Published in final edited form as:

Stat Med. 2009 May 30; 28(12): 1668–1679. doi:10.1002/sim.3580.

Non-inferiority tests for clustered matched pair data

Jun-mo Nam^{1,*,+} and Deukwoo Kwon²

¹Biostatistics Branch, DCEG, National Cancer Institute, National Institutes of Health, DHHS, Rockville, Maryland 20852-7244, U.S.A.

²Radiation Epidemiology Branch, DCEG, National Cancer Institute, National Institutes of Health, DHHS, Rockville, Maryland 20852-7244, U.S.A.

SUMMARY

Non-inferiority tests for matched-pair data where pairs are mutually independent may not be appropriate when pairs are clustered. The tests may require an adjustment to account for the correlation within a cluster. We consider the adjusted score and Wald-type tests, and a modification of Obuchowski's method for non-inferiority and compare them with the non-inferiority test based on a method of moments estimate in terms of Type 1 error rate and power by simulations for a small cluster size under various correlation structures. In general, the score test adjusted by an inflation factor and the modified Obuchowski's method perform as good as the test based on moments estimate in the accuracy of Type 1 error rates. The latter does not provide reasonably close Type 1 error rates to the nominal level when a number of clusters is 25 or smaller and a positive response rate for the standard procedure is 20% or lower. The adjusted score test, the method based on moments estimate and the modified test are comparable in power. The adjusted Wald-type test is too anti-conservative and we should caution use of the test. Since number of clusters is strongly related to the accuracy of empirical Type 1 error rate and power, it is very important to have a sufficiently large number of clusters in designing a clustered matched-pair study for non-inferiority.

Keywords

adjusted score test; adjusted Wald-type test; clustered matched pair; method of moments estimate; non-inferiority

1. INTRODUCTION

When the standard procedure is highly effective or accurate but it is expensive, toxic or invasive, researchers are interested in finding an inexpensive, less toxic or non-invasive new procedure which is no worse than a pre-specified amount in effectiveness or accuracy. A statistical test of significance for this research goal is called a non-inferiority test. McNemar's statistic [1] for testing a difference between two treatments is not applicable for this situation. Lu and Bean [2] and Nam [3] have provided statistical methods to establish non-inferiority for matched-pair data when pairs are independent. However, these tests may not be appropriate for clustered matched pair data where pairs in a cluster are correlated. Durkalski et al [4] proposed a statistical method testing non-inferiority for clustered data using the method of moments [5] and examined the performance of the method for sparse data by simulations. The standard error calculated under the assumption of independent pairs is smaller than that under

*Correspondence to : Jun-mo Nam, Biostatistics Branch, DCEG, National Cancer Institute, 6120 Executive Blvd, Rockville, Maryland 20852-7244, U.S.A.

[†]E-mail: namj@mail.nih.gov

non-independent pairs when pairs in a cluster are positively correlated [6–11]. Eliasziw and Donner [12] adjusted McNemar's test by utilizing an inflation factor to account for the underestimated variance resulting from clustering. We investigate the non-inferiority test statistics for clustered matched pair samples using a similar approach.

Section 2 provides notations and a review of non-inferiority tests, and presents an application of an inflation factor to adjust non-inferiority tests [2,3] for clustered data. It also includes a modification of Obuchowski's method [13] for non-inferiority. Section 3 gives evaluation of the performance of each of the non-inferiority tests in terms of the accuracy of empirical Type 1 error rates and power by simulations under the general correlation structure. Sections 4 and 5 contain a numerical example for an illustration and discussion.

2. NON-INFERIORITY TESTS

We are interested in establishing non-inferiority of a new procedure comparing with the standard procedure in clustered matched-pair studies. Consider a random sampling of K clusters from a population, where there are n_k units in the k^{th} cluster for $k=1, 2, \dots, K$. The new and standard procedures are administered to each unit. We have n_k pairs of matched samples (Y_{1ik}, Y_{0ik}) for $i=1, 2, \dots, n_k$ in the k^{th} cluster. Y_{1ik} and Y_{0ik} are binary responses (i.e. 1 or 0 as respond or not-respond) of the new and standard procedures in the i^{th} pair and the k^{th} cluster. There are four possible pair types, i.e., (1, 1), (1, 0), (0, 1) and (0, 0). The matched observations and corresponding probabilities in the k^{th} cluster are shown in Table 1. We consider a case where the number of clusters is large and a number of units in each cluster is small. Assume a common difference between two procedures in proportions for all clusters, i.e., $p_{1\bullet k} - p_{0\bullet k} = \delta$ (or $p_{10k} - p_{01k} = \delta$) for $k=1, 2, \dots, K$. We want to test the null hypothesis, $H_0: \delta = \delta_0 (< 0)$ against an alternative, $H_1: \delta > \delta_0$. Letting $\hat{t}_k = \hat{p}_{10k} - \hat{p}_{01k} - \delta_0$ where $\hat{p}_{10k} = x_{10k} / n_k$ and $\hat{p}_{01k} = x_{01k} / n_k$ for every k , we have $E(\hat{t}_k) = 0$ under H_0 . Applying a method of moments estimate (Royal [5]), Durkalski et al [4] proposed a test statistic for non-inferiority of clustered matched pair data,

$$Z_{EV} = \sum_{k=1}^K \left(\frac{x_{10k} - x_{01k}}{n_k} - \delta_0 \right) / \left\{ \sum_{k=1}^K \left(\frac{x_{10k} - x_{01k}}{n_k} - \delta_0 \right)^2 \right\}^{1/2}. \quad (1)$$

