Published in final edited form as:

Comput Stat Data Anal. 2011 October 1; 55(10): 2880–2887. doi:10.1016/j.csda.2011.04.020.

Power and sample size requirements for non-inferiority in studies comparing two matched proportions where the events are correlated

Jun-mo Nam*

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute
National Institutes of Health, DHHS, 6120 Executive Blvd, Rockville, MD 20852-7244, U.S.A

Abstract

Consider clustered matched-pair studies for non-inferiority where clusters are independent but units in a cluster are correlated. An inexpensive new procedure and the expensive standard one are applied to each unit and outcomes are binary responses. Appropriate statistics testing non-inferiority of a new procedure have been developed recently by several investigators. In this note, we investigate power and sample size requirement of the clustered matched pair study for non-inferiority. Power of a test is related primarily to the number of clusters. The effect of a cluster size on power is secondary. The efficiency of a clustered matched-pair design is inversely related to the intra-class correlation coefficient within a cluster. We present an explicit formula for obtaining the number of clusters for given a cluster size and the cluster size for a given number of clusters for a specific power. We also provide alternative sample size calculations when available information regarding parameters are limited. The formulae can be useful in designing a clustered matched-pair study for non-inferiority. An example for determining sample size to establish non-inferiority for a clustered matched-pair study is illustrated.

Keywords

binary outcomes; clustered matched-pair; intra-class correlation coefficient; non-inferiority; power; sample size

1. Introduction

When the standard treatment is highly effective but it is expensive, toxic, or difficult to administer, researchers are interested in finding an inexpensive, less toxic, or easily administrable new treatment that is no worse than a pre-specified amount in effectiveness. A statistical test of significance for this research goal is called a non-inferiority test. Lu and Bean (1995) and Nam (1997) presented statistical methods to establish non-inferiority for matched-pair data where the pairs are independent. These methods may not be applicable for clustered matched-pair data where pairs in a cluster are correlated. If the correlation within a cluster is ignored, then the p-value of a test based on no correlation could be biased downward and it may lead to spurious statistical significance, and sample size that is underestimated for a given power. Durkalski et al (2003) provided a proper statistical

*Corresponding author. Tel: +1-301-496-3232; fax +1-301-402-0081. namj@mail.nih.gov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

method for testing non-inferiority for clustered matched-pair data using method of moments (Royal, 1986). Alternatively, Nam and Kwon (2009) proposed a modification of Obuchowski's method (1998) for testing non-inferiority. Their simulations showed that the alternative test is robust with an empirical Type 1 error rate close to nominal. In designing a clustered matched pair study, researchers would want to determine an appropriate sample size required for a specific power, but such a sample size formula has not been fully investigated.

Section 2 gives the notation and model. Section 3 provides the power function of the modified Obuchowski's test, sample size formulae in designing a clustered matched-pair study for non-inferiority and examine the relation of the number of clusters for a given intra-class correlation coefficient and cluster size. Sections 4 and 5 contain an example for illustration and concluding remarks.

2. Model and notations

Consider a random sample 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 to the i^{th} unit contained in the k^{th} cluster. The matched samples can be categorized in four 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, where x_{ijk} is the number of the pair-type (i,j) for $i, j = 0, 1$ in the k^{th} cluster:

The dot in the subscript in the Table 1 denotes a sum over corresponding subscripts, e.g.,

$$x_{1\bullet k} = x_{11k} + x_{10k} \text{ and } n_{\bullet} = \sum_{k=1}^K n_k. \text{ Define overall response rates of the new and standard}$$

$$\text{procedures as } \hat{p}_{1\bullet\bullet} = x_{1\bullet\bullet}/n_{\bullet} \text{ and } \hat{p}_{\bullet 1\bullet} = x_{\bullet 1\bullet}/n_{\bullet} \text{ where } x_{1\bullet\bullet} = \sum_{k=1}^K x_{1\bullet k} \text{ and } x_{\bullet 1\bullet} = \sum_{k=1}^K x_{\bullet 1k}.$$

3. Power of non-inferiority test and sample size

Let the overall difference between the new and standard procedures in the proportions be $\hat{d} = \hat{p}_{1\bullet\bullet} - \hat{p}_{\bullet 1\bullet}$ and its expectation be $E(\hat{d}) = p_{1\bullet\bullet} - p_{\bullet 1\bullet} = \delta$. Denote $\bar{p}_{1\bullet\bullet} = (\hat{p}_{1\bullet\bullet} + \delta_0)/2$ and $\bar{p}_{\bullet 1\bullet} = (\hat{p}_{\bullet 1\bullet} + \delta_0)/2$. When the formulation by Obuchowski (1998) using Rao and Scott (1992) is applied to a non-inferiority setting, the z_O statistic for testing

$$H_0: p_{1\bullet\bullet} - p_{\bullet 1\bullet} = \delta (<0) \text{ against } H_1: p_{1\bullet\bullet} - p_{\bullet 1\bullet} = \delta_1 (>\delta_0) \text{ is}$$

$$z_o = (\hat{d} - \delta_0) / \left\{ \widehat{\text{var}}(\hat{d} - \delta_0)_{p=\bar{p}} \right\}^{1/2} \quad (1)$$

