To cite this article: Yan D. Zhao, Dewi Rahardja, De-Hui Wang & Haili Shen (2014) Testing Homogeneity of Stratum Effects in Stratified Paired Binary Data, Journal of Biopharmaceutical Statistics, 24:3, 600-607, DOI: 10.1080/10543406.2014.888440

To link to this article: http://dx.doi.org/10.1080/10543406.2014.888440

Journal of Biopharmaceutical Statistics, 24: 600–607, 2014

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2014.888440

Taylor & Francis
Taylor & Francis Group

TESTING HOMOGENEITY OF STRATUM EFFECTS IN STRATIFIED PAIRED BINARY DATA

Yan D. Zhao[1], Dewi Rahardja[2], De-Hui Wang[3], and Haili Shen[4]

[1]*Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, USA*
[2]*US Food and Drug Administration, Silver Spring, Maryland, USA*
[3]*College of Mathematics, Jilin University, Changchun, Jilin, China*
[4]*Department of Pain, The Second Hospital of Lanzhou University, Lanzhou, Gansu, China*

For paired binary data, McNemar's test is widely used to test marginal homogeneity or symmetry for a 2 by 2 contingency table. In this article, we extend McNemar's test by considering a series of paired binary data in which the series is defined by a stratification factor. We provide a test for testing homogeneous stratum effects. For illustration, we apply our test to a cancer epidemiology study. Finally, we conduct simulations to show that our test preserves the nominal type I error level and evaluate the power of our test under various scenarios.

Key Words: Bowker's test; Marginal homogeneity; McNemar's test; Paired binary data; Stratified test; Stuart–Maxwell test; Symmetry.

1. INTRODUCTION

In this article we consider binary data collected on matched pairs. The sampling unit is not one individual but a pair of related individuals, which could be two parts of or two occasions for the same individual. For example, the binary response is a voter's choice from two presidential candidates and the two occasions could be two different time points before the presidential election.

Let random variables X and Y be the binary responses (1 or 2) of the matched pairs, and n_{ij} be the number of pairs with $X = i$ and $Y = j$, where $i, j \in \{1, 2\}$. Also, let $p_{ij} = \Pr(X = i, Y = j)$; then McNemar's test (McNemar, 1947) is commonly used to test the following null hypothesis of homogeneity: $H_0 : p_{1\bullet} = p_{\bullet 1}$, where $p_{1\bullet} = p_{11} + p_{12}$ and $p_{\bullet 1} = p_{11} + p_{21}$, or equivalently, the null hypothesis of symmetry, $H_0 : p_{12} = p_{21}$.

When the categorical random variables X and Y take K ($K > 2$) values, Bowker's test (Bowker, 1948) allows for testing the null hypothesis of symmetry $H_0 : p_{ij} = p_{ji}$, for all $i \neq j$, where $i, j \in \{1, 2, \ldots, K\}$. Additionally, the Stuart–Maxwell test (Stuart, 1955; Maxwell, 1970) provides for testing the null hypothesis of marginal homogeneity

Received March 27, 2012; Accepted February 28, 2013
Address correspondence to Haili Shen, Department of Pain, The Second Hospital of Lanzhou University, Lanzhou, Gansu 730030, China; E-mail: shenhl@lzu.edu.cn

$H_0 : p_{i\bullet} = p_{\bullet i}$, where $i = 1, 2, \ldots, K$. Note that for $K > 2$, the null hypothesis of symmetry is not equivalent to the null hypothesis of homogeneity. In fact, rejection of marginal homogeneity implies rejection of symmetry, but not vice versa. Therefore, practitioners need to decide which hypothesis to test for a particular application.

We now consider applications with a series of independent paired binary data that is defined by a stratification factor. We aim to provide a test for testing homogeneous stratum effects in this case. In analogy, this is similar to the Breslow–Day test (Breslow and Day, 1980) for homogeneous odds ratios across a series of stratified 2 by 2 contingency tables in which the binary data are unpaired.