For a large number of clusters, the statistic (1) is approximately normal with mean zero and variance one under H_0 . We reject H_0 against H_1 at α when $Z_{EV} > z_{(1-\alpha)}$ where $z_{(1-\alpha)}$ is the $100 \times (1-\alpha)$ percentile point of the standard normal distribution. Let the dot denote a sum over k ,

e.g., $n_{\bullet} = \sum_{k=1}^K n_k$ which is a total number of units across clusters. If all matched pairs are independent, we may pool $K \times 2 \times 2$ tables (Table 1) into a large single 2×2 table and apply the test statistics Z_{LB} (Lu and Bean [2]) or Z_N (Nam [3]) for non-inferiority:

$$Z_{LB} = (x_{10\bullet} - x_{01\bullet} - n_{\bullet}\delta_0) / (x_{10\bullet} + x_{01\bullet} - n_{\bullet}\delta_0)^{1/2} \quad (2)$$

or

$$Z_N = (x_{10\bullet} - x_{01\bullet} - n_{\bullet}\delta_0) / \{n_{\bullet}(\tilde{p}_{10} + \tilde{p}_{01} - \delta_0^2)\}^{1/2} \quad (3)$$

where $\tilde{p}_{10} = \tilde{p}_{01} = +\delta$ and $\tilde{p}_{01}$ is the restricted maximum likelihood estimator of p_{01} for given $\delta = \delta_0$, i.e., $\tilde{p}_{01} = \{-b + (b^2 - 4ac)^{1/2}\} / (2a)$ where $a = 2n_{\bullet}$, $b = (2n_{\bullet} + x_{01\bullet} - x_{10\bullet})\delta_0 - (x_{10\bullet} +$

$x_{01\bullet}$) and $c = -\delta_0(1 - \delta_0)x_{01\bullet}$. The statistics (2) and (3) for testing non-inferiority have been used for data where all units are randomly sampled from the same population. If matched pairs are positively correlated in a cluster, the test statistics based on pooled data ignoring the positive correlation are inflated and p-values are distorted downward. It may falsely reject non-inferiority. A correlation coefficient between units in a cluster has been called the intraclass correlation coefficient. A simulation study by Durkalski et al [4] showed an actual Type 1 error rate of an unadjusted test statistic is larger than a nominal level and it is greater as an intraclass correlation coefficient increases. Our simulation indicates the same observation. The intraclass correlation coefficient may be estimated using the mean square error of analysis of variance for mixed model (Snedecor and Cochran [14] and Donner [6]). A variance of sample mean under the assumption of independent pairs can be adjusted by multiplying the variance inflation factor (or design effect) for clustered data, e.g., Kish [15], Cochran [16] and Cornfield [17]. This type of adjustment has been applied in analysis of data from random group sampling, community intervention trials, longitudinal studies and sample survey. Applying a similar procedure to adjust McNemar test statistic for non-independent matched pair data (Eliasziw and Donner [12] and Durkalski et al [18]), the test statistic for non-inferiority based on pooled data may be adjusted using an inflation factor, $\bar{c} = 1 + (n_c - 1) \hat{\rho}$ where n_c is the adjusted mean number of discordant pairs and $\hat{\rho}$ is a consistent estimator of the intraclass correlation coefficient ρ (Donner [6]), (See APPENDIX for n_c and $\hat{\rho}$). The adjusted Z_{LB} and Z_N statistics testing $H_0: p_{10} - p_{01} = \delta_0 (< 0)$ against $H_1: p_{10} - p_{01} > \delta_0$ for clustered matched pair data are

$$Z_{ALB} = Z_{LB} / \{1 + (n_c - 1) \hat{\rho}\}^{1/2} \quad (4)$$

and

$$Z_{AN} = Z_N / \{1 + (n_c - 1) \hat{\rho}\}^{1/2}. \quad (5)$$

When $n_c = 1$ or $\hat{\rho} = 0$, an inflation factor is unity and no adjustment is needed. For a large number of clusters, the adjusted statistics are distributed asymptotically normal mean zero and variance one. As an alternative method to an adjusted McNemar's statistic for testing the conventional null hypothesis (Eliasziw and Donner [12]), Obuchowski [13] presented a statistic using Rao and Scott's formulation [19]. Denote that $\hat{p}_{1\bullet\bullet} = x_{1\bullet\bullet}/n_{\bullet\bullet}$, $\hat{p}_{\bullet 1\bullet} = x_{\bullet 1\bullet}/n_{\bullet\bullet}$ and $\bar{p}_{1\bullet\bullet} = \{(x_{1\bullet\bullet} + x_{\bullet 1\bullet}) / n_{\bullet\bullet} + \delta_0\} / 2$, $\bar{p}_{\bullet 1\bullet} = \{(x_{1\bullet\bullet} + x_{\bullet 1\bullet}) / n_{\bullet\bullet} - \delta_0\} / 2$, $p' = (p_{1\bullet\bullet}, p_{\bullet 1\bullet})$ and $\bar{p}' = (\bar{p}_{1\bullet\bullet}, \bar{p}_{\bullet 1\bullet})$. When a similar approach applied to a non-inferiority setting, we have

$$Z_o = (\hat{p}_{1\bullet\bullet} - \hat{p}_{\bullet 1\bullet} - \delta_0) / \{\widehat{\text{var}}(\hat{p}_{1\bullet\bullet} - \hat{p}_{\bullet 1\bullet} - \delta_0)_{p=\bar{p}}\}^{1/2} \quad (6)$$

where $\widehat{\text{var}}(\hat{p}_{1\bullet\bullet} - \hat{p}_{\bullet 1\bullet} - \delta_0)_{p=\bar{p}} = \widehat{\text{var}}(\hat{p}_{1\bullet\bullet}) + \widehat{\text{var}}(\hat{p}_{\bullet 1\bullet}) - 2 \widehat{\text{cov}}(\hat{p}_{1\bullet\bullet}, \hat{p}_{\bullet 1\bullet})_{p=\bar{p}}$

$$\begin{aligned} \widehat{\text{var}}(\hat{p}_{1\bullet\bullet})_{p=\bar{p}} &= K(K-1)^{-1} \left\{ \sum_{k=1}^K (x_{1\bullet k} - n_k \bar{p}_{1\bullet\bullet})^2 / n_{\bullet\bullet}^2 \right\}, \\ \widehat{\text{var}}(\hat{p}_{\bullet 1\bullet})_{p=\bar{p}} &= K(K-1)^{-1} \left\{ \sum_{k=1}^K (x_{\bullet 1 k} - n_k \bar{p}_{\bullet 1\bullet})^2 / n_{\bullet\bullet}^2 \right\} \end{aligned}$$

and $\widehat{\text{cov}}(\widehat{p}_{1\bullet\bullet}, \widehat{p}_{\bullet 1\bullet})_{p=\bar{p}} = K(K-1)^{-1} \left\{ \sum_{k=1}^K (x_{1\bullet k} - n_k \bar{p}_{1\bullet\bullet})(x_{\bullet 1k} - n_k \bar{p}_{\bullet 1\bullet}) / n_{\bullet}^2 \right\}$ (See Table 1 for notations). For a large number of clusters, the statistic (6) is approximately a standard normal variate. We reject the null hypothesis at α when $Z_O \geq Z(1-\alpha)$.

