American Journal of Biostatistics

Original Research Paper

# Adaptive Superiority and Noninferiority Trial Design with Paired Binary Data

[1,2]Mark Chang and [2]Jing Wang

[1]AMAG Pharmaceuticals, Inc., Cambridge, MA, USA
[2]Boston University, Boston, MA, USA

*Article history*
Received: 31-01-2015
Revised: 05-05-2015
Accepted: 28-08-2015

Corresponding Author:
Jing Wang
Boston University, Boston, MA, USA
Email: jwa222@bu.edu

**Abstract:** Non-inferiority of a diagnostic test to the standard is a common issue in medical research. For instance, we may be interested in determining if a new diagnostic test is noninferior to the standard reference test because the new test might be inexpensive to the extent that some small inferior margin in sensitivity or specificity may be acceptable. Noninferiority trials are also found to be useful in clinical trials, such as image studies, where the data are collected in pairs. Conventional noninferiority trials for paired binary data are designed with a fixed sample size and no interim analysis is allowed. Adaptive design which allows for interim modifications of the trial becomes very popular in recent years and are widely used in clinical trials because of its efficiency. However, to our knowledge there is no adaptive design method available for noninferiority trial with paired binary data. In this study, we developed an adaptive design method for non-inferiority trials with paired binary data, which can also be used for superiority trials when the noninferiority margin is set to zero. We included a trial example and provided the SAS program for the design simulations.

**Keywords:** Non-Inferiority, Adaptive Design, Power, Sample Size, Paired Data, Matched Data

## Introduction

### *Noninferiority Design*

As the European regulatory agency, Committee for Medicinal Products for Human Use (CHMP, 2005) stated, "Many clinical trials comparing a test product with an active comparator are designed as noninferiority trials. The term 'noninferiority' is now well established, but if taken literally could be misleading. The objective of a noninferiority trial is sometimes stated as being to demonstrate that the test product is not inferior to the comparator. However, only a superiority trial can demonstrate this. In fact a noninferiority trial aims to demonstrate that the test product is not worse than the comparator by more than a pre-specified, small amount. This amount is known as the noninferiority margin, or delta."

Until recent years, the majority of clinical trials were designed for superiority to a comparative drug (the control group). A statistic shows that only 23% of all NDAs from 1998 to 2002 were innovative drugs and the rest were accounted for as "me-too" drugs (Chang, 2010). The "me-too" drugs are judged based on noninferiority criteria. The increasing popularity of noninferiority trials is a reflection of regulatory and industry adjustments in response to increasing challenges in drug development.

From a methodological perspective, Chan (2001) derived power and sample size formulations for noninferiority trials using an exact method. Kong *et al*. (2004), studied noninferiority diagnostic test for using a bivariate normal distribution. Wiens and Heyes (2003) proposed analysus strategy that allows to consider interactions in noninferiority trials. Liu *et al*. (2002) investigated two asymptotic test statistics, a Wald-type test statistic (sample-based) and a Restricted Maximum Likelihood Estimation (RMLE-based) test statistic, to assess non-inferiority based on paired binary endpoints. They found that the RMLE-based test controls type I error better than the sample-based test. Lu and Bean (1995) and Nam (1997) proposed test statistics and sample size determination for comparing two diagnostic methods for the non-inferiority test of sensitivity. Lu *et al*. (2003) discussed simultaneous comparisons of

sensitivity and specificity. However, all these methods are only applicable for classical design with fixed sample size. We will develop in this study an adaptive design method for noninferiority trials with paired binary endpoint and discuss its application in diagnosis test.

There are three major sources of uncertainty about the conclusions from a non-inferiority (NI) study: (1) The uncertainty of the active-control effect over a placebo, which is estimated from historical data, (2) the possibility that the control effect may change over time, violating the "constancy assumption" and (3) the risk of making a wrong decision from the test of the noninferiority hypothesis in the NI study, i.e., the type-I error. These three uncertainties have to be considered in developing a noninferiority design method.

## Commonly Used Noninferiority Design Methods

Most commonly used noninferiority trials are based on parallel, two-group designs. Three-group designs with a placebo may sometimes be used, but they are not very cost-effective and often face ethical challenges when including a placebo group, especially in the United States.

There are three commonly used methods of noninferiority designs: The fixed-margin method, the $\lambda$-portion method and the synthesis method (in original and log scales). We denote the test and the active-control groups by subscripts $T$ and $C$, respectively. Where there is no confusion, the letter $T$ will also be used for test statistics. We will use the hat "^" to represent an estimate of the corresponding parameter, e.g., $\hat{\theta}$ is an estimate of $\theta$.

### *Fixed-Margin Method*

The null hypothesis for the fixed-margin method can be defined as:

$$H_o : \theta_T - \theta_C - \delta_{NI} \le 0 \tag{1}$$

where, $\theta$ can be the mean, hazard rate, adverse event rate, recurrent events rate, or the mean number of events. The constant noninferiority margin $\delta_{NI}\le 0$ (assuming a larger value of the parameter is desirable; otherwise, $\delta_{NI}$ should be larger than zero) is usually determined based on a historical placebo control study (see more discussions later). When $\delta_{NI} = 0$, (1) becomes a null hypothesis test for superiority.