For example, we consider an epidemiological study in which patients with breast cancer were allocated based on physician's choice to one of two treatment groups: chemotherapy or control (no chemotherapy). Each patient's estrogen receptor (ER) biomarker status (positive or negative) was assessed at two time points: before and after the treatment (chemotherapy or control). The research hypothesis is to test whether the change in biomarker status is the same for the chemotherapy and control groups. In this article we develop a testing procedure for such a hypothesis.

The remainder of the article is organized as follows. In section 2, we describe the data and derive a statistic for testing homogeneous stratum effects in stratified paired binary data. In section 3, we illustrate the test by using a cancer epidemiological study. The type I error rates of the test are examined using simulations in section 4, and a discussion can be found in section 5.

2. DATA AND TEST

As we introduced in section 1, we consider a series of independent paired binary data in which the series is defined by a stratification factor. For the kth stratum, $k = 1, \ldots, K$, let random variables X_{k} and Y_{k} be the binary responses (1 or 2) of the matched pairs, cell counts n_{ijk} be the number of pairs with $X_{\mathrm{k}} = i$ and $Y_{\mathrm{k}} = j$, cell proportions $p_{ijk} = \Pr(X_{\mathrm{k}} = i, Y_{\mathrm{k}} = j)$, where $i, j \in \{1, 2\}$. Furthermore, we adopt a common notation of a dot in the subscript that denotes the summation over the subscript. For example, $n_{1\bullet k} = n_{11k} + n_{12k}$. The data for the kth stratum are displayed in Table 1.

The null hypothesis of interest is to test the homogeneous stratum effects. In mathematical terms, we are interested in testing $H_0 : p_{1\bullet 1} - p_{\bullet 11} = \cdots = p_{1\bullet K} - p_{\bullet 1K}$, or equivalently, $H_0 : p_{121} - p_{211} = \cdots = p_{12K} - p_{21K}$. In the matrix form, this null hypothesis is

$$H_0 : \mathbf{A}\delta = \mathbf{0}, \tag{1}$$

Table 1 Cross-classification contingency table for the kth stratum

	Y_k		
X_k	1	2	Total
1	$n_{11\mathrm{k}}$ ($p_{11\mathrm{k}}$)	$n_{12\mathrm{k}}$ ($p_{12\mathrm{k}}$)	$n_{1\bullet k}$ ($p_{1\bullet k}$)
2	$n_{21\mathrm{k}}$ ($p_{21\mathrm{k}}$)	$n_{22\mathrm{k}}$ ($p_{22\mathrm{k}}$)	$n_{2\bullet k}$ ($p_{2\bullet k}$)
Total	$n_{\bullet 1k}$($p_{\bullet 1k}$)	$n_{\bullet 2k}$($p_{\bullet 2k}$)	$n_{\bullet\bullet k}$($p_{\bullet\bullet k}$)

where

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & -1 \end{bmatrix}, \delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_K \end{bmatrix},$$

and $\delta_k = p_{12k} - p_{21k}, k = 1, \cdots, K$.

Next, we develop a test statistic for H_0 in Eq. (1). Because the cell counts (n_{11k}, n_{12k}, n_{21k}, n_{22k}) in Table 1 have a multinomial distribution with the parameters $n_{\bullet\bullet k}$ and (p_{11k}, p_{12k}, p_{21k}, p_{22k}), the maximum likelihood estimator (MLE) for p_{ijk} is $\hat{p}_{ijk} = n_{ijk}/n_{\bullet\bullet k}$ and the MLE for δ_k is $\hat{\delta}_k = \hat{p}_{12k} - \hat{p}_{21k}$. Clearly, we have

$$E\left(\hat{\delta}_k\right) = \delta_k.$$

Using the property of a multinomial distribution, we have $V\left(n_{ijk}\right) = n_{\bullet\bullet k}\, p_{ijk}(1 - p_{ijk})$ and $COV\,(n_{12k}, n_{21k}) = -n_{\bullet\bullet k} p_{12k} p_{21k}$. Therefore, we obtain

$$\sigma_k^2 \equiv V\left(\hat{\delta}_k\right) = n_{\bullet\bullet k}^{-1}\left(p_{12k} + p_{21k} - (p_{12k} - p_{21k})^2\right).$$

Also, asymptotically, $\hat{\delta}_k$ is normally distributed with a mean δ_k and a variance σ_k^2.