Denote $v = (p_{10} + p_{01} - \delta^2)$, $\hat{t} = (\hat{p}_{10\bullet} + \hat{p}_{01\bullet} - \delta)$, and $n_{\bullet} = K\bar{n}$ where $\bar{n} = \sum_{k=1}^K n_k / K$. As $\bar{n} \rightarrow \infty$, $n_c / \bar{n} \rightarrow f = p_{10} + p_{01}$ (APPENDIX). The variance of $\hat{t}$ adjusted with an inflation factor is expressed as $\text{var}(\hat{t}) = v \{ 1 + (n_c - 1)\rho \} / (K\bar{n}) = (v/K) \{ (1-\rho) / \bar{n} + (n_c / \bar{n}) \rho \}$. The above variance approaches to zero as $K \rightarrow \infty$ while it converges to $v\rho / K$ as $\bar{n} \rightarrow \infty$. It suggests that a large number of clusters provides more strong impact than a larger cluster size does in precision of estimation of t and power of a test based on $\hat{t}$. Note that we consider non-inferiority tests on clustered matched pair data with a large value of K and a small cluster size in this paper.

3. SIMULATIONS

We conducted a simulation study to assess the performance of statistics, Z_{EV} , Z_{ALB} , Z_{AN} and Z_O for testing non-inferiority of a new procedure compared with the standard one in empirical Type 1 error rate and power for clustered matched pair data. Pre-specified parameters are the number of cluster (K), the number of units in the k^{th} cluster (n_k), the probabilities of a positive response for the new and standard procedures (p_1 and p_0 , respectively) and a negative value which is materially unacceptable difference between the new and standard procedures ($p_{10} - p_{01} = \delta_0 = -0.1$). We applied a specified correlation structures similar to those by Obuchowski [13] and Durkalski et al [4]. The structure specified four within cluster correlations are: between the responses for the new procedure (r_1); between responses for the standard procedure (r_2); between responses for new and standard procedures on the same units (r_3); and between those responses of different units (r_4). We set $r_1 = r_2 = r$ and $r_4 = r/4$ since r_4 is likely much smaller relative to r , and consider various values of r and r_3 . Using Matlab subroutine, we generate a random vector from a $(n_k \times 2)$ -variate normal distribution with mean $\mathbf{0}$ and variance-covariance Σ where correlation coefficients (r , r_3 and r_4) are specified. The 1st n_k variates are the outcome of n_k units with the k^{th} cluster given the new procedure and the last n_k variates are those given the standard procedure. Denote clustered normal variates as z . Then, we generate a binomial response for each procedure. Define y as $y=1$ if $z \leq c$ or $y=0$ otherwise where c is a cut-off point which satisfies $\Pr(z \leq c) = p$. When the number of units per cluster (n_k) varies, we consider two kinds of distributions: the value of n_k is generated from a uniform distribution and also from a beta distribution. In our simulation study, 10,000 data sets are generated for each configuration. If the expected Type 1 error rate of a test is indeed a nominal 0.05 level, then a 95% confidence interval for $\alpha=0.05$ is (0.046, 0.054) from 10,000 simulated data. Those empirical Type 1 error rates outside of the interval are shown with an asterisk sign (*). When there are no discordant pairs in a cluster in simulations, an intraclass correlation coefficient is not estimable and such a case is excluded in Type 1 error rate and power calculations. If undefined cases occur, then they are excluded from both numerator and denominator in Type 1 error rate and power computations and empirical coverage rate and power are adjusted accordingly.

Table II-A, B and C summarize simulated Type 1 error rates and power of tests for $K=(100, 50, 25)$, $r_1=r_2=(0, 0.1, 0.4, 0.6)$, $r_3=0.5$, $\delta_0 = -0.1$ and $\delta_1=0$ with $n_k=2$ units per cluster and those with uniformly as well as non-uniformly distributed $n_k \leq 5$ per cluster, respectively. Table II-A for $n_k=2$ shows that Z_{AN} and Z_O tests provide empirical Type 1 error rates which are satisfactorily close to a nominal 0.05 level except those cases $K \leq 50$ and $p_0=0.2$. In these cases, Z_{AN} test is conservative while Z_O test is anti-conservative. The Z_{EV} test gives reasonably close Type 1 error rates but it does not give reliable Type 1 error rates for $K \leq 50$ with $p_0=0.2$ or for

$K=25$ with $p_0=0.8$. The performance of the Z_{ALB} test is always anti-conservative and less satisfactory in a comparison with the Z_{EV} , Z_{AN} and Z_O tests. When a cluster size is uniformly distributed, Table II-B for $n_k \leq 5$ indicates that Z_{AN} and Z_O tests provide satisfactorily accurate Type 1 error rates except $K \leq 50$ with $p_0=0.2$. A size of the Z_{EV} test is reasonably close to a nominal 0.05 level except $K \leq 25$, or $K \leq 50$ with $p_0=0.2$ and that of the Z_{ALB} test is generally not satisfactory. When a cluster size is distributed non-uniformly, Table II-C showed that the Z_O test gives a reasonably accurate Type 1 error rate unless $K \leq 50$ with $p_0=0.2$, and the Z_{AN} test is satisfactory except those cases of $K=25$ or $K=50$ with $p_0=0.8$. The Z_{EV} test is satisfactory unless $K=25$ and $p_0=0.2$. In summary, the Z_{EV} , Z_{AN} and Z_O tests are generally robust and comparable unless a number of clusters is small and they perform clearly better than the Z_{ALB} test in the accuracy of Type 1 error rates to the nominal one.

Empirical power of each of the four tests is decreasing as an intraclass correlation coefficient increases, and it is greater as a number of clusters increases. Power is related to a response rate of the standard procedure p_0 : it is greater as a departure of p_0 from 0.5 increases. Power is also inversely related to a size of the test. In assessment of power along with empirical Type 1 error rates, we may consider Z_{EV} , Z_{AN} and Z_O tests are competitive in power for $n_k=2$ and $n_k \leq 5$. Although power of Z_{ALB} method is slightly greater than those other tests, it does not imply that the Z_{ALB} test is more powerful since a Type 1 error rate of the Z_{ALB} test is always more anti-conservative and power is inversely related with Type 1 error rate. Simulations for those with various values of r_3 , and $r_4 = r_1/8$ yield similar results in statistical properties.

