International Journal of Statistics and Probability; Vol. 5, No. 5; September 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Considering Both Statistical and Clinical Significance in the Analysis of Clinical Data

Guolong Zhao[1]

[1] Henan Institute of Medical Sciences, Henan Medical University, 40 University Road, Zhengzhou, Henan, 450052, China

Correspondence: Guolong Zhao, Henan Institute of Medical Sciences, Henan Medical University, 40 University Road, Zhengzhou, Henan, 450052, China. E-mail: zhaogzu@hotmail.com

Received: October 29, 2015 Accepted: November 16, 2015 Online Published: August 9, 2016

doi:10.5539/ijsp.v5n5p16 URL: http://dx.doi.org/10.5539/ijsp.v5n5p16

Abstract

To evaluate a drug, statistical significance alone is insufficient and clinical significance is also necessary. This paper explains how to analyze clinical data with considering both statistical and clinical significance. The analysis is practiced by combining a confidence interval under null hypothesis with that under non-null hypothesis. The combination conveys one of the four possible results: (i) both significant, (ii) only significant in the former, (iii) only significant in the latter or (iv) neither significant. The four results constitute a quadripartite procedure. Corresponding tests are mentioned for describing Type I error rates and power. The empirical coverage is exhibited by Monte Carlo simulations. In superiority trials, the four results are interpreted as clinical superiority, statistical superiority, non-superiority and indeterminate respectively. The interpretation is opposite in inferiority trials. The combination poses a deflated Type I error rate, a decreased power and an increased sample size. The four results may helpful for a meticulous evaluation of drugs. Of these, non-superiority is another profile of equivalence and so it can also be used to interpret equivalence. This approach may prepare a convenience for interpreting discordant cases. Nevertheless, a larger data set is usually needed. An example is taken from a real trial in naturally acquired influenza.

Keywords: clinical significance, confidence interval, equivalence trial, non-inferiority trial, non-null hypothesis, quadripartite procedure

1. Introduction

Statistical significance is measured by a p-value or a confidence interval (CI). Food and Drug Administration (FDA, 2010) recommends the use of CIs. A CI tends to be interpreted either as statistically significant (if the CI does not include zero) or not significant (if the CI includes zero) in FDA (2010). Such an interpretation is similar to that in Cox and Hinkley (1974), which involves a consideration of the relation between CIs and tests. For example, if for some estimated parameter θ one wants to test the null hypothesis that $\theta = 0$ against the alternative that $\theta \neq 0$, then this test can be performed by determining whether the CI for θ crosses zero. Thus a treatment effect is established by using a CI under certain hypotheses.

Statistical significance does not yield information about magnitude of effect (Haase et al, 1989; Greenstein, 2003). Clinical significance measures how large the differences in treatment effects are in practice. It was proposed by Jacobson et al (1984). Its definition varies with the clinical field being addressed (Lindgren et al, 1993; Greenstein, 2003; Kraemer et al, 2003). A common way is to establish a cutoff point (see, for example, Jacobson & Truax, 1991), threshold or margin (see, for example, D'Agostino Sr et al, 2003; Blackwelder, 2004; Dann & Koch, 2008; Goeman et al, 2010). The difference in excess of the margin is considered to be clinically significant.

Using the margin in place of zero forms non-null hypothesis. Perhaps this term is difficult to accept to some statisticians. It was introduced by Egon Pearson (1939) and used by Fisher (Good, 1992). Its definition is found in the book A Dictionary of Statistical Terms (Marriott, 1990)(http://stats.oecd.org /glossary /detail.asp? ID=3737). In Dunnett and Gent (1977), where both zero and non-zero appeared, they used "null" to refer zero and "non-null" to non-zero $H : \delta = \Delta$, $H^1 : \delta < \Delta$ so as to avoid confusion. Tests of non-null hypothesis have developed (see, for example, Dunnett & Gent, 1977; Mehta et al, 1984; Holmgren, 2001; Zhao, 2003; Dann & Koch, 2008; Zhao, 2008; Goeman et al, 2010; Zhao, 2015) but CIs under non-null hypothesis have not been widely noted. They may be used as measures of clinical significance.