The rejection of (1) can be expressed a simple way: The test drug T is not inferior to C by $\delta_{NI}$ or more.

### *$\lambda$-Portion Method*

The null hypothesis for the $\lambda$-portion method is given by:

$$H_o : \theta_T - \lambda_{NI}\theta_C \le 0 \tag{2}$$

where, $0<\lambda_{NI}<1$. For the superiority test, $\lambda_{NI} = 1$.

The rejection of (2) can be interpreted in layman's terms: Drug $T$ is at least $100\lambda_{NI}$% as effective as drug $C$.

### *Synthesis Method*

The null hypothesis for the synthesis method is given by:

$$H_o : \frac{\theta_T - \theta_P}{\theta_C - \theta_P} - \lambda_{NI} \le 0 \tag{3}$$

Assuming we have proved $\theta_C-\theta_P>0$, (3) is then equivalent to:

$$H_o : \theta_T - \theta_C + \left(1-\lambda_{NI}\right)\left(\theta_C - \theta_P\right) \le 0 \tag{4}$$

where, $0<\lambda_{NI}<1$. For the superiority test, $\lambda_{NI} = 1$.

The rejection of (3) can summed up in these terms: The test drug $T$ is at least $100\lambda_{NI}$% as effective as $C$ after subtracting the placebo effect. When $\lambda_{NI} = 0$, (3) represents a null hypothesis for a putative placebo-control trial.

## Non-Inferiority Design with Fixed-Margin Method for Paired Data

### *Classical Design*

Let $Y_1$ and $Y_2$ be, respectively, binary response variables of treatments 1 and 2 with the joint distribution $P(Y_1 = i;\ Y_2 = j) = p_{ij}$ for $i = 0, 1;\ j = 0, 1$. $\sum_{i=0}^{1}\sum_{j=0}^{1} p_{ij} = 1$. Paired data are commonly displayed in a 2×2 contingency table (Table 1).

Nam (1997; Tango, 1998) proposed the following asymptotic test for paired data:

$$H_o : p_{10} - p_{01} - \delta_{NI} \le 0 \ vs.\ H_a : \bar{H}_o \tag{5}$$

where, $\delta_{NI}<0$ is the noninferiority margin. The test statistic is defined as:

$$Z = \frac{\hat{\varepsilon}\sqrt{n}}{\hat{\sigma}} \tag{6}$$

Table 1. Matched-pair data

| | Test | Total |
|---|---|---|
| Control | 1 0 | |
| 1 | $x_{11}$ $x_{10}$ | |
| 0 | $x_{01}$ $x_{00}$ | |
| Total | | n |

Where:

$$\begin{cases} \hat{\varepsilon} = \hat{p}_{10} - \hat{p}_{01} - \delta_{NI}, \\ \hat{p}_{ij} = x_{ij} / n \\ \hat{\sigma}^2 = 2\tilde{p}_{01} + \delta_{NI} - \delta_{NI}^2 \end{cases} \tag{7}$$

And $\hat{p}_{10}$ is the restricted MLE of $p_{10}$:

$$\begin{cases} \tilde{p}_{01} = \dfrac{-b + \sqrt{b^2 - 8c}}{4} \\ b = \left(2 + \hat{p}_{01} - \hat{p}_{10}\right)\delta_{NI} - \hat{p}_{01} - \hat{p}_{10} \\ c = -\hat{p}_{01}\delta_{NI}\left(1 - \delta_{NI}\right) \end{cases} \tag{8}$$

Nam (1997) proved that under the constraint $p_{10}$-$p_{01}$-$\delta_{NI}$ = 0, $Z$ in (6) follows approximately the normal distribution for large $n$:

$$Z \sim N\left(\frac{\sqrt{n\varepsilon}}{\sigma}, 1\right) \tag{9}$$

where, $\varepsilon = E(\hat{\varepsilon})$ and $\sigma = E(\hat{\sigma})$ can be obtained by replacing $\hat{p}_{ij}$ with $p_{ij}$ ($i$ = 0, 1; $j$ = 0, 1) in the corresponding expression (7).

The rejection rule is specified as follows (assuming a larger $\theta$ is preferred):

$$\begin{cases} \text{Re}\,ject\ H_o\ if\ Z \geq z_{1-\alpha} \\ Accept\ H_o\ otherwise \end{cases} \tag{10}$$

Equivalently, we can use the confidence interval of $\varepsilon$:

$$1 - \beta = \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma} - z_{1-\alpha}\right) \tag{11}$$

The power of the test statistic $T$ under a particular $H_a$ can be expressed as:

$$1 - \beta = \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma} - z_{1-\alpha}\right) \tag{12}$$

where, $\varepsilon$ and $\sigma$ are estimated by (7).

Solving (12) for the sample size, we obtain:

$$n = \begin{cases} \dfrac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2 \sigma^2}{\varepsilon^2}, & for\ \varepsilon > 0 \\ \infty, & for \varepsilon > 0 \end{cases} \tag{13}$$

Equation 13 is a general sample size formulation for a trial with a normal, binary, or survival endpoint (Chang, 2007a).