We examine a relative importance of a number of clusters and cluster size for a fixed total number of units. Table III shows average empirical Type 1 error rates and power of a test over various values of p_0 and r_1 . It demonstrates that, for a fixed total number of units (e.g., $N=200$), reliability of type 1 error and power of a non-inferiority test based on clustered matched pair data are more closely related with a number of clusters than a cluster size. Simulations show that a large number of clusters is pivotal to obtain accurate Type 1 error rate and good power of each of the four tests particularly when a cluster size is small.

4. AN EXAMPLE

Consider a diagnostic study for comparing two imaging techniques, i.e., positron emission tomography (PET) and single photon emission CT scan (SPECT), Neumann et al [20]. In 21 patients, there were 51 glands confirmed at surgery not to have hyperparathyroidism. The number of glands for each patient ranges from one to four with the average 2.4 glands. Consider a patient and glands in a patient as a cluster and units in a cluster, respectively. (See Table IV, Obuchowski [13]). PET provides good resolution images in detecting abnormal parathyroid but is expensive while SPECT is inexpensive and quickly performed but gives lower resolution. Weighting cost and time as well as accuracy, we may consider that SPECT is non-inferior to PET when the former must be no more than 10% lower than the latter in specificity, i.e., testing null against alternative hypotheses are $H_0 : \delta = -0.1$ vs. $H_1 : \delta > -0.1$ where $p_{10k} - p_{01k} = \delta$ for $k=1, 2, \dots, 21$. When a possible correlation within cluster is ignored, we have $Z_{LB}=4.06$ and $Z_N=3.30$ based on pooled data from (2) and (3). When they are adjusted by using an inflation factor based on discordant pairs, we have $Z_{ALB}=3.31$ ($p=0.0005$) and $Z_{AN}=2.68$ ($p=0.004$) from (4) and (5). Values of the adjusted statistics are smaller than those unadjusted ones. From (1) and (6), the statistic proposed by Durkalski et al [4] and the modified Obuchowski one are $Z_{EV}=2.61$ ($p=0.005$) and $Z_O=2.62$ ($p=0.005$), and two values are essentially identical. They are similar to the value of Z_{AN} . These three statistics are smaller than Z_{ALB} which is known to be anti-conservative in Section 3.

5. DISCUSSION

Lu and Bean [2] and Nam [3] have developed test statistics for non-inferiority when matched pairs are independent. The former is based on a Wald-type method and the latter on the likelihood score method. These test statistics are not applicable for clustered matched pair data. However, they can be adjusted properly using an inflation factor resulting from correlated pairs in a cluster. The adjusted statistics are appropriate for testing non-inferiority in clustered pair data. Durkalski et al [4] have presented a proper test statistic based on moment estimate. In addition, we also consider a modification of Obuchowski's method [13]. We investigate four statistical methods applicable to analysis of clustered matched pair data for non-inferiority. For a large number of clusters, empirical Type 1 error rates of the test based on moments estimate, the adjusted score test by an inflation factor and a modified Obuchowski's test for non-inferiority are satisfactorily close to a nominal level. When a number of clusters is small and a positive response rate for the standard procedure is low, the test based on moments estimate and the modified test are anti-conservative. The adjusted Wald-type test for non-inferiority is always anti-conservative. We may not recommend the Wald-type method for testing non-inferiority in clustered matched-pair studies. The adjusted score, moment based tests and the modified method are reasonably robust. The latter two tests are computationally simple and easy to apply while the former requires more computations but it yield information regarding an intra-class correlation coefficient in clustered matched-pairs. If tests ignoring the positive correlation within cluster do not reject non-inferiority, then tests adjusted by an inflation factor also do not reject the null hypothesis when the estimated correlation is positive. If an unadjusted test rejects the null, then the adjusted test may or may not reject the null and it is necessary to conduct the adjusted test for a statistical significance. The intra-class correlation coefficient involving an inflation factor for an adjustment may be also estimated using information from all pair types factor instead of those of discordant pairs only [12,18]. However, our empirical results indicate that this type of adjustment does not improve the accuracy of Type 1 error rates. When an intraclass coefficient is positive but close to zero, an estimated intraclass correlation coefficient can be negative due to sampling variations. We did not exclude such a case for empirical Type 1 error rate and power calculations. The exclusion may be resulted to bias of empirical error rate and power. For no intraclass correlation, a test without adjustment is appropriate. We employed the analysis of variance estimator of intraclass correlation coefficient to adjust Z_{LB} and Z_N test statistics in this paper. Although this is a commonly used estimator, there are also alternative estimators of the intraclass correlation coefficient. Donner [21] and Ridout et al [22] provide extensive reviews on estimators of intraclass correlation coefficient. The variance inflation factor is also called the design effect in sampling survey. Respondents in the same cluster are likely to be somewhat similar to one another. Selecting an additional member from the same cluster provides less information than would be a completely independent selection. The loss of effectiveness by cluster sampling, instead of simple random sampling is the design effect. The inflation factor (design effect) is the ratio of the variance of a statistic under use of cluster sampling to that under use of simple random sampling. The inflation factor is greater as an intraclass correlation increases. In cluster sampling, the clusters are treated as the sampling elements so analysis is done on a population of clusters. In stratified sampling, a random sample is drawn from each of strata and analysis is done on elements within strata. In this paper, we investigate non-inferiority tests for clustered matched pair data. Nam [24] studied noninferiority tests for stratified pair design. Liu et al [25] examined equivalence and non-inferiority when pairs are independent. For a small sample size, a test with a continuity correction leads to the test conservative. When a number of clusters is large, a continuity correction is not necessary. For a fixed total number of units, the accuracy of Type 1 error rate and power of a test is more closely related to a number of clusters than cluster size. In designing a clustered matched-pair study for non-inferiority, one should make a great effort to have a sufficiently large number of clusters. The asymptotic normality of a test statistic is based on

the central limit theory and a large number of clusters is essential for a reliable p -value and a good power of the test.