To evaluate a drug, statistical significance alone is insufficient and clinical significance is also necessary (Laupacis et al, 1988). This idea is in line with our general intuition. In drug development, one may imagine two possible conclusions. Conclusion 1, new drug is more effective than old one or 2, new drug is as effective as old one. Conclusion 1 is usually drawn from statistical significance but 2 involves clinical significance. Hence we need CIs under non-null hypothesis. Clinical significance is a subjective evaluation and before a finding can be clinically significant, it must have achieved statistical significance, which is indicated by Killoy (2002), among others. Put another way, in dealing with clinical

significance, statistical significance is already assessed. This indication claims that a CI under non-null hypothesis is actually combined with a CI under null hypothesis. Lu and Fang (2003) show that the combination conveys one of four possible outcomes. It is not yet clear, however, how the four outcomes are set up, how they are interpreted, what benefits they have in practice and what influences they have on the size and power.

This paper explains how to analyze clinical data with considering both statistical and clinical significance. This will be a two-part endeavor - first it will be necessary to define the combination of a CI under null hypothesis with a CI under non-null hypothesis; then we will look for the process that generates the four results. Corresponding tests will be mentioned for dealing with Type I error rates and power. Monte Carlo simulation will be taken up to exhibit the empirical coverage of CIs. It does not involve any new statistical principle but offers a wider field of vision on the analysis of clinical data.

2. Confidence Intervals

2.1 A Brief Overview of Confidence Intervals

CIs are usually used under null hypothesis. For simplicity, the two-proportion CI is taken as an example. CIs in other distributions may be deduced by analogy.

Consider a model in which Y_j, $j = 1, 2$, represents two independent binomial observations with sample sizes n_j and success probability π_j. The total sample size is $N = n_1 + n_2$ with the sample fractions $k_j = n_j/N$. Writing $Y_j = y_j$, the proportion is $\hat{p}_j = y_j/n_j$ with the treatment difference $D = \hat{p}_1 - \hat{p}_2$ and the average $\bar{p} = k_1\hat{p}_1 + k_2\hat{p}_2$.

Their expectations are $E[\hat{p}_j] = \pi_j$, $E[D] = \mu = \pi_1 - \pi_2$ and $E[\bar{p}] = \bar{\pi} = k_1\pi_1 + k_2\pi_2$ respectively. The variance of $\hat{p}_j$ is known to be $V[\hat{p}_j] = \pi_j(1 - \pi_j)/(k_jN)$. The variance of D for two independent groups is given by $V[D] = \sum_j V[\hat{p}_j]$. We denote it briefly by $V[D] = \Sigma^2 = \sigma^2/N$, where $\sigma^2 = \sum_j \pi_j(1 - \pi_j)/k_j$.

Let α be the probability of Type I error and β that of Type II. The null hypothesis is $H_0 : \mu = 0$ and its alternative $H_1 : \mu > 0$ ($\mu \in (0, 1]$), $\mu < 0$ ($\mu \in [-1, 0)$) or $\mu \neq 0$ ($\mu \in [-1, 0) \cup (0, 1]$). If H_0 is true, we have $\pi_1 = \bar{\pi} + k_2\mu$ and $\pi_2 = \bar{\pi} - k_1\mu$, where $\pi_1 - \pi_2 = 0$. The variance is $\Sigma_0^2 = \sigma_0^2/N$, $\sigma_0^2 = \bar{\pi}(1 - \bar{\pi})(1/k_1 + 1/k_2)$ under H_0 and $\Sigma_1^2 = \sigma_1^2/N$, $\sigma_1^2 = \sum_j \pi_j(1 - \pi_j)/k_j$ under H_1. The distribution of D is described along the asymptotic normal method (Fleiss, 1981) with the density $f_D(d; 0, \Sigma_0^2)$ under H_0 and $f_D(d; \mu, \Sigma_1^2)$ under H_1. Let d_α be the critical value of D at α level. We have $d_\alpha = 0 + z_\alpha\Sigma_0$ and $d_\alpha = \mu - z_\beta\Sigma_1$. It follows that $0 + z_\alpha\Sigma_0 = \mu - z_\beta\Sigma_1$ or