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

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

$$\text{and } \widehat{\text{cov}}(\widehat{p}_{1\bullet\bullet}, \widehat{p}_{\bullet 1\bullet})_{\widetilde{p}=\widetilde{p}} = K(K-1)^{-1} \left\{ \sum_k (x_{1\bullet k} - n_k \bar{p}_{1\bullet k})(x_{\bullet 1k} - n_k \bar{p}_{\bullet 1k}) \right\}$$

where $\widetilde{p}' = (p_{1\bullet\bullet}, p_{\bullet 1\bullet})$ and $\widetilde{p} = (\bar{p}_{1\bullet\bullet}, \bar{p}_{\bullet 1\bullet})$ (Nam and Kwon, 2009). For a large number of clusters, the statistic (1) is approximately a standard normal deviate. We reject H_0 at a level of α when $z_O \geq z_{(1-\alpha)}$ where $z_{(1-\alpha)}$ is the $100 \times (1-\alpha)$ percentile point of the standard normal distribution. Consider equal cluster sizes, i.e., $n_k = n$ for $k=1, 2, \dots, K$, for simplicity. From (A3) and (A4) in Appendix 1, we can show that the variance of $\hat{d}$ under $\delta = \delta_0$ can be expressed as

$$v_0 \equiv \text{var}(\hat{d})_0 = (\sigma_{b0}^2 + \sigma_w^2/n)/K \quad (2)$$

where $\sigma_{b0}^2 = \sum_k (p_{10k} - p_{01k} - \delta_0)^2 / (K-1)$ and $\sigma_w^2 = \sum_k \{p_{10k} + p_{01k} - (p_{10k} - p_{01k})^2\} / (K-1)$ are the between and the within cluster components of $\text{var}(\hat{d})_0$, and the variance of $\hat{d}$ under $\delta = \delta_1$, is written as

$$v_1 \equiv \text{var}(\hat{d})_1 = (\sigma_{b1}^2 + \sigma_w^2/n)/K \quad (3)$$

where $\sigma_{b1}^2 = \sum_k (p_{10k} - p_{01k} - \delta_1)^2 / (K-1)$ is the between cluster component of $\text{var}(\hat{d})_1$.

When $n = 1$, $v_0 = (\sigma_{b0}^2 + \sigma_w^2)/K$. If $p_{10k} - p_{01k} = \delta_0$ for $k=1, 2, \dots, K$ under H_0 , then $\sigma_{b0}^2 = 0$ so that $v_0 = (2p_{01\bullet} + \delta_0 - \delta_0^2)/K$ for a large K , from (2). Similarly, if $p_{10k} - p_{01k} = \delta_1$ across clusters under H_1 , then $\sigma_{b1}^2 = 0$ and $v_1 = (2p_{01\bullet} + \delta_1 - \delta_1^2)/K$ from (3). The variances, v_0 and v_1 are identical to those of Wald-type test statistic for non-inferiority under null and alternative hypotheses in K independent matched-pair studies (Nam, 1997). The expression of (2) and (3) in terms of the between and within cluster components of $\text{var}(\hat{d})$ under null and alternative hypotheses are roughly the same as the expression of the variance of the estimated mean in two-stage sampling described by Cochran (1977).

3.1. Power

The power of the z_O test for $H_0: \delta = \delta_0 (<0)$ against $H_1: \delta = \delta_1 (>\delta_0)$ at a level of α is expressed as

$$\text{power} = P(z_O \geq z_{(1-\alpha)} | H_1: \delta = \delta_1) = 1 - \Phi(u) \quad (4)$$

where $u = \{v_0^{1/2} \cdot z_{(1-\alpha)} - (\delta_1 - \delta_0)\} / v_1^{1/2}$ and $\Phi(u)$ is the cumulative normal density function of u .