Appendix

APPENDIX

Define $S_k = x_{10k} + x_{01k}$ for $k=1, 2, \dots, K$ and K_d as the number of clusters at least one discordant

pairs: $K_d = \sum_{k=1}^K I_k$ where $I_k=0$ when $S_k=0$ and $I_k=1$ when $S_k \geq 1$. The unadjusted and adjusted mean

numbers of discordant pairs are $\bar{S} = \sum_{k=1}^K S_k / K_d$ and

$S_0 = \bar{S} - \left\{ \sum_{k=1}^K (S_k - \bar{S})^2 - (K - K_d) \bar{S}^2 \right\} / \{K_d(K_d - 1) \bar{S}\}$, e.g., [12,23]. The mean squares

between clusters and within clusters are expressed as $BMS = \left\{ \sum_{k=1}^K (x_{10k} - S_k \bar{p})^2 / S_k \right\} / K_d$ where

$\bar{p} = \sum_{k=1}^K x_{10k} / \sum_{k=1}^K S_k$ and $WMS = \sum_{k=1}^K (x_{10k} x_{01k} / S_k) / \{K_d(\bar{S} - 1)\}$ where $S_k \geq 1$. An estimator of the intra-class correlation coefficient is $\hat{\rho} = (BMS - WMS) / \{BMS + (S_0 - 1)WMS\}$, e.g., [12,18, 23]. If $S_k=0$, then the k^{th} component of WMS is undefined and excluded from analysis. When all S_k 's are zero across clusters, the $\hat{\rho}$ is not estimable.

An inflation factor resulting from intra-class correlation is written as $\bar{c} = 1 + (n_c - 1) \hat{\rho}$ where the adjusted mean number of discordant pairs is

$$n_c = S_0 + K_d(S_0 - \bar{S}) = \bar{S} + \hat{\sigma}^2 / \bar{S} \text{ where } \hat{\sigma}^2 = \left\{ \sum_{k=1}^K (S_k - \bar{S})^2 - (K - K_d) \bar{S}^2 \right\} / K_d. \quad (\text{A1})$$

The 2nd term of n_c is the variance-to-mean ratio which is a measure of a relative dispersion of

S_k 's. Denote the mean cluster size as $\bar{n} = \sum_{k=1}^K n_k / K$. The ratio of an adjusted mean number of discordant pairs to the mean cluster size is $n_c / \bar{n} = \bar{S} / \bar{n} + \hat{\sigma}^2 / (\bar{n} \bar{S})$ from (A1). Assume $p_{10k} = p_{10}$ and $p_{01k} = p_{01}$ for every k . We have $E(I_k) = \Pr(S_k \geq 1) = 1 - \Pr(S_k = 0) = 1 - (1 - p_{10} - p_{01})^{n_k}$. When a cluster size is the same ($n_k = n$ for every k , i.e., $\bar{n} = n$), we have

$E(K_d) = \sum_{k=1}^K E(I_k) \rightarrow K$ as $n \rightarrow \infty$. Thus, $n_c / \bar{n} \rightarrow f = p_{10} + p_{01}$ as $n \rightarrow \infty$. Similarly, we have $n_c / \bar{n} \rightarrow f$ as $\bar{n} \rightarrow \infty$ when a cluster size varies.

ACKNOWLEDGEMENT

The authors thank two reviewers and an associate editor for their helpful suggestions and constructive comments that improve the manuscript. This research was supported by the Intramural Research Program of the NIH National Cancer Institute.

REFERENCES

1. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947;12(2):153–157.

2. Lu Y, Bean JA. On the sample size for one-sided equivalence of sensitivities based upon McNemar's Test. *Statistics in Medicine* 1995;14:1831–1839. [PubMed: 7481214]
3. Nam J. Establishing equivalence of two treatments and sample size requirements in matched-pair data. *Biometrics* 1997;53:1422–1230. [PubMed: 9423257]
4. Durkalski VL, Palesch YY, Lipsitz SR, Rust PF. Analysis of clustered matched-pair data for a non-inferiority study design. *Statistics in Medicine* 2003;22:279–290. [PubMed: 12520562]
5. Royall RM. The prediction approach to robust variance estimation in two-stage cluster sampling. *Journal of American Statistical Association* 1986;81:119–123.
6. Donner A. The analysis of intraclass correlation in multiple samples. *Annals of Human Genetics* 1985;49:75–82. [PubMed: 4073828]
7. Donald A, Donner A. Adjustments to the Mantel-Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. *Statistics in Medicine* 1987;6:491–499. [PubMed: 3629050]
8. Donner A, Banting D. Analysis of site-specific data in dental studies. *Journal of Dental Research* 1988;67(11):1392–1395. [PubMed: 3183156]
9. Donner A, Donald A. The statistical analysis of multiple binary measurements. *Journal of Clinical Epidemiology* 1988;41(9):899–905. [PubMed: 3183697]
10. Donner A, Banting D. Adjustment of frequently used chi-square procedures for the effect of site-to-site dependencies in the analysis of dental data. *Journal of Dental Research* 1989;68(9):1350–1354. [PubMed: 2778178]
11. Donner A. Statistical methods in ophthalmology: an adjusted chi-square approach. *Biometrics* 1989;45:605–611. [PubMed: 2765640]
12. Eliasziw M, Donner A. Application of the McNemar test to non-independent matched-pair data. *Statistics in Medicine* 1991;10:1981–1991. [PubMed: 1805322]
13. Obuchowski NA. On the comparison of correlated proportions for clustered data. *Statistics in Medicine* 1998;17:1495–1507. [PubMed: 9695194]
14. Snedecor, GW.; Cochran, WG. *Statistical Methods*. Vol. 7th ed.. Ames: Iowa State University Press; 1967.
15. Kish, L. *Survey Sampling*. New York: Wiley; 1965.
16. Cochran, WG. *Sampling Techniques*. New York: Wiley; 1977.
17. Cornfield J. Randomization by groups: A formal analysis. *American Journal of Epidemiology* 1978;108:100–102. [PubMed: 707470]
18. Durkalski VL, Palesch YY, Lipsitz SR, Rust PF. Analysis of clustered matched-pair data. *Statistics in Medicine* 2003;22:2417–2428. [PubMed: 12872299]
19. Rao JNK, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics* 1992;48:577–585. [PubMed: 1637980]
20. Neumann DR, Esselstyn CB, MacIntyre WJ, Go RT, Obuchowski NA, Chen EQ, Licata AA. Comparison of FDG PET and sestamibi-SPECT in primary hyperparathyroidism. *Journal of Nuclear Medicine* 1996;37:1809–1815. [PubMed: 8917180]
21. Donner A. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review* 1986;54:67–82.
22. Ridout MS, Demitrio CGB, Firth D. Estimating intraclass correlation for binary data. *Biometrics* 1999;55:137–148. [PubMed: 11318148]
23. Fleiss, JL. *Statistical Methods for rates and Proportions*. Vol. 2nd ed.. New York: Wiley; 1981.
24. Nam J. Non-inferiority of new procedure to standard procedure in stratified matched-pair design. *Biometrical Journal* 2006;6:966–977. [PubMed: 17243295]
25. Liu J, Hsueh H, Hsieh H, Chen J. Tests for equivalence or non-inferiority for paired binary data. *Statistics in Medicine* 2002;21:231–245. [PubMed: 11782062]