For the test statistic given by (9), the p-value is given by:

$$p = 1 - \Phi\left(\frac{\hat{\varepsilon}\sqrt{n}}{\hat{\sigma}}\right) \tag{14}$$

where, $\Phi$ is the standard normal cdf.

### *Remark*

A common misconception is that for an NI trial the sample size calculation must assume $\theta_T = \theta_C$ or $p_{01} = p_{10}$, which is not true at all.

One can choose an NI design because the difference $\theta_I$-$\theta_C$ is positive but too small for a superiority test with reasonable power or unreasonably large sample size. The treatment difference can be positive or negative depending on the particular situation. The power and sample size calculation should be based on the best knowledge about the value of $\theta_I$-$\theta_C$ and this knowledge should not change because of the different choice of hypothesis test. Therefore, for a given value of $\theta_I$-$\theta_C$ and power, superiority testing always requires a larger sample size than noninferiority testing.

### *Adaptive Design*

We now discuss how to incorporate Nam's formulation (Nam, 1997) into group sequential and adaptive designs. Let $T_k$ be a test statistic on p-value scale at the $k$th stage. The stopping rules are given by:

$$\begin{cases} Stop\ for\ efficacy & if\ T_k \leq \alpha_k \\ Stop\ for\ futility & if\ T_k > \beta_k \\ Continue\ with\ adaptations & if\ \alpha_k < T_k \leq \beta_k \end{cases} \tag{15}$$

where, $\alpha_k$<$\beta_k$ ($k$ = 1, ..., $K$-1) and $\alpha_K = \beta_K$. For convenience, $\alpha_k$ and $\beta_k$ are called the efficacy and futility boundaries, respectively. The adoptions can be changes in the timing and the number of interim analyses, sample-size re-estimation, etc.

To reach the $k$th stage, a trial has to pass the 1st to ($k$-1)$^{th}$ stages.

Therefore the c.d.f. of $T_k$ is given by:

$$\begin{aligned} \psi_k(t) &= \Pr\left(\alpha_1 < T_1 < \beta_1, ..., \alpha_{k-1} < T_{k-1} < \beta_{k-1}, T_k < t\right) \\ &= \int_{\alpha_1}^{\beta_1} ... \int_{\alpha_{k-1}}^{\beta_{k-1}} \int_{\infty}^{t} f_{T_1 ... T_k}\, dt_k dt_{k-1} ... dt_1 \end{aligned} \tag{16}$$

where, $f_{T_1 \ldots T_k}$ is the joint p.d.f. of $T_1$, ... and $T_k$.

In a classic group sequential or adaptive design, the test statistic on p-value scale can be expressed as (Chang, 2007b):

$$T_k = 1 - \Phi\left(\sum_{i=1}^{k} w_{ki}\Phi^{-1}\left(1-p_i\right)\right) \tag{17}$$

where the weights $w_{ki}$ satisfy $\sum_{i=1}^{k} w_{ki}^2 = 1$. The weights $w_{ki}$ can be functions of the information time or sample-size fraction. For the error-spending approach, $w_{ki} = \sqrt{\frac{n_i}{\sum_{j=1}^{k} n_j}}, i = 1..,k, k = 1,...,K,$ where $n_i$ is the sub-sample size (not cumulative sample size) at stage $i$. Using the error-spending approach, the changes in the timing (information time) of the interim analyses and the total number of analyses can be changed after the initiation of the trial as long as the change is independent of treatment difference. For sample size re-estimation, we use fixed weights $w_{ki}$, i.e., the weight will not change even when the sample size is modified.

Two other commonly used test statistics for adaptive designs are the product of stagewise p-values and the linear combination of stagewise p-values:

$$T_k = \prod_{i=1}^{k} p_i \tag{18}$$

And:

$$T_k = \frac{1}{k}\sum_{i=1}^{k} p_i \tag{19}$$

The two-stage design stopping boundaries for (15) can be calculated using numerical integration or simulation, whereas the stopping boundaries for (18) and (19) can be analytically obtained for two-stage designs. Specifically, for the test statistic defined by (18), after choosing the efficacy stopping boundary $\alpha_1$ and futility stopping ($\beta_1 = 1$), the efficacy stopping boundary for the 2nd stage is given by (Chang, 2007b):

$$\alpha_2 = \frac{\alpha_1 - \alpha}{\ln \alpha_1} \tag{20}$$

Similarly, for the test statistic defined by (19), the stopping boundary is given by:

$$\alpha_2 = \frac{\sqrt{2\left(\alpha - \alpha_1\right)} + \alpha_1}{2} \tag{21}$$

For the error-spending approach numerical integrations gives the OF-like boundary, Pocock-like boundary and power-function boundary (with $\rho = 0.2$) as follows: $\alpha_1$ = 0.00260 and $\alpha_2$ = 0.0240 (OF), $\alpha_1$ = 0.0147 and $\alpha_2$ = 0.0147 (Pocock) and $\alpha_1$ = 0.00625 and $\alpha_2$ = 0.02173 (PF). These stopping boundaries will be used later in our trial example.