Then u can be rewritten in terms of K and n as

$$u = \{(\sigma_{b0}^2 + \sigma_w^2/n)^{1/2} \cdot z_{(1-\alpha)} - K^{1/2}(\delta_1 - \delta_0)\} / (\sigma_{b1}^2 + \sigma_w^2/n)^{1/2}. \quad (5)$$

From (5), we have $u \rightarrow -\infty$ as $K \rightarrow \infty$. Namely, the power of the test approaches one (100%) as the number of clusters increases to infinity for any fixed value of an alternative from (4). As $n \rightarrow \infty$, however, we have $u \rightarrow \{\sigma_{b0} \cdot z_{(1-\alpha)} - K^{1/2} (\delta_1 - \delta_0)\} / \sigma_{b1}$ for a given number of clusters K and the power approaches to a limit which is less than one. Even when the cluster size n is infinitively large, the power does not reach one. The number of clusters has a major effect on power while cluster sizes have a limited effect on the power of a test.

3.2. Sample size determination

From (4) and (5), the number of clusters and the cluster size required for power = $1 - \beta$ of the z_O test at level α can be obtained by the following relation:

$$v_0^{1/2} \cdot z_{(1-\alpha)} + v_1^{1/2} \cdot z_{(1-\beta)} = \delta_1 - \delta_0. \quad (6)$$

From (5) and (6), the number of clusters for a given cluster size n is

$$K = \{(\sigma_{b0}^2 + \sigma_w^2/n)^{1/2} \cdot z_{(1-\alpha)} + (\sigma_{b1}^2 + \sigma_w^2/n)^{1/2} \cdot z_{(1-\beta)}\}^2 / (\delta_1 - \delta_0)^2. \quad (7)$$

The cluster size n for a given number of clusters K is numerically obtained from (7). Cluster sizes beyond four ($n > 4$) contribute little in reducing the number of clusters for a given power when σ_w is small (see Table 2A). The test based on the z_O statistic cannot achieve a desired power such as 80% or 90% even when the cluster size is infinitively large if the number of clusters is less than a certain number. For example, as indicated in Table 2A, the power of a test cannot reach 80% when K is less than or equal to 18. The expanded Table 2B shows that the minimum number of clusters required for a specific power is related to the variance components and the non-inferiority margin. Table 2A also shows that the total number of units where the two procedures are applied to each unit increases as the cluster size increases for a given power and it reaches the minimum when the cluster size $n=2$. The form (7) indicates the required number of clusters is inversely related with the square of a difference between null and alternative hypotheses (i.e., non-inferiority margin).

For non-inferiority testing, δ_0 and δ_1 are usually small negative and non-negative values, respectively. We may use the between cluster component of the variance under $\delta = 0$, i.e.,

$$\sigma_b^2 = \sum_k (p_{1\bullet k} - p_{\bullet 1k})^2 / K \quad \text{for an approximation. Then the number of clusters is found as}$$

$$K = \{z_{(1-\alpha)} + z_{(1-\beta)}\}^2 \cdot (\sigma_b^2 + \sigma_w^2/n) / (\delta_1 - \delta_0)^2. \quad (8)$$

Alternatively, σ_b^2 can be replaced by $\sigma_b'^2$ which it is an average of σ_{b0}^2 and σ_{b1}^2 .

The approximation (8) is crude but simple, and it is convenient in examining factors related to a number of clusters. Equation (8) can be rewritten as

$z_{(1-\alpha)} + z_{(1-\beta)} = K^{1/2} (\delta_1 - \delta_0) / (\sigma_b^2 + \sigma_w^2/n)^{1/2}$, where the right-hand side is called the non-central parameter of the asymptotical normal distribution of the statistic z_O , e.g., Odeh and Fox (1991). The asymptotic power of the z_O test is an increasing function of the non-central parameter. We examine the form (8) for a relation between a required number of clusters and an intra-class correlation. In many cases, estimates of σ_w^2 and intra-class correlation

coefficient ρ are available but an estimate of σ_b^2 may not be given. Since $\rho = \sigma_b^2 / (\sigma_b^2 + \sigma_w^2)$, σ_b^2 can be obtained from σ_w^2 and ρ as $\sigma_b^2 = \{\rho / (1 - \rho)\} \bullet \sigma_w^2$. From the equation (8), we have

$$K = f(\alpha, \beta, \delta_0, \delta_1) \bullet \{\rho / (1 - \rho) + 1/n\} \bullet \sigma_w^2$$

and

$$n = f(\alpha, \beta, \delta_0, \delta_1) \bullet \sigma_w^2 / \left[K - f(\alpha, \beta, \delta_0, \delta_1) \bullet \{\rho / (1 - \rho)\} \bullet \sigma_w^2 \right].$$

where $f(\alpha, \beta, \delta_0, \delta_1) \equiv \{z_{(1-\alpha)} + z_{(1-\beta)}\}^2 / (\delta_1 - \delta_0)^2$.