We define $\hat{\boldsymbol{\delta}} = \left(\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_K\right)$. Because $\hat{\delta}_1$, $\hat{\delta}_2$, ..., and $\hat{\delta}_K$ are independent, the vector $\hat{\boldsymbol{\delta}}$ has an asymptotic normal distribution with mean $\boldsymbol{\delta}$ and variance $\boldsymbol{\Delta}$, a diagonal matrix with σ_k^2 being the kth diagonal element. Consequently, the vector $\mathbf{A}\hat{\boldsymbol{\delta}}$ has an asymptotic normal distribution with the mean $E(\mathbf{A}\hat{\boldsymbol{\delta}}) = \mathbf{A}\boldsymbol{\delta}$ and the variance $\Sigma \equiv V(\mathbf{A}\hat{\boldsymbol{\delta}}) = \mathbf{A}\boldsymbol{\Delta}\mathbf{A}^T$.

For the H_0 in Eq. (1), we have $\mathbf{A}\boldsymbol{\delta} = \mathbf{0}$. There remains to be derived an estimator for σ_k^2. For the H_0, all of the δ_k are equal to the common risk difference, which we denote as δ. Consequently, $\sigma_k^2 = n_{\bullet\bullet k}^{-1}\left(p_{12k} + p_{21k} - \delta^2\right)$. First, we can simply estimate $p_{12k} + p_{21k}$ by using $\hat{p}_{12k} + \hat{p}_{21k}$. Then, for δ, we consider the estimators $\hat{\delta}$ of δ in the forms of a weighted average of $\hat{\delta}_1, \ldots, \hat{\delta}_K$, namely,

$$\hat{\delta} = \sum_{k=1}^{K} w_k \hat{\delta}_k \Big/ \sum_{k=1}^{K} w_k,$$

where w_k terms are the weights. By the Cauchy theorem, if we choose $w_k = 1/\tilde{\sigma}_k^2$ where $\tilde{\sigma}_k^2 = n_{\bullet\bullet k}^{-1}\left(\hat{p}_{12k} + \hat{p}_{21k} - \left(\hat{p}_{12k} - \hat{p}_{21k}\right)^2\right)$, then the resulting $\hat{\delta}$ will approximately have the smallest variance among all of the linear estimators. Note that if any of the $\tilde{\sigma}_k^2$ terms is zero, we can simply set all of the w_k to be 1 (i.e., equal weights). Finally, we estimate σ_k^2 by using $\hat{\sigma}_k^2 = n_{\bullet\bullet k}^{-1}\left(\hat{p}_{12k} + \hat{p}_{21k} - \hat{\delta}^2\right)$. If this formula results in a negative $\hat{\sigma}_k^2$, then we use $\hat{\sigma}_k^2 = \tilde{\sigma}_k^2$. Then the statistic $T \equiv (\mathbf{A}\hat{\boldsymbol{\delta}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{A}\hat{\boldsymbol{\delta}})$ has an asymptotic chi-squared distribution with $(K-1)$ degrees of freedom, where $\hat{\boldsymbol{\Sigma}} = \mathbf{A}\hat{\boldsymbol{\Delta}}\mathbf{A}^T$ and $\hat{\boldsymbol{\Delta}}$ is a $K \times K$ diagonal matrix with $\hat{\sigma}_k^2$ being the kth diagonal element. A level α test of the H_0 is given

by rejecting the H_0 if $T > \chi^2_{K-1;1-\alpha}$, where $\chi^2_{K-1;1-\alpha}$ is the $(1-\alpha)$ th quantile of the chi-squared distribution with $(K-1)$ degrees of freedom. The appendix has the R code for computing this test statistic and associated p-value.

3. EXAMPLE

In this section we provide an example of the use of the developed test in an epidemiological study of breast cancer patients. There were 232 breast cancer patients in the study with 133 and 99 patients treated with or without chemotherapy. The decision for patient allocation to the treatment or control group was not random but was determined based on the oncologists' opinions and with patients' agreement.