If we use Nam's test statistic defined by (6)-(8) for the subsumple at the $i$th stage, we then can calculate the "stagewise" p-value for the ith stage based on (14), that is:

$$p_i = 1 - \Phi\left(\frac{\hat{\varepsilon}_i\sqrt{n_i}}{\hat{\sigma}_i}\right) \tag{22}$$

where, $\hat{\varepsilon}_i, n_i$ and $\hat{\sigma}_i$ are the corresponding quantities in (14) but calculated based on a subsample at the $i$th stage. (22) is valid as long as $n_i$ is large.

## Conditional Power and Sample-Size Re-estimation

The general expression of conditional power at the interim analysis for a two stage adaptive design can be written as (Chang, 2007b):

$$cP_\delta\left(p_1\right) = 1 - \Phi\left(B\left(\alpha_2, p_1\right) - \frac{\hat{\varepsilon}_1\sqrt{2_n}}{\hat{\sigma}_1}\right), \alpha_1 < p_1 \le \beta_1 \tag{23}$$

Where:

$$B\left(\alpha_2, p_1\right) = \begin{cases} \dfrac{\Phi^{-1}\left(1-\alpha_2\right) - w_1\Phi^{-1}\left(1-p_1\right)}{w_2} & for\ the\ test\ statistic\ (17) \\ \Phi^{-1}\left(1-\dfrac{\alpha_2}{p_1}\right) & for\ the\ test\ statistic\ (18) \\ \Phi^{-1}\left(1-\max\left(0, \dfrac{\alpha_2}{2} - p_1\right)\right) & for\ the\ test\ statistic\ (19) \end{cases}$$

If the trial continues, i.e., $\alpha_1 < p_1 \le \beta_1$, for a given conditional power $cP$, we can solve (23) for the adjusted sample-size for the second stage:

$$n_2 = \begin{cases} \dfrac{\sigma_1^2}{\varepsilon_1^2}\left(B\left(\alpha_2, z_1\right) - \Phi^{-1}\left(1-cP\right)\right)^2, if\ \varepsilon > 0 \\ \infty, if\ \varepsilon \le 0 \end{cases} \tag{24}$$

## Type-I Error Control

We have used an approximation of the normal distribution for $z$ given by (6) and (22) for the classic and adaptive designs, respectively. We want to check how

well such approximations work in terms of type-I error control. Various scenarios have been checked with 1,000,000 simulation runs for each scenario. The scenarios with larger type-I errors are presented in Table 2 (sample size = 3000 pairs). For a classic design, we use 3000 pairs. For an adaptive design with sample size re-estimation, we use 1500 pairs for the interim analysis and the maximum sample size allowed is $N_{max}$ = 6000. We can see from the table that type-I error is well controlled when the proportion $p_{10} \geq 2\%$. When $p_{10} < 2\%$, there is a slight inflation of the error.

When we run the same set of simulations with a smaller sample size of 300 pairs and $N_{max}$ = 600 pairs, the type-I error is far below 2.5% for all cases. For $p_{10} \geq 2\%$, smaller sample sizes give smaller error but the difference is small; for $p_{10} < 2\%$, the error is much smaller than 2.5% with 300 pairs. Therefore, we can say the method can be applied to NI adaptive designs.

## Trial Example

### *Preliminary Data for Trial Design*

The adaptive design considerations will be oriented toward comparisons of the diagnostic performance of two scanning methods, separately for sensitivity (using data from positive patients) and specificity (using data from negative patients).

The two methods (Method 1 is a good standard) for the detection of metastatic disease in a group of subjects with known prostate cancer use standardized clinical end-points of documented disease including clinical outcome, serial PSA levels, contrast enhanced CT scans and radionuclide bone scans. A small study was conducted on a group of matched patients. The sensitivities are 63 and 84% for method 1 and method 2, respectively. The specificity is 80% for both methods.

The patients per CT/bone scan data are presented in Table 3.

### *The Effectiveness Requirements*

The requirements for gaining the regulatory approval are defined as follows:

- Superiority on sensitivity with 10% margin (point estimate) and NI on specificity with 7.5% margin (CI); the hypothesis testing is based on the results from 2 out of 3 image readers
- Statistical methods: McNemar's test with and without cluster adjustment. However, since we don't have data about the cluster, our sample size calculation will be based on testing without considering clustering

The effectiveness claim will be based primarily on subject level results, that is, a diagnosis of whether or not the patient has any evidence of metastatic prostate cancer, disregarding the number of sites of disease. The analyses of lesions will provide additional information on the ability of the diagnostic tests to determine localization and staging of the disease. For this reason, the sample size will be based on analysis results on the subject level. It is required that Method 2 has at least a 10% improvement (based on a point estimate) over Method 1 in sensitivity and is non-inferior to 1 in specificity with a margin of 7.5%.