Table I
Observations and probabilities for matched pairs in the k^{th} cluster

New procedure	<u>Standard procedure</u>		sum
	1	0	
1	$x_{11k}(p_{11k})$	$x_{10k}(p_{10k})$	$x_{1\bullet k}(p_{1\bullet k})$
0	$x_{01k}(p_{01k})$	$x_{00k}(p_{00k})$	$x_{0\bullet k}(p_{0\bullet k})$
sum	$x_{\bullet 1k}(p_{\bullet 1k})$	$x_{\bullet 0k}(p_{\bullet 0k})$	n_k

Table II

Table II-A. Empirical Type I Error rates and power: $n_k=2, r_3=0.5, r_4=r_1/4, \delta_0=-0.1$ ($p_1 - p_0 = -0.1$) and $\delta_1 = p_1 - p_0 = 0$. Asterisk indicates a value is outside the 95% confidence interval on the nominal level of 0.05.

K	p_0	$r_1=r_2$	Type I error rate (%)				Power (%)			
			Z_{EV}	Z_{ALB}	Z_{AN}	Z_0	Z_{EV}	Z_{ALB}	Z_{AN}	Z_0
100	0.8	0	4.8	5.3	4.7	4.6	90.8	91.6	90.6	90.5
		0.1	4.8	5.2	4.7	4.6	89.1	90.1	88.8	88.8
		0.4	4.9	5.3	4.8	4.8	84.2	85.2	89.6	83.7
		0.6	4.8	5.3	4.6	4.6	80.1	81.5	79.1	79.5
50	0.8	0	5.4	6.0*	4.9	4.9	68.2	70.6	66.7	66.1
		0.1	5.4	6.0*	5.1	4.9	65.7	68.6	64.1	64.0
		0.4	5.4	6.1*	5.2	5.2	59.3	62.3	57.7	58.1
		0.6	5.3	6.4*	5.1	5.2	54.6	57.9	53.3	54.1
25	0.8	0	5.9*	6.0*	5.3	5.1	44.5	45.3	42.1	41.1
		0.1	5.9*	5.8*	5.1	4.9	42.9	43.4	40.0	39.4
		0.4	5.9*	6.2*	5.0	5.2	39.1	39.2	35.4	36.3
		0.6	5.8*	6.4*	5.0	5.1	36.3	37.4	32.9	33.6
100	0.5	0	5.2	5.7*	5.3	5.1	78.5	80.0	78.7	78.0
		0.1	5.0	5.5*	5.1	4.9	76.0	77.2	46.1	75.6
		0.4	5.2	5.5*	5.2	5.0	69.3	70.6	69.3	68.5
		0.6	5.1	5.5*	5.1	4.9	64.6	65.	64.8	64.0
50	0.5	0	4.6	5.4	4.9	4.5*	53.0	55.9	54.2	52.7
		0.1	4.9	5.7*	5.1	4.8	50.8	53.8	51.8	50.6
		0.4	4.8	5.7*	4.9	4.7	45.1	48.3	45.7	44.5
		0.6	5.1	5.9*	5.2	5.0	41.8	44.6	42.0	41.1
25	0.5	0	5.5*	6.2*	5.4	4.8	33.1	35.7	33.0	30.6

Table II-A. Empirical Type 1 Error rates and power: $n_k=2, r_3=0.5, r_4=r_1/4, \delta_0=-0.1$ ($p_1 - p_0 = -0.1$) and $\delta_1 = p_1 - p_0 = 0$. Asterisk indicates a value is outside the 95% confidence interval on the nominal level of 0.05.

K	p_0	$r_1=r_2$	Type 1 error rate (%)				Power (%)			
			Z_{EV}	Z_{ALB}	Z_{AN}	Z_0	Z_{EV}	Z_{ALB}	Z_{AN}	Z_0
100	0.2	0.1	5.7*	6.4*	5.6*	5.0	31.6	34.2	31.5	29.1
		0.4	5.3	6.2*	5.2	4.7	28.4	31.4	28.5	26.4
		0.6	5.2	6.2*	5.3	4.7	26.2	28.9	26.7	24.9
100	0.2	0	5.1	5.4	4.5*	5.1	90.5	91.3	90.3	90.3
		0.1	5.3	5.7*	4.8	5.2	89.1	89.9	88.7	88.8
		0.4	5.6*	6.0*	4.9	5.5*	84.5	85.6	84.0	84.1
		0.6	5.7*	6.2*	4.9	5.6*	80.4	81.8	79.6	79.9
50	0.2	0	5.8*	5.9*	4.3*	5.3	67.6	69.9	66.2	65.8
		0.1	5.8*	5.8*	3.9*	5.1	65.5	68.0	63.8	63.7
		0.4	6.2*	6.4*	4.5*	5.7*	59.8	62.3	58.3	58.9
		0.6	5.9*	6.6*	4.5*	5.5*	54.8	58.2	53.4	54.2
25	0.2	0	6.7*	5.1	4.3*	5.8*	45.2	46.1	43.0	42.0
		0.1	7.2*	5.5*	4.6	6.2*	44.3	44.5	41.2	40.8
		0.4	7.3*	5.5*	4.3*	6.6*	39.6	40.1	36.1	36.9
		0.6	7.2*	6.2*	4.2*	6.4*	37.3	37.9	33.9	34.8

Table II-B. Empirical Type 1 error rates and power: $n_k \leq 5$ where n_k is uniformly distributed, $r_3=0.5, r_4=r_1/4, \delta_0 = -0.1, (p_1 - p_0 = -0.1)$ and $\delta_1 = p_1 - p_0 = 0$. Asterisk indicates a value is outside the 95% confidence interval on the nominal level of 0.05.