One research objective of this study was to examine whether the change in biomarker status was the same in the chemotherapy and control groups. One biomarker of particular interest was the ER, which had been shown to be a predictive and prognostic factor for breast cancer survival. For each patient, the ER status (positive, or ER+; negative, or ER−) was determined before and after the treatment (chemotherapy or control). Random variables X_1 and Y_1 are defined as the ER status before and after treatment for patients in the chemotherapy group, respectively. Random variables X_2 and Y_2 are similarly defined for patients in the control group. A value of 1 (or 2) for these four random variables indicated ER− (or ER+). The number and proportions for the cross-classification of ER status for both groups are presented in Table 2.

Using the notations introduced in section 2, the research null hypothesis to be tested was $H_0 : p_{121} - p_{211} = p_{122} - p_{212}$. In other word, the differences in proportion of changing from ER+ to ER− and proportion of changing from ER− to ER+ were the same for chemotherapy and control groups. From Table 2, the differences were −7.5% (= .8% − 8.3%) and 1.0% (= 1.0% – 0.0%) for the chemotherapy and control groups, respectively. Thus, a numerical difference existed. Then, using the test developed in section 2, we computed the test statistic $T = 9.32$ with a p-value of 0.002. Therefore, we concluded that there was a statistically significant difference in the change in ER status in the women allocated to the chemotherapy and control groups. In the chemotherapy group, the proportion of patients changing from ER+ to ER− was higher than the proportion of patients changing from ER− to ER+, while these two proportions were about the same in the control group.

Table 2 Cross-classification contingency table for ER status

		Y_1		
Stratum	X_1	1 (ER−)	2 (ER+)	Total
Chemotherapy	1 (ER−)	54 (40.6%)	1 (0.8%)	55 (41.4%)
	2 (ER+)	11 (8.3%)	67 (50.4%)	78 (58.7%)
	Total	65 (48.9%)	68 (51.1%)	133
		Y_2		
	X_2	1 (ER−)	2 (ER+)	Total
Control	1 (ER−)	18 (18.2%)	1 (1.0%)	19 (19.2%)
	2 (ER+)	0 (0.0%)	80 (80.8%)	80 (80.8%)
	Total	18 (18.2%)	81 (81.8%)	99

4. SIMULATION

We report the results from two simulation studies. In the first simulation study, we examined whether our test of homogeneous stratum effects preserved the nominal type I error rate under various scenarios. In the second simulation study, we evaluated the power of our test. In the simulations, we considered the two-sided nominal α level to be 5% and fixed the number of strata to 2. We chose two stratum fractions (SF = (0.2, 0.8), and (0.5, 0.5)), which were defined as the fractions of stratum sample sizes over the total sample size. For each simulation scenario, we generated 10,000 data sets.

In the first simulation study, we used three values for the total sample size (N = 50, 100, and 200). We chose three sets of contingency table proportions as displayed in Table 3 (Set 1 to Set 3). The proportions were chosen under the null hypothesis (1) which indicated no stratum effect ($p_{121} - p_{211} = p_{122} - p_{212}$). In Set 1, the contingency tables were exactly the same for two strata; in Set 2, the contingency tables were slight different for two strata; in Set 3, the contingency tables were largely different for two strata. In total, there were 18 simulation scenarios under the null hypothesis (Eq. 1).

The empirical rejection rates of the null hypothesis and associated standard errors were computed and are displayed in Table 4. Our test preserved the nominal type I error rate when the total sample size was 200 for both stratum fractions (Table 4). Our test continued to preserve the nominal type I error rate when the total sample size was 100 and the stratum