### *Design for Sensitivity*

For the sensitivity requirement, we use group sequential design to handle the uncertain information with high power 95%. The simulation is done by setting the noninferiority margin to zero in the SAS program in the appendix, which was also verified using the commercial software package ExpDesign Studio 5.0.

Table 2. Type-I error rate control (%) against $\alpha = 2.5\%$

| | Proportion p10 (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Design | 0.5 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10 | 20 | 30 | 50 |
| Classic Sup | 2.9 | 2.6 | 2.4 | 2.4 | 2.3 | 2.3 | 2.0 | 1.4 | 0.9 | 0.3 |
| Classic NI | 2.6 | 2.6 | 2.4 | 2.4 | 2.3 | 2.2 | 1.9 | 1.2 | 0.7 | 0.1 |
| GSD Sup | 2.9 | 2.7 | 2.5 | 2.4 | 2.3 | 2.2 | 1.9 | 1.4 | 0.9 | 0.3 |
| SSR Sup | 2.8 | 2.6 | 2.4 | 2.4 | 2.3 | 2.2 | 1.9 | 1.4 | 0.9 | 0.3 |
| SSR NI | 2.7 | 2.5 | 2.4 | 2.3 | 2.2 | 2.2 | 1.8 | 1.7 | 0.7 | 0.1 |

Note: N = 3000, $N_{max}$ = 6000, NI margin $\delta_{NI} = -0.5p_{10}$.
For superiority design, $p_{10} = p_{01}$. SSR = Sample Size Re-estimation

Table 3. CT/Bone scan data

| Positive patients | | | Negative patients | | |
|---|---|---|---|---|---|
| | Method 1 | | | Method 1 | |
| Method 2 | Positive | Negative | Method 2 | Nagative | Positive |
| Positive | 62% | 20% | Negative | 60% | 10% |
| Negative | 3% | 15% | Positive | 10% | 20% |

For the purpose of comparison, we first calculate the sample size required for the classical design. Given the data in Table 3, i.e., $p_{10} = 0.2$ and $p_{01} = 0.03$, for a 95% power at a level of significance 2.5% (one-sided), 82 pairs are required based on McNemar's test with data provided in Table 3.

For group sequential designs (GSD), three different error-spending functions are considered: (1) The O'Brien-Fleming-like error-spending function (OF), (2) the power-function with $\rho = 2$ (PF) and (3) the Pocock-like errors pending function (Pocock).

Given the data in Table 3 and a 95% power, we design the group sequential trial with one interim analysis at 50% information time. The simulation results are presented in Table 4. To choose an "optimal" design, we perform the following comparisons:

- Comparing the results from the OF and the PF designs, we can see that the latter requires a smaller expected sample size ($\bar{N}_a$), a 7.5% reduction (73 versus 67.5 pairs) because the PF design has a larger Early Efficacy stopping Probability (EEP = 0.429) than the OF design (EEP = 0.263). The maximum sample size is almost the same for the two designs. Therefore, the PF design with $\rho = 2$ is a better design than the OF design
- Comparing the results from the Pocock and PF designs, we can see that the latter requires a smaller maximum sample-size (86 versus 92) and a smaller expected sample-size (67.5 versus 63.2). We further compare the sample sizes required under other conditions, such as $H_o$
- Under $H_o$: $p_{10} = p_{01} = 0.2$, the expected sample sizes are 65.3, 66.7 and 71 pairs for the OF, the PF and the Pocock designs, respectively. The expected sample sizes under Ho are thus similar for the OF and PF designs while being smaller than that for the Pocock design. The Early Futility stopping Probabilities (EFP) are almost identical, i.e., 45% for all three designs, which deviates from the theoretical value 50% due to approximation in normality. Based on these comparisons, we believe the design with PF ($\rho = 2$) is the best design among the three. The design can save about 18% in the expected sample size from the classical design (67 versus 82 pairs)

## *Design for Specificity*

For specificity, due to large uncertainty in the information (rates in Table 3), our design starts with a lower power 85%, then uses sample-size re-estimation at interim with 50% information time and the targeted conditional power 90%.

Like the GSD for sensitivity, we start with a classical design for specificity. Given the data in Table 3, i.e., $p_{10} = 0.1$ and $p_{01} = 0.1$, the calculation indicates that 322 pairs are required for an 85% power at a level of significance 2.5% (one-sided) based on Nam's test (1997) and the sample size calculation method presented earlier.

We use the same three error-spending functions for the adaptive trial for specificity: (1) OF, (2) PF with $\rho = 2$ and (3) the Pocock. All designs have two stages and the interim analysis will be performed at 50% information time with a sample size of 161 pairs. The sample size adjustment is based on a targeted conditional power of 90% and the maximum sample size $N_{\max}$ is 500 pairs. In all designs we use the futility boundary $\alpha_1 = 0.5$ which means approximately that if at interim analysis we observe $\hat{p}_{10} - \hat{p}_{01} - \delta_{NI} \le 0$, we will stop the trial for futility. The simulation results are presented in Table 5, where EEP and $\bar{N}_a$ are the early efficacy stopping probability and expected sample size, respectively, when $H_a$ ($p_{10} = p_{01} = 0.1$) is true.