Table 3 shows the number of clusters divided by σ_w^2 (i.e., $K' = K / \sigma_w^2$) for 80% power of the z_O test at 0.05 level for testing the null hypothesis $\delta_0 = -0.1$ against an alternative $\delta_1 = 0$ for various values of cluster size and intra-class correlation coefficient. When $\sigma_w^2 = 0.05$, the numbers of clusters and cluster sizes for 80% power of the test are illustrated in Figure 1, which demonstrates a strong positive relationship between K and ρ . The relative difference between a clustered matched-pair design and the independent matched-pair design in terms of the number of clusters for the same Type 1 error rate, power, null and alternative hypotheses can be expressed as

$$r = (K - K_1) / K_1 = -(1 - \rho)(1 - 1/n)$$

where K is the number of clusters required for a clustered matched-pair design with a size n per cluster and K_1 is the number of units (i.e., $n=1$) for the independent matched-pair design (Appendix 2). The relative benefit of a clustered matched-pair design in terms of the number of clusters becomes greater as $\rho \rightarrow 0$ and/or $n \rightarrow \infty$. Fig 2 shows the relationship of r with ρ and n . There is no gain by measuring more than one unit per cluster when $\rho = 1$. When $\rho = 0$, relative reduction rates due to a clustered matched-pair design for cluster sizes $n=2, 5$ and 10 are 50%, 80% and 90% respectively. The relative reduction rate of a typical case is in between those of the extreme cases ($\rho = 0$ and 1), and the rate in the relative reduction increases as $n \rightarrow \infty$ but its increment diminishes. Equations (7) and (8) can be also applicable for sample size determination for conventional one-sided testing of $\delta_0 = 0$ against $\delta_1 > 0$ in clustered matched-pair studies when there is suitable modification of the between cluster components under the null and alternative hypotheses.

4. An example

Consider a study comparing the specificity of two diagnostic procedures, i.e., positron emission tomography (PET) and single photo emission CT scan (SPECT), for the detection of abnormal parathyroid tissues in patient with primary hyperparathyroidism (Neuman et al., 1996). Specificity is the proportion of true negative test results among glands confirmed not to have parathyroidism at surgery. 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 an average of 2.4 glands (See Table VI, Obuchowski (1998)). The data are organized in Table 4 where $\hat{p}_{ijk} = x_{ijk} / n_k$ for $i, j = 0, 1$ and $k = 1, 2, \dots, K$. Consider a patient and glands in a patient as a cluster and units, respectively.

The SPECT is inexpensive and is performed quickly while the PET is expensive and provides a high resolution image. Considering factors involving cost, time and accuracy, we may make a value judgment that the SPECT is non-inferior to the PET when the former must be no more than 10 percent lower than the latter in specificity, i.e., testing null and alternative hypotheses for non-inferiority are $H_0 : \delta = -0.1$ and $H_1 : \delta > -0.1$. From (1), we have the $z_O = 2.68$ ($p=0.004$) and reject $\delta_0 = -0.1$ in a favor of $\delta_1 > -0.1$ at $\alpha=0.05$.

The between cluster component of $\text{var}(\hat{p}_{1..} - \hat{p}_{.1.})$ under the null hypothesis of $\delta_0 = -0.1$ and

under an alternative of $\delta_1 = 0.1$ are estimated as $\hat{\sigma}_{b0}^2 = \sum_{k=1}^{21} (\hat{p}_{10k} - \hat{p}_{01k} + 0.1)^2 / 20 = (0.365)^2$ and

$\hat{\sigma}_{b1}^2 = \sum_{k=1}^{21} (\hat{p}_{10k} - \hat{p}_{01k} - 0.1)^2 / 20 = (0.300)^2$, respectively, from (A3) and $\hat{\sigma}_b^2 = (0.318)^2$. The within cluster component of the variance is estimated as

$\hat{\sigma}_w^2 = \sum_k \{ \hat{p}_{10k} - \hat{p}_{01k} - (\hat{p}_{10k} - \hat{p}_{01k})^2 \} / 20 = (0.239)^2$ from (A3). Since $u = -0.788$, the power for the z_O test is 78% from (4) and (5). For a closer alternative hypothesis, $\delta_1 = 0.05$, the power of the test is 54%. Note that estimates of the between and within cluster components of the variance are used as values of the parameters in power calculation. When varying cluster sizes are considered, instead of equal cluster sizes, estimates of parameters are

$\hat{\sigma}_{b0}^2 = (0.377)^2$ under $\delta_0 = -0.1$, $\hat{\sigma}_{b1}^2 = (0.291)^2$ under $\delta_1 = 0.1$ and $\hat{\sigma}_w^2 = (0.255)^2$ from (A5) and (A6). From (4) and (5), we have $u = -0.712$ and the power of the test is 76% which is slightly lower than the power assuming equal cluster sizes for $\delta_1 = 0.1$. The slightly greater power as a result of equal cluster sizes is explained by an increase in efficiency from there being a balanced design and from the relative insensitivity of cluster sizes on power.