Table 3 Cross-classification contingency tables for simulations

Set	Stratum		Y_1		Set	Stratum		Y_1	
1	1	X_1	1	2	4	1	X_1	1	2
		1	.1	.2			1	.1	.1
		2	.1	.6			2	.1	.7
			Y_2					Y_2	
	2	X_2	1	2		2	X_2	1	2
		1	.1	.2			1	.1	.2
		2	.1	.6			2	.1	.6
Set	Stratum		Y_1		Set	Stratum		Y_1	
2	1	X_1	1	2	5	1	X_1	1	2
		1	.1	.2			1	.1	.1
		2	.1	.6			2	.1	.7
			Y_2					Y_2	
	2	X_2	1	2		2	X_2	1	2
		1	.1	.3			1	.1	.4
		2	.2	.4			2	.1	.4
Set	Stratum		Y_1		Set	Stratum		Y_1	
3	1	X_1	1	2	6	1	X_1	1	2
		1	.1	.2			1	.1	.1
		2	.1	.6			2	.1	.7
			Y_2					Y_2	
	2	X_2	1	2		2	X_2	1	2
		1	.1	.4			1	.1	.6
		2	.3	.2			2	.1	.2

Table 4 Empirical rejection rates (standard error) of the null hypothesis for two-sided nominal $\alpha = 5\%$

Set	Stratum fraction	Total sample size		
		50	100	200
1	(0.2, 0.8)	5.58 (0.23)	4.93 (0.22)	5.10 (0.22)
	(0.5, 0.5)	5.22 (0.22)	5.00 (0.22)	5.11 (0.22)
2	(0.2, 0.8)	5.48 (0.23)	5.12 (0.22)	4.76 (0.21)
	(0.5, 0.5)	5.64 (0.23)	5.09 (0.22)	4.71 (0.21)
3	(0.2, 0.8)	5.30 (0.22)	5.12 (0.22)	4.90 (0.22)
	(0.5, 0.5)	5.43 (0.23)	5.05 (0.22)	5.08 (0.22)

fraction was (0.5, 0.5). These conclusions were consistent across three sets of contingency tables.

In the second simulation study, for the total sample size N, we used values from 50 to 200 with an increment of 10. We chose three sets of contingency table proportions as displayed in Table 3 (Set 4 to Set 6). The proportions were chosen under the alternative hypothesis, which indicated a stratum effect ($p_{121} - p_{211} \neq p_{122} - p_{212}$). In Set 4, Set 5, and Set 6, the contingency tables were slightly, moderately, and largely different for two strata, respectively.

The power curves are plotted in Fig. 1. As expected, the power curves increased as the sample size increased. Comparing the left and right panels of Fig. 1 for the same set of contingency tables, the stratum fractions only slightly affected the power curves, with the balanced stratum fractions mostly having larger powers. For the Set 6 contingency tables in which there was a large difference between the two strata, a total sample size of 50 was needed to achieve 80% power; for Set 5 contingency tables in which there was a moderate difference between two strata, a total sample size of 150 was needed to achieve 80% power;

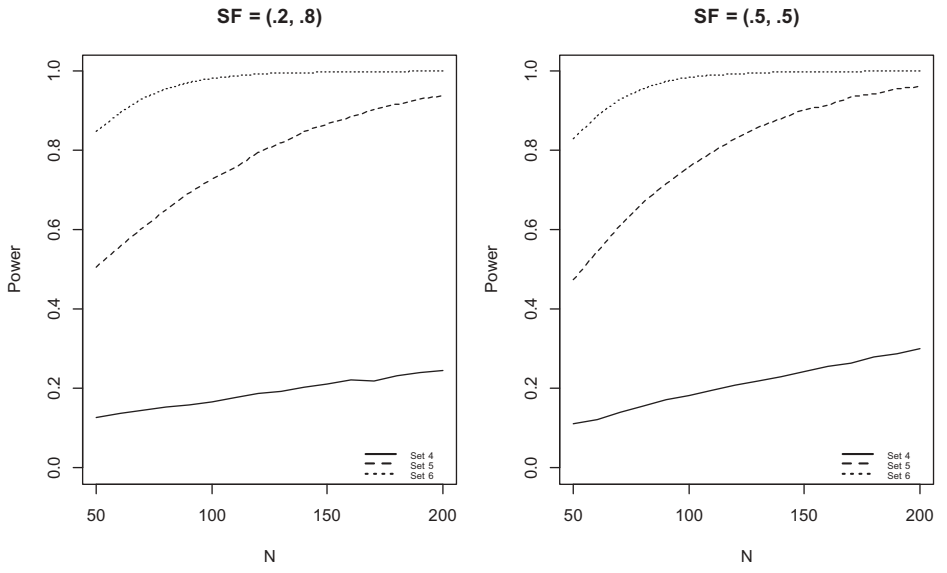

Figure 1 Power curves for our test.