Table 4. Operating Characteristics of AD Under Ha for Sensitivity

| | $\alpha_1$ | $\alpha_2$ | EEP | Power | $\bar{N}_a$ | EFP | $\bar{N}_O$ | $N_{\max}$ |
|---|---|---|---|---|---|---|---|---|
| OF | 0.00260 | 0.02400 | 0.263 | 0.95 | 73.0 | 0.45 | 65.3 | 84 |
| PF | 0.00625 | 0.02173 | 0.429 | 0.95 | 67.5 | 0.45 | 66.7 | 86 |
| Pocock | 0.01470 | 0.01470 | 0.625 | 0.95 | 63.2 | 0.45 | 71.0 | 92 |

Note: $\beta_1 = 0.5$, the proportions of shifting: $p_{10} = 0.2$, $p_{01} = 0.03$

Table 5. Operating characteristics of adaptive design for specificity

| | $\alpha_1$ | $\alpha_2$ | $N_{\max}$ | EEP | Power | $\bar{N}_a$ | EFP | $\bar{N}_o$ |
|---|---|---|---|---|---|---|---|---|
| OF | 0.00260 | 0.02400 | 500 | 0.206 | 0.942 | 354 | 0.47 | 335 |
| PF | 0.00625 | 0.02173 | 500 | 0.328 | 0.940 | 336 | 0.47 | 335 |
| Pocock | 0.01470 | 0.01470 | 500 | 0.454 | 0.931 | 324 | 0.47 | 335 |

Table 6. Power Preserved by GSD and SSR Designs for Specificity

| Boundary | Design | $\alpha_1$ | $\alpha_2$ | $N_{max}$ | $\bar{N}$ | Power |
|---|---|---|---|---|---|---|
| OF | GSD | 0.00260 | 0.02400 | 322 | 299 | 0.716 |
| OF | SSR | 0.00260 | 0.02400 | 500 | 386 | 0.847 |
| PF | GSD | 0.00625 | 0.02173 | 322 | 284 | 0.705 |
| PF | SSR | 0.00625 | 0.02173 | 500 | 374 | 0.842 |
| Pocock | GSD | 0.01470 | 0.01470 | 322 | 266 | 0.659 |
| Pocock | SSR | 0.01470 | 0.01470 | 500 | 364 | 0.814 |

The simulation results are summarized in Table 5. Following the same steps for comparing different adaptive designs in sensitivity, we find the PF design is better than the OF design. To evaluate the PF design against the Pocock design, we need to perform the simulations under $H_o$: $p_{10}$-$p_{01}$-$\delta_{NI}$ = 0 ($p_{10}$ = 0.1, $p_{01}$ = 0.175 and $\delta_{NI}$ = 0.075). Under this null hypothesis, the OF, PF and Pocock designs have almost the same expected sample size ($\bar{N}_o$) 335 with futility stopping probability 47%. This is because they use the same futility boundary and same sample size at the interim analysis, while the efficacy stopping boundary has virtually no effect on sample size.

We also studied the effect of SSR. We assume there is a small difference in proportions but within the noninferiority margin: $p_{10}$ = 0.1, $p_{01}$ = 0.11. We want to know if the power is reasonably preserved in this case.

The simulation results (Table 6) show that that GSD cannot well preserve power in this case. The effect of sample size adjustment on power is higher for the OB and FP designs than the Pocock designs because the OB and PF designs spend more alpha on stage 2. The Pocock design has already spent 50% alpha before the interim analysis; therefore, the sample-size adjustment at stage 2 has less effect on the power. Compared with the OB design with SSR, the PF design with SSR has a smaller expected sample size $\bar{N}_S$ (374 versus 386).

We noticed that the expected sample size under $H_o$ is high even when the null hypothesis is true. Therefore, we ran simulations with an aggressive futility boundary $\beta_1$ = 0.25 (less than original 0.5). The sample size under $H_o$ reduces from 335 to 265. However, the reduction is at the cost of power: The power is reduced from 84 to 79% when $p_{10}$ = 0.1, $p_{01}$ = 0.11. Therefore we still recommend using $\beta_1$ = 0.5, which means that if at interim analysis the observed difference is at the non-inferiority margin, we will stop the trial for futility.

Through these comparisons, we can conclude that the PF design with SSR is most preferable for the specificity design.

### *Summary of Design*

For sensitivity, totally 86 positive patients with one interim analysis will provide 95% power for the superiority test. The error-spending function for the stopping boundary is $\alpha t^2$, where t is information time or sample-size fraction and the futility stopping rule is $p_1$>$\beta_1$ = 0.5. The design features a 43% early efficacy stopping probability if the alternative hypothesis is true, a 45% early futility stopping probability if the null hypothesis is true. The expected sample size is 68 and 67 under Ha and Ho, respectively, an 18% savings in comparison to 82 pairs for the classic design.