K	p_0	$r_1=r_2$	Type 1 error rate (%)				Power (%)			
			Z_{EV}	Z_{ALB}	Z_{AN}	Z_0	Z_{EV}	Z_{ALB}	Z_{AN}	Z_0
100	0.8	0	5.0	5.5	4.9	4.8	91.1	97.6	97.2	97.2

Table II-B. Empirical Type I error rates and power: $n_k \leq 5$ where n_k is uniformly distributed, $r_3=0.5$, $r_4=r_1/4$, $\delta_0 = -0.1$, $(p_1 - p_0 = -0.1)$ and $\delta_1 = p_1 - p_0 = 0$. Asterisk indicates a value is outside the 95% confidence interval on the nominal level of 0.05.

K	p_0	$r_1=r_2$	Type I error rate (%)				Power (%)			
			Z_{RV}	Z_{ALB}	Z_{AN}	Z_O	Z_{RV}	Z_{ALB}	Z_{AN}	Z_O
50	0.8	0.1	5.0	5.4	4.9	4.8	89.8	95.9	95.3	95.2
		0.4	4.9	5.3	4.7	4.7	84.3	87.6	86.3	86.3
		0.6	4.6	5.1	4.4*	4.4*	80.3	80.6	78.5	78.8
50	0.8	0	5.2	5.9*	5.0	4.8	69.2	83.0	80.9	79.9
		0.1	5.3	6.1*	5.0	4.8	66.9	78.4	75.9	75.2
		0.4	5.0	6.0*	4.9	4.9	59.9	66.5	63.0	62.2
		0.6	5.2	6.1*	5.0	5.1	55.2	58.6	54.5	54.6
25	0.8	0	6.1*	6.2*	5.5*	5.1	47.3	60.7	56.9	53.5
		0.1	5.9*	5.9*	5.0	5.0	45.7	56.4	52.4	49.5
		0.4	5.7*	5.9*	4.9	5.0	40.1	46.3	41.2	39.7
		0.6	5.9*	6.4*	5.1	5.0	36.0	40.9	35.0	34.4
100	0.5	0	5.4	5.8*	5.5*	5.3	78.8	90.3	89.9	89.4
		0.1	5.2	5.6*	5.2	5.0	76.4	85.9	85.4	84.6
		0.4	5.0	5.5*	5.0	4.8	68.9	72.4	71.2	70.0
		0.6	5.1	5.5*	5.1	4.9	63.9	64.8	63.4	62.5
50	0.5	0	5.2	6.0*	5.5*	5.1	54.1	69.0	67.5	65.3
		0.1	5.2	6.0*	5.4	5.1	51.2	63.2	61.5	59.1
		0.4	5.5*	6.4*	5.6*	5.4	44.7	49.4	47.4	44.8
		0.6	5.0	5.7*	5.1	4.8	40.6	42.0	40.0	38.2
25	0.5	0	5.5*	6.1*	5.4	4.9	35.8	48.2	54.9	41.1
		0.1	5.5*	6.2*	5.4	5.0	33.5	43.3	40.9	36.5
		0.4	5.6*	6.3*	5.5*	4.8	28.5	34.1	31.7	27.9

Table II-B. Empirical Type 1 error rates and power: $n_k \leq 5$ where n_k is uniformly distributed, $r_3=0.5, r_4=r_1/4, \delta_0 = -0.1, (p_1 - p_0 = -0.1)$ and $\delta_1 = p_1 - p_0 = 0$. Asterisk indicates a value is outside the 95% confidence interval on the nominal level of 0.05.

K	p_0	$r_1=r_2$	Type 1 error rate (%)				Power (%)			
			Z_{EV}	Z_{ALB}	Z_{AN}	Z_0	Z_{EV}	Z_{ALB}	Z_{AN}	Z_0
100	0.2	0.6	5.2	6.2*	5.2	4.6	26.3	29.8	27.2	24.5
		0	5.1	5.6*	4.6	5.1	90.9	97.4	97.1	97.0
		0.1	5.1	5.6*	4.7	5.1	89.4	95.8	95.1	95.1
		0.4	5.2	5.5*	4.4*	5.1	84.0	87.6	86.3	86.3
50	0.2	0.6	5.4	5.9*	4.7	5.3	79.7	80.7	78.5	78.9
		0	5.9*	6.0*	4.3*	5.3	68.8	82.2	80.4	79.4
		0.1	5.8*	5.8*	4.1*	5.2	66.9	77.8	75.6	74.7
		0.4	6.3*	6.6*	4.5*	5.6*	60.5	66.3	63.0	62.3
25	0.2	0.6	6.7*	7.4*	5.1	6.1*	55.7	58.3	54.2	54.4
		0	6.8*	5.8*	4.7	5.8*	47.2	60.7	56.8	53.2
		0.1	6.9*	5.4	4.4*	5.8*	45.5	56.7	52.4	49.6
		0.4	7.5*	5.5*	4.0*	6.5*	39.9	47.1	41.6	39.9
		0.6	7.6*	6.9*	4.5*	6.8*	36.1	40.8	34.8	34.3

Table II-C. Empirical Type 1 error rates and power: $n_k \leq 5$ where n_k is non-uniformly distributed, $r_3=0.5, r_4=r_1/4, \delta_0 = -0.1, (p_1 - p_0 = -0.1)$ and $\delta_1 = p_1 - p_0 = 0$. Asterisk indicates a value is outside the 95% confidence interval on the nominal level of 0.05.