Suppose we want to conduct a definitive study involving non-inferiority of SPECT of specificity using the same values of parameters obtained from this study. From (7), (8) and $\sigma_b'^2$ replacing σ_b^2 in (8), the number of clusters required for a power = 80% using the statistic z_o for testing $\delta_0 = -0.1$ against $\delta_1 = 0.05$ for a given average cluster size $n = 51/21 = 2.43$ are $K = 39, 34$ and 37 , and for testing $\delta_0 = -0.10$ against $\delta_1 = 0.10$ are $K = 22, 19$ and 21 , respectively. In a range of alternative hypotheses and power that are of interested in this example, the number of clusters determined by (8) and obtained by replacing σ_b^2 as $\sigma_b'^2$ in (8) are smaller than the number of clusters determined by (7). A discrepancy in the required number is due to the use of different estimates of the between cluster component of the variance. The number of clusters required for a given power is very sensitive to the magnitude of the difference between the null and alternative hypotheses.

5. Remarks

The clustered matched-pair design is basically different from the matched-pair cluster design. In the matched-pair cluster design, a cluster is matched with a similar cluster, and one cluster is assigned randomly to intervention and the other to control. Statistical methods applicable for the analysis of data arising from the matched-pair cluster design have been investigated by, e.g., Donner (1987) and Thompson et al (1997), and sample size requirements for stratified cluster randomization designs for a conventional hypothesis testing by, e.g. Hsieh (1998) and Donner (1992). In this paper, we consider the clustered matched-pair design where clusters are randomly sampled and units in a cluster are measured by both new and standard procedures and the two measures on the same unit are matched. We are interested in determining the number of clusters and cluster size for designing a clustered matched-pair study to establish non-inferiority of a new procedure.

Power of a non-inferiority test with clustered matched-pair data is primarily dependent on a number of clusters, and the cluster size plays a secondary role with limited effect on power. Clusters are independent but units in a cluster are correlated. When a cluster size $n \geq 2$ is considered, the number of clusters given cluster size $n=2$ for a specific power of a non-inferiority test at level α results in the minimum number of total units (which is the number of clusters multiplied by the cluster size) to achieve the required power. It may be called optimal in terms of number of measurements required. The cluster sizes may be fixed or may vary. For variable cluster sizes, we may use an average cluster size to find the number of clusters for a given power. It is simpler and provides a good approximation. If the number of clusters is smaller than a minimum boundary point corresponding to a specified power for given values of variance components and non-inferiority margin, then we cannot attain the power no matter how large the cluster size is. Thus, it makes more sense to determine the number of clusters for a required power given the cluster size instead of the other way round in designing studies. In designing non-inferiority studies, we have to specify the indifference margin. Determination of a non-inferiority margin requires a value judgment that incorporates factors such as safety, ease of administration, cost and effectiveness. Recently, the US Food Drug Administration (2010) issued a draft report providing guidance on non-inferiority trials in which the analytical and statistical approaches used to form the non-inferiority margin with examples to determine non-inferiority margins for conducting non-inferiority studies.

Acknowledgments

The author thanks two referees for their careful review and constructive comments that improve the manuscript. This research was supported by the Intramural Research Program of the NIH, National Cancer Institute.

References

- Cochran, WG. Sampling Techniques. 3. Wiley; New York: 1977.
- Donner A. Statistical methodology for paired cluster designs. American Journal of Epidemiology. 1987; 126:972–979. [PubMed: 3661544]
- Donner A. Sample size requirements for stratified cluster randomization designs. Statistics in Medicine. 1992; 11:743–750. [PubMed: 1594813]
- Durkalski VL, Palesch YY, Lipsitz SR, Rust PE. Analysis of clustered matched-pair data for non-inferiority study design. Statistics in Medicine. 2003; 22:279–290. [PubMed: 12520562]
- Hsieh FY. Sample size formulae for intervention studies with the cluster as unit of randomization. Statistics in medicine. 1988; 8:1195–1201. [PubMed: 3201045]
- 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]
- Nam J. Establishing equivalence of two treatments and sample size requirements in matched-pair data. Biometrics. 1997; 53:1422–1430. [PubMed: 9423257]
- Nam J, Kwon D. Non-inferiority tests for clustered matched-pair data. Statistics in Medicine. 2009; 28:1668–1679. [PubMed: 19326387]
- Neumann DR, Esselstyn CB, MacIntyre WJ, Go RT, Obuchowski NA, Chen EQ, Licate AA. Comparison of FDG PET and Sestamibi-SPECT in primary hyperparathyroidism. Journal of Nuclear Medicine. 1996; 37:1809–1815. [PubMed: 8917180]
- Obuchowski NA. On the comparison of correlated proportions for clustered data. Statistics in Medicine. 1998; 17:1495–1507. [PubMed: 9695194]
- Odeh, RE.; Fox, M. Sample Size Choice: Charts for Experiments with Linear Models. Marcel Dekker; New York: 1991.
- Rao JNK, Scott AJA. A simple method for the analysis of clustered binary data. Biometrics. 1992; 48:577–585. [PubMed: 1637980]