For specificity, we use the two-stage design, featuring sample size re-estimation at interim analysis with 161 pairs. The sample size re-estimation will be based on a 90% conditional power with a cap of 500 pairs. The two-stage adaptive design has 94% power for the non-inferiority test with an NI margin of 7.5%. The error-spending function for the stopping boundary is $\alpha t^2$, where t is information time and the futility stopping rule is $p_1$>$\beta_1$ = 0.5.

The design features a 33% early efficacy stopping probability when the alternative hypothesis is true, a 47% early futility stopping probability if the null hypothesis is true. The expected sample size is 336 and 335 under $H_a$ and $H_o$, respectively, a 23% savings as compared to the classical design ($N$ = 438) with the same 94% power.

Given a 95% power for the sensitivity test and a 94% power for the specificity test, which are assumed to be independent, the overall probability of getting an effectiveness claim for the diagnosis test (Method 2) is about 90%.

The stopping rules for sensitivity and specificity are the same but sample size re-estimation is allowed for the design for specificity:

If the interim p-value for the sensitivity (specificity) test is $p_1$≤0.00625, the null hypothesis for sensitivity (specificity) will be rejected. If the p-value for sensitivity (specificity) test is $p_1$>0.5, stop recruiting positive (negative) patients. If 0.5≥$p_1$>0.00625, we continue to recruit positive (negative) patients and the sample size will be reestimated for negative patients based on a 90% conditional power. At the final analysis, if the p-value for the sensitivity (specificity) is $p_1$≤0.02173, then the null hypothesis for sensitivity (specificity) will be rejected. In the end, if both null hypothesis tests for sensitivity and specificity are rejected, then the new diagnosis test (Method 2) will be claimed effective.

## Conclusion

We have developed a simple method for adaptive trial design with binary paired data. We illustrate an application of the adaptive method for an image study, in which both superiority in sensitivity and non-inferiority in specificity are required. Using the adaptive design, the savings in the expected sample size is about 20%. The method can easily be used for other cases with paired data. For convenience, we have provided the SAS program for the classical and adaptive non-inferiority/superiority designs.

## Acknowledgement

Thank Robert Pierce for reviewing and commenting on the manuscript.

## Author's Contributions

**Mark Chang:** Methodology development of this study, draft and approval of thismanuscript.

**Jing Wang:** Data analyses, interpretation of results, review, revise and approval of thismanuscript.

## Ethics

This article is original and contains unpublished material. The corresponding author confirms that all authors have read and approved the manuscriptand no ethical issues involved.

## References

Chan, I.S.F., 2002. Power and sample size determination for noninferiority trials using an exact method. J. Biopharmaceutical Stat., 12: 457-469. PMID: 12477069

Chang, M., 2007a. Multiple-arm superiority and non-inferiority designs with various endpoints. Pharmaceutical Stat., 6: 43-52. PMID: 17323311

Chang, M., 2007b. Adaptive Design Theory and Implementation using SAS and R. 1st Edn., Chapman and Hall/CRC, ISBN-10: 1584889624, pp: 440.

Chang, M., 2010. Monte Carlo Simulation for the Pharmaceutical Industry: Concepts, Algorithms and Case Studies. 1st Edn., CRC Press, ISBN-10: 1439835934, pp: 564.

CHMP, 2005. Guideline on the choice of the noninferiority margin. EMEA/EWP/2158/99. London.

Kong, L,, R.C. Kohberger and G.G. Koch, 2004. Type I error and power in noninferiority/equivalence trials with correlated multiple endpoints: An example from vaccine development trials. J. Biopharmaceutical Stat., 14: 893-907. PMID: 15587971

Liu, J.P., H.M. Hsueh, E. Hsieh and J.J. Chen, 2002. Tests for equivalence or non-inferiority for paired binary data. Stat. Med., 21: 231-245. DOI: 10.1002/sim.1012

Lu, Y. and J.A. Bean, 1995. On the sample size for one-sided equivalence of sensitivities based upon McNemar's Test. Stat. Med., 14: 1831-1839. PMID: 7481214

Lu, Y., H. Jin and H.K. Genant, 2003. On the non-inferiority of a diagnostic test based on paired observations. Stat. Med., 22: 3029-3044. DOI: 10.1002/sim.1569

Nam, J., 1997. Establishing equivalence of two treatments and sample size requirements in matched-pairs design. Biometrics, 53: 1422-1430. PMID: 9423257

Tango, T., 1998. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. Stat. Med., 17: 891-908. DOI: 10.1002/(SICI)1097-0258(19980430)17:8<891::AID-SIM780>3.0.CO;2-B

Wiens, B.L. and J.F. Heyes, 2003. Testing for interaction in studies of noninferiority. J. Biopharmaceutical Stat., 13: 103-115. DOI: 10.1081/BIP-120017729