K	p_0	$r_1=r_2$	Type 1 error rate (%)				Power (%)			
			Z_{EV}	Z_{ALB}	Z_{AN}	Z_0	Z_{EV}	Z_{ALB}	Z_{AN}	Z_0
100	0.8	0	5.2	5.6*	5.2	5.1	97.1	98.2	98.0	98.0
		0.1	4.9	5.5*	5.1	4.9	95.9	97.0	96.7	96.7
		0.4	5.0	5.6*	5.0	4.9	89.9	91.2	90.1	90.3

Table II-C. Empirical Type 1 error rates and power: $n_k \leq 5$ where n_k is non-uniformly distributed, $r_3=0.5$, $r_4=\Gamma_1/4$, $\delta_0 = -0.1$, $(p_1 - p_0 = -0.1)$ and $\delta_1 = p_1 - p_0 = 0$. Asterisk indicates a value is outside the 95% confidence interval on the nominal level of 0.05.

K	p_0	$r_1=r_2$	Type 1 error rate (%)			Power (%)			
			Z_{EV}	Z_{ALB}	Z_{AN}	Z_0	Z_{EV}	Z_{ALB}	Z_{AN}
50	0.8	0.6	5.0	5.4	4.8	4.7	84.9	85.5	89.8
		0	5.1	6.2*	5.6*	5.2	79.1	83.4	81.5
		0.1	5.4	6.2*	5.5*	5.2	76.0	79.8	77.7
		0.4	5.3	6.2*	5.4	5.2	66.7	69.4	66.2
25	0.8	0.6	5.6*	6.6*	5.5*	5.3	59.2	61.9	58.4
		0	5.2	7.0*	5.9*	5.2	55.1	61.4	57.0
		0.1	5.0	6.7*	5.6*	4.9	51.6	57.6	53.1
		0.4	5.5*	7.1*	5.8*	5.2	43.3	48.4	43.2
100	0.5	0.6	5.5*	7.1*	5.6*	5.1	38.2	42.6	37.4
		0	5.0	5.3	5.1	4.8	89.2	92.0	91.5
		0.1	4.9	5.5*	5.2	4.9	85.7	88.0	87.5
		0.4	4.8	5.5*	5.1	4.8	74.9	76.7	75.6
50	0.5	0.6	5.2	5.7*	5.3	5.1	67.7	68.5	67.2
		0	5.3	6.1*	5.7*	5.1	64.9	69.6	68.0
		0.1	5.1	6.0*	5.5*	5.0	60.2	64.3	62.8
		0.4	5.0	5.8*	5.3	4.7	49.4	52.3	50.3
25	0.5	0.6	4.7	5.7*	5.0	4.6	43.6	46.2	44.3
		0	5.0	6.4*	5.7*	4.8	42.5	48.8	46.3
		0.1	5.1	6.4*	5.8*	4.6	38.9	44.0	41.7
		0.4	5.0	6.6*	5.7*	4.7	31.0	35.9	33.4
100	0.5	0.6	5.1	6.6*	5.7*	4.9	27.6	31.3	28.7
		0	5.0	6.4*	5.7*	4.8	42.5	48.8	46.3
		0.1	5.1	6.4*	5.8*	4.6	38.9	44.0	41.7
		0.4	5.0	6.6*	5.7*	4.7	31.0	35.9	33.4

Table II-C. Empirical Type I error rates and power: $n_k \leq 5$ where n_k is non-uniformly distributed, $r_3=0.5$, $r_4=\tau_1/4$, $\delta_0 = -0.1$, $(p_1 - p_0 = -0.1)$ and $\delta_1 = p_1 - p_0 = 0$. Asterisk indicates a value is outside the 95% confidence interval on the nominal level of 0.05.

K	p_0	$r_1=r_2$	Type I error rate (%)				Power (%)			
			Z_{EV}	Z_{ALB}	Z_{AN}	Z_0	Z_{EV}	Z_{ALB}	Z_{AN}	Z_0
100	0.2	0	5.4	6.1*	5.1	5.3	96.7	97.7	97.4	97.3
		0.1	5.4	6.1*	5.2	5.4	95.4	96.4	96.0	95.9
		0.4	5.4	6.2*	5.1	5.5*	89.8	90.8	89.7	89.7
		0.6	5.6*	6.4*	5.3	5.7*	84.1	85.1	83.4	83.7
50	0.2	0	5.6	6.5*	5.3	5.4	79.8	83.8	81.7	81.2
		0.1	5.6	6.6*	5.6*	5.7*	76.9	80.5	78.6	78.0
		0.4	6.1	7.1*	5.5*	5.9*	66.1	68.9	65.7	65.2
		0.6	5.8	6.9*	5.2	5.8*	59.3	61.6	58.0	57.9
25	0.2	0	6.0*	7.6*	5.6*	5.7*	55.3	62.1	57.5	55.4
		0.1	6.2*	7.6*	5.6*	5.8*	52.3	58.2	53.5	51.7
		0.4	6.7*	7.8*	5.5*	6.4*	43.7	48.9	43.6	52.2
		0.6	6.7*	8.1*	5.5*	6.4*	38.8	43.6	38.1	37.3

Table III
Average empirical Type 1 error rate and power of each test over $P_0 = (0.8, 0.5, 0.2)$, $r_1 = r_2 = (0, 0.1, 0.4, 0.6)$, $r_3 = 0.5$ and $r_4 = r_1 / 4$ Those number in a dotted bracket are Type 1 error rates and power for the total number of units $N = k \times n = 200$.

Clusters <i>K</i>		Average Type 1 error rate (%)			Average power (%)		
		<i>n</i> =2	<i>n</i> =4	<i>n</i> =8	<i>n</i> =2	<i>n</i> =4	<i>n</i> =8
100	Z_{Ev}	5.11	5.39	5.25	81.4	91.6	94.9
	Z_{ALB}	5.54	5.89	5.84	82.5	92.3	95.4
	Z_{AN}	4.86	5.34	5.36	81.1	91.7	95.1
	Z_O	4.98	5.33	5.18	81.0	91.5	94.8
50	Z_{Ev}	5.38	5.26	5.16	57.2	72.7	82.3
	Z_{ALB}	5.98	6.16	6.02	60.0	75.1	84.1
	Z_{AN}	4.80	5.32	5.38	56.4	73.2	82.9
	Z_O	5.05	5.12	4.97	56.1	72.3	81.7
25	Z_{Ev}	6.14	5.41	5.27	37.4	48.9	62.5
	Z_{ALB}	5.96	6.94	6.69	38.7	53.8	66.6
	Z_{AN}	4.94	5.63	5.75	35.4	50.3	64.2
	Z_O	5.37	5.13	4.85	34.7	47.8	61.2