- Royall RM. The prediction approach to robust variance estimation in two-stage cluster sampling. *Journal of American Statistical Association*. 1986; 81:119–123.
- Thompson SG, Pyke SDM, Hardy RJ. The design and analysis of paired cluster randomized trials: an application of meta-analysis techniques. *Statistics in Medicine*. 1997; 16:2063–2079. [PubMed: 9308132]
- US Food and Drug Administration. Guidance for Industry Non-Inferiority Clinical Trials. Center for Drug Evaluation and Research, Silver Spring, Maryland, and Center for Biologic Evaluation and Research; Rockville, Maryland: 2010.

Appendix 1: Within and between cluster components of $\text{var}(\hat{p}_{1\bullet\bullet} - \hat{p}_{\bullet 1\bullet} - \delta)$

Consider a balanced design where the cluster sizes are the same, i.e., $n_k = n$ for $k=1, 2, \dots, K$.

From (1), the estimated variances and covariance are written as

$$\begin{aligned}\widehat{\text{var}}(\hat{p}_{1\bullet\bullet}) &= \sum_k \{(x_{1\bullet k} - np_{1\bullet k}) + n(p_{1\bullet k} - \bar{p}_{1\bullet\bullet})\}^2 / \{(K-1)Kn^2\}, \\ \widehat{\text{var}}(\hat{p}_{\bullet 1\bullet}) &= \sum_k \{(x_{\bullet 1k} - np_{\bullet 1k}) + n(p_{\bullet 1k} - \bar{p}_{\bullet 1\bullet})\}^2 / \{(K-1)Kn^2\}\end{aligned}\quad (\text{A1})$$

and

$$\widehat{\text{cov}}(\hat{p}_{1\bullet\bullet}, \hat{p}_{\bullet 1\bullet}) = \sum_k \{(x_{1\bullet k} - np_{1\bullet k}) + n(p_{1\bullet k} - \bar{p}_{1\bullet\bullet})\} \{(x_{\bullet 1k} - np_{\bullet 1k}) + n(p_{\bullet 1k} - \bar{p}_{\bullet 1\bullet})\} / \{(K-1)Kn^2\}$$

where $\bar{p}_{1\bullet\bullet} = \{(x_{1\bullet\bullet} + x_{\bullet 1\bullet})/(Kn) + \delta\}/2$ and $\bar{p}_{\bullet 1\bullet} = \{(x_{1\bullet\bullet} + x_{\bullet 1\bullet})/(Kn) - \delta\}/2$ so that $\bar{p}_{1\bullet\bullet} - \bar{p}_{\bullet 1\bullet} = \delta_0$. After the expansion of (A1), summarize $\widehat{\text{var}}(\hat{p}_{1\bullet\bullet} - \hat{p}_{\bullet 1\bullet} - \delta)$ as $\widehat{\text{var}}(\hat{p}_{1\bullet\bullet}) + \widehat{\text{var}}(\hat{p}_{\bullet 1\bullet}) - 2\widehat{\text{cov}}(\hat{p}_{1\bullet\bullet}, \hat{p}_{\bullet 1\bullet}) = B1 + B2 + B3$

where

$$\begin{aligned}B1 &= \sum_k \{(x_{1\bullet k} - np_{1\bullet k})^2 + (x_{\bullet 1k} - np_{\bullet 1k})^2 - 2(x_{1\bullet k} - np_{1\bullet k})(x_{\bullet 1k} - np_{\bullet 1k})\} / \{(K-1)Kn^2\}, \\ B2 &= \sum_k \{(p_{1\bullet k} - \bar{p}_{1\bullet\bullet})^2 + (p_{\bullet 1k} - \bar{p}_{\bullet 1\bullet})^2 - 2(p_{1\bullet k} - \bar{p}_{1\bullet\bullet})(p_{\bullet 1k} - \bar{p}_{\bullet 1\bullet})\} / \{(K-1)Kn\}\end{aligned}$$

and

$$B3 = 2 \sum_k \{(x_{1\bullet k} - np_{1\bullet k})(p_{1\bullet k} - p_{\bullet 1k} - \delta) - (x_{\bullet 1k} - np_{\bullet 1k})(p_{1\bullet k} - p_{\bullet 1k} + \delta)\} / \{(K-1)Kn\}.$$