## Appendix: SAS Program

```
/*————————————————————————————————————*/
/* Adaptive Noninferiority Design with Paired Data */
/* Ho: p10-p01-delNI <= 0 */
/* p10 and p01 are the % of disconcordant pairs */
/* Sample size: nPairs = nPairs1 + nPairs2 from stage 1 and 2 */
/* ExpN = the expected sample size (pairs) nPairs/nRuns; */
/* nRuns = number of simulation runs */
/* alpha = one-sided significance level */
/* RejPr1 and RejPr2 = Rejection probability at stage 1 and 2 */
/* Power = probability of rejecting Ho. */
/* alpha1, alpha2=, beta1 = Stopping boundaries on p-scale. */
/*————————————————————————————————————*/
%Macro McNemarAD(alpha1 = 0.0026, alpha2 = 0.024, beta1 = 1, p10 = 0.125,
p01 = 0.125, delNI = 0.1, nPairs1 = 154, nPairs2 = 154, nPairsMax = 600,
TargetcPow = 0.90, w1 = 0.707, w2 = 0.707, nRuns = 1000000);
Data RnMvars;
Retain Power1 Power2 Futile nPairs;
alpha1 = &alpha1; alpha2 = &alpha2; beta1 = &beta1;
p10 = &p10; p01 = &p01; delNI = &delNI;
nPairs1 = &nPairs1; nPairs20 = &nPairs2;
TargetcPow = &TargetcPow; nPairsMax = &nPairsMax;
```

```
nRuns = &nRuns;
w1 = &w1/(&w1**2+&w2**2)**0.5;
w2=&w2/(&w1**2+&w2**2)**0.5;
Power1 = 0; Power2 = 0; Futile = 0; nPairs = 0;
Do iRun = 1 To nRuns;
nPairs = nPairs+nPairs1;
n10Stg1 = RANBIN(0,nPairs1,p10);
n01Stg1 = RANBIN(0,nPairs1,p01);
p10obsStg1 = n10Stg1/nPairs1;
p01obsStg1 = n01Stg1/nPairs1;
epsStg1 = p10obsStg1-p01obsStg1-delNI;
b = (2+p01obsStg1-p10obsStg1)*delNI-p01obsStg1-
p10obsStg1;
c = -p01obsStg1*delNI*(1-delNI);
p01Wave = (-b+sqrt(b*b-8*c))/4;
sigma2Stg1 = 2*p01Wave+delNI-delNI**2;
z1 = 0;
If sigma2Stg1 ^= 0 Then z1 =
epsStg1*Sqrt(nPairs1/sigma2Stg1);
pValue1 = 1-CDF('Normal',z1);
T1 = pValue1;
If T1<= alpha1 Then Power1 = Power1+1;
If T1>beta1 Then Futile = Futile+1;
If alpha1<T1<= beta1 Then Do;
** Sample size reestimation based on conditional power
**;
Bval = (Probit(1-alpha2)-w1*Probit(1-pValue1))/w2;
    nPairs2 = nPairs20;
    If epsStg1>0 Then
    nPairs2 = sigma2Stg1/epsStg1**2*(Bval-Probit(1-
    TargetcPow))**2;
    nPairs2 = Min(nPairsMax-nPairs1,nPairs2);
    nPairs2 = Round(Max(nPairs2, nPairs20));
    n10Stg2 = RANBIN(0,nPairs2,p10);
    n01Stg2 = RANBIN(0,nPairs2,p01);
    p10obsStg2 = n10Stg2/nPairs2;
    p01obsStg2 = n01Stg2/nPairs2;
    epsStg2 = p10obsStg2-p01obsStg2-delNI;
    b = (2+p01obsStg2-p10obsStg2)*delNI-
    p01obsStg2-p10obsStg2;
    c = -p01obsStg2*delNI*(1-delNI);
    p01Wave = (-b+sqrt(b*b-8*c))/4;
    sigma2Stg2 = 2*p01Wave+delNI-delNI**2;
    z2 = 0;
    If sigma2Stg2 ^= 0 Then z2 =
    epsStg2*Sqrt(nPairs2/sigma2Stg2);
    pValue2 = 1-CDF('Normal',z2);
    T2 = 1-CDF('NORMAL', w1*z1+w2*z2);
    If T2<= alpha2 Then Power2 = Power2+1;
    nPairs = nPairs+nPairs2;
    End;
End;
ExpN = nPairs/nRuns;
RejPr1 = Power1/nRuns;
RejPr2 = Power2/nRuns;
Power = RejPr1+RejPr2;
FutilePr = Futile/nRuns;
Output;
Run;
proc print data = RnMvars;
 var alpha1 alpha2 beta1 nPairs1 nPairs20 nPairs2 w1
w2
 TargetcPow nRuns p10 p01 FutilePr RejPr1 RejPr2
Power ExpN;
Run;
%Mend McNemarAD;
Title2 "Classic 1-Stage Design: Type-I erorr rate p10-
p01-delNI = 0";
%McNemarAD(alpha1 = 0.025, alpha2 = 0, beta1 = 0,
p10=0.1,
    p01 = 0.175, delNI = -0.075, nPairs1=322, nPairs2 =
    0);
Title2 "PF (rho=2), beta1 = 0.5 with SSR (Nmax>N0)";
%McNemarAD(alpha1 = 0.00625, alpha2=0.02173,
beta1 = 0.5, p10 = 0.1,
   p01 = 0.1, delNI = -0.075, nPairs1 = 161, nPairs2 =
161, nPairsMax = 500);
```