for Set 4 contingency tables in which there was a small difference between two strata, a total sample size of 1000 was needed to achieve 80% power.

5. DISCUSSION

The Breslow–Day test is useful for testing the homogeneity of odds ratios for a series of 2 by 2 contingency tables in which each table is formed by two independent binary variables. With everything else being equal, our test was developed for data in which the two binary variables are measured on matched pairs. We have applied our test to an epidemiological study to show the utility of our test. The common nonzero risk difference described in H_0 indicated no interaction between treatments and change in biomarker status in our example. Such a null hypothesis is less common than a null hypothesis with the common risk difference of zero, which means no change in marker status after adjusting for the stratum (treatment) effect.

We conducted simulations to demonstrate that our test preserves the type I error rate. Our test is most sensitive when the difference across strata is large and when the stratum fractions are balanced. The Cochran–Mantel–Haenszel (CMH) test (Mantel and Haenszel, 1959) can be used when the Breslow–Day test is not rejected for a particular data set. The CMH test combines odds ratios from all strata and calculates a common odds ratio. For future research on data with matched pairs, we plan to develop a test, analogous to the CMH test, for when our test of homogeneous stratum effect is not rejected.

APPENDIX: R CODE FOR COMPUTING TEST STATISTIC AND *p*-VALUE FOR TESTING HOMOGENEITY OF STRATUM EFFECTS IN STRATIFIED PAIRED BINARY DATA

```
    HSE = function (Tab){
    # HSE: Testing homogeneity of stratum effects in
stratified binary Data
    # Tab: 2kx2 matrix composed of k 2x2 contingency tables
across strata 1 to k
    # Output: test statistic and p-value
    k=nrow(Tab)/2
    delta=sigtilde=sighat=w=N=rep(NA,k)
    for (i in 1:k){
    Tabi=Tab[(2*i-1):(2*i),]
    N[i]=sum(Tabi)
    pi12=Tabi[1,2]/N[i]
    pi21=Tabi[2,1]/N[i]
    delta[i]=pi12-pi21
    sigtilde[i]=(pi12+pi21-(pi12-pi21)^2)/N[i]
    sighat[i]=(pi12+pi21)/N[i]
    if (sigtilde[i]!=0) w[i]=1/sigtilde[i]
    }
    if (any(sigtilde==0)) w=1
    A = cbind (diag(k-1), -1)
    deltahat=sum(w*delta)/sum(w)
    sighat=sighat-deltahat^2/N
```

```
negindex=(sighat <= 0)
sighat[negindex]=sigtilde[negindex]
Sigmahat=A%*%diag(sighat)%*%t(A)
Adelta=A%*%delta
T=t(Adelta)%*%solve(Sigmahat)%*%Adelta
p=1-pchisq(T, k-1)
list(T=T, p=p)
}
```

ACKNOWLEDGMENTS

We thank Dr. Kristina Wasson-Blader for editorial assistance. In addition, we thank two anonymous referees for their constructive comments, which helped improve the content of our article. This article represents the authors' own work and opinion. It does not reflect any policy nor represent the official position of the US Food and Drug Administration.

REFERENCES

Bowker, A. H. (1948). Bowker's test for symmetry. *Journal of the American Statistical Association* 43:572–574.

Breslow, N. E., Day, N. E. (1980). *Statistical methods in cancer research, Volume I: The analysis of case-control studies*. IARC Scientific Publications No. 32. Lyon, France: International Agency for Research on Cancer.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12:153–157.

Mantel, N., Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22:719–748.

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry* 116:651–655.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* 42:412–416.