The expectation of $B1$, $B2$ and $B3$ are

$$E(B1) = \sum_k \{p_{10k} + p_{01k} - (p_{10k} - p_{01k})^2\} / \{(K-1)Kn\} \quad (\text{A2})$$

and $E(B2) = \sum (p_{10k} - p_{01k} - \delta)^2 / \{(K-1)K\}$ and $E(B3) = 0$.

From (A2), the variance of $(\hat{p}_{1..} - \hat{p}_{.1.} - \delta_0)$ under $H_0 : \delta = \delta_0$ is expressed as

$$\text{var}(\hat{p}_{1..} - \hat{p}_{.1.}) = (\sigma_{b0}^2 + \sigma_w^2/n)/K \quad (\text{A3})$$

$$\text{where } \sigma_{b0}^2 = \sum_k \{(p_{10k} - p_{01k} - \delta_0)^2\}/(K-1) \text{ and } \sigma_w^2 = \sum_k \{p_{10k} + p_{01k} - (p_{10k} - p_{01k})^2\}/(K-1)$$

Similarly, the variance of $(\hat{p}_{1..} - \hat{p}_{.1.} - \delta_1)$ under $H_1 : \delta = \delta_1$ is written as

$$\text{var}(\hat{p}_{1..} - \hat{p}_{.1.} - \delta_1) = (\sigma_{b1}^2 + \sigma_w^2/n)/K \quad (\text{A4})$$

$$\text{where } \sigma_{b1}^2 = \sum_k (p_{10k} - p_{01k} - \delta_1)^2/(K-1).$$

When the cluster size varies over k , the expectations of B 's are

$$E(B1) = \left[K \sum_k n_k \{p_{10k} + p_{01k} - (p_{10k} - p_{01k})^2\}/(K-1) \right] / n_{..}^2, \\ E(B2) = \left[K \sum_k n_k^2 (p_{10k} - p_{01k} - \delta^2)/(K-1) \right] / K \text{ and } E(B3) = 0.$$

Denote the fraction of n_k to $n_{..}$ as $f_k = n_k/n_{..}$ for $k = 1, 2, \dots, K$.

The $\text{var}(\hat{p}_{1..} - \hat{p}_{.1.})$ can be expressed as

$$\text{var}(\hat{p}_{1..} - \hat{p}_{.1.}) = \sigma_{b0}^2/K + \sigma_w^2/n_{..} \quad (\text{A5})$$

under the null hypothesis and

$$\text{var}(\hat{p}_{1..} - \hat{p}_{.1.}) = \sigma_{b1}^2/K + \sigma_w^2/n_{..} \quad (\text{A6})$$

under the alternative hypothesis where

$$\sigma_{b0}^2 = K^2 \sum_k f_k^2 (p_{10k} - p_{01k} - \delta_0)^2/(K-1), \\ \sigma_{b1}^2 = K^2 \sum_k f_k^2 (p_{10k} - p_{01k} - \delta_1)^2/(K-1),$$

and

$$\sigma_w^2 = K \sum_k f_k \{p_{10k} + p_{01k} - (p_{10k} - p_{01k})^2\}/(K-1).$$

Note that the forms (A3) and (A4) are those special cases of (A5) and (A6) with $n_k = n$, i.e., $f_k = 1/K$ for $k = 1, 2, \dots, K$.

Appendix 2. Relative difference between clustered matched-pair design and non-clustered matched-pair design (n=1) in a number of clusters

From (8), we have $K = \{z_{(1-\alpha)} + z_{(1-\beta)}\}^2 \cdot (\sigma_b^2 + \sigma_w^2/n) / (\delta_1 - \delta_0)^2$ and

$K_1 = \{z_{(1-\alpha)} + z_{(1-\beta)}\}^2 \cdot (\sigma_b^2 + \sigma_w^2) / (\delta_1 - \delta_0)^2$ for $n=1$. The relative difference between K and K_1 for the same type 1 error probability, power and null and alternative hypotheses is

$$r = (K - K_1) / K_1 = \{\sigma_w^2 / (\sigma_b^2 + \sigma_w^2)\} \cdot (1/n - 1) = -(1 - \rho)(1 - 1/n).$$

Fig 1.
Number of clusters for 80% power of z_0 test at 0.05 level testing $\delta_0 = -0.1$ against $\delta_1 = 0$
when $\sigma_w^2=0.05$

Figure 2. Relative reduction rate of number of clusters by clustered matched-pair design ($n \geq 2$) to that by independent matched-pair design ($n = 1$)

Table 1
Observations and probabilities for matched pairs in the k^{th} cluster

Standard procedure			
New procedure	1	0	sum
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	$x_{\bullet 1k} (p_{\bullet 1k})$	$x_{\bullet 0k} (p_{\bullet 0k})$	$n_k (1)$

Table 2A

Number of clusters (K), cluster size (n) and total number of units required for power=(80%, 90%) of z_0 test for $H_0: \delta_0 = -0.1$ against $H_1: \delta_1 = 0.1$ at a level of 0.05 when $\sigma_{b0} = 0.365$, $\sigma_{b1} = 0.300$, $\sigma_w = 0.239$

Cluster size (n)	80% power		90% power	
	# of clusters (K)	Total units (nK)	# of clusters (K)	Total units (nK)
2	23	46	30	60
3	21	63	28	84
4	20	80	27	108
5	20	100	27	135
6	20	120	26	156
10	19	190	26	260
∞	18	∞	24	∞

Number of clusters (K), required for power = (80%, 90%) of z_0 test for $H_0 : \delta_0 = -0.1$ against $H_1 : \delta_1 = (0,0.05,0.1)$ when $\sigma_{b0} = \sigma_{b1} = \sigma_b$. $\Delta = \delta_1 - \delta_0 = (0.1,0.15,0.2)$

Table 2B

		80% power						90% power					
σ_v	σ_b	$\Delta = 0.1$		$\Delta = 0.15$		$\Delta = 0.2$		$\Delta = 0.1$		$\Delta = 0.15$		$\Delta = 0.2$	
		$n = 4$	$n = \infty$	$n = 4$	$n = \infty$	$n = 4$	$n = \infty$	$n = 4$	$n = \infty$	$n = 4$	$n = \infty$	$n = 4$	$n = \infty$
0.1	0.1	8	6	3	3	2	2	11	9	5	4	3	2
	0.3	57	56	25	25	14	14	79	77	35	34	20	19
	0.5	156	155	69	69	39	39	216	214	96	95	54	54
0.3	0.1	20	6	9	3	5	2	28	9	12	4	7	2
	0.3	70	56	31	25	17	14	96	77	43	34	24	19
	0.5	169	155	75	69	42	39	233	214	104	95	58	54

Table 3

Number of clusters divided by $\sigma_w^2(K'=K/\sigma_w^2)$ for 80% power of z_0 test at 0.05 testing $\delta_0 = -0.1$ against $\delta_1 = 0$ ($\Delta = \delta_1 - \delta_0 = 0.1$). K is obtained by $K' \sigma_w^2$.

n	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.9$
2	480	729	1,178	2,225	7,459
3	349	598	1,046	2,094	7,328
4	284	533	981	2,028	7,262
5	244	494	942	1,990	7,223
6	218	467	916	1,962	7,197
7	199	449	897	1,944	7,178
8	185	435	883	1,930	7,164
9	174	424	872	1,919	7,153
10	166	415	864	1,910	7,145
∞	87	336	785	1,832	7,066

Note: K' value for testing $\delta_0 = -0.1$ against $\delta_1 = 0.1$ ($\Delta = 0.2$) is one quarter of the corresponding value in Table 3.

Table 4
Matched pairs of SPECT and PET in specificity by patient (Obuchowski, 1998)

Patient number	Number of glands n_k	Matched pair-types (SPECT, PET) Observed number			
		(1, 1) X_{11k}	(1, 0) X_{10k}	(0, 1) X_{01k}	(0, 0) X_{00k}
1	3	0	2	0	1
2	3	2	1	0	0
3	3	3	0	0	0
4	1	1	0	0	0
5	3	2	1	0	0
6	4	4	0	0	0
7	3	3	0	0	0
8	2	2	0	0	0
9	2	1	0	1	0
10	1	1	0	0	0
11	3	2	0	0	1
12	2	2	0	0	0
13	3	3	0	0	0
14	2	2	0	0	0
15	2	0	2	0	0
16	3	2	0	0	1
17	3	2	0	0	1
18	3	2	1	0	0
19	2	2	0	0	0
20	1	1	0	0	0
21	2	2	0	0	0
sum	51	39	7	1	4