$$\mu - 0 = z_\alpha\sigma_0/\sqrt{N} + z_\beta\sigma_1/\sqrt{N}\,. \tag{2.1}$$

Equation (2.1) can then be used to evaluate sample size or power once σ_0, σ_1 and α have been specified. A $100(1 - \alpha)\%$ CI for $\mu = 0$ is of the form

$$D - z_{\alpha/2}S_1 \leq 0 \leq D + z_{\alpha/2}S_1 \tag{2.2}$$

where S_1 comes from $S_1^2 = s_1^2/N$, $s_1^2 = \sum_j \hat{p}_j(1 - \hat{p}_j)/k_j$ the estimator of $\Sigma_1^2 = \sigma_1^2/N$.

The corresponding test is

$$Z_0 = (D - 0)/\sqrt{s_0^2/N}, \tag{2.3}$$

$Z_0 \sim N(0, 1]$ (Fleiss, 1981) where $s_0^2 = \bar{p}(1 - \bar{p})(1/k_1 + 1/k_2)$ is the estimator of σ_0^2. They are just the two-proportion CI and the Z-test.

2.2 Confidence Intervals under Non-Null Hypothesis

We are then led to consider the use of a CI under non-null hypothesis. Suppose Δ is the difference margin. It is called non-inferiority margin, equivalence margin and so on in the definite field being addressed (see, for example, D'Agostino Sr et al, 2003; Blackwelder, 2004; Dann & Koch, 2008; Goeman et al, 2010). Since the margin is a subjective concept (Killoy, 2002; Scott, 2009), there are some explanatory notes:

(i) The valuation of Δ is based on professional knowledge, references, academic conferences, prior experience, past trials, or pilot studies, etc. Generally a proper valuation comes from the process of repeated practices and experiences.

(ii) Also in light of Killoy (2002), the parameter μ defines the range in which the margin Δ has its being so that the sign of Δ is always kept identical to that of μ. In a superiority trial with $\mu \in [0, 1]$, for instance, we appoint $\Delta \geq 0$.

The non-null hypothesis is $H : \mu = \Delta$. Its alternative is $H_a : \mu > \Delta$ ($\mu \in [\Delta, 1]$) or $\mu < \Delta$ ($\mu \in [0, \Delta)$) in $\mu \in [0, 1]$ with $\Delta \geq 0$ and $H_a : \mu < \Delta$ ($\mu \in [-1, \Delta]$) or $\mu > \Delta$ ($\mu \in [\Delta, 0]$) in $\mu \in [-1, 0]$ with $\Delta \leq 0$. If H is true, we have $\pi_1^* = \bar{\pi} + k_2\Delta$ and $\pi_2^* = \bar{\pi} - k_1\Delta$, where $\pi_1^* - \pi_2^* = \Delta$. The variance is $\Sigma^{*2} = \sigma^{*2}/N$, $\sigma^{*2} = \sum_j \pi_j^*(1 - \pi_j^*)/k_j$ under H and $\Sigma_a^2 = \sigma_a^2/N$, $\sigma_a^2 = \sum_j \pi_j(1 - \pi_j)/k_j$ under H_a. Using $\mu - \Delta$, σ^{*2} and σ_a^2 in place of $\mu - 0$, σ_0^2 and σ_1^2 respectively, (2.1) becomes

$$\mu - \Delta = z_\alpha\sigma^*/\sqrt{N} + z_\beta\sigma_a/\sqrt{N}\,. \tag{2.4}$$

Let the variance Σ^{*2} be estimated as $S^{*2} = s^{*2}/N$, $s^{*2} = \sum_j \hat{p}_j^*(1-\hat{p}_j^*)/k_j$, where $\hat{p}_1^* = \bar{p} + k_2\Delta$ and $\hat{p}_2^* = \bar{p} - k_1\Delta$ and Σ_a^2 as $S_a^2 = s_a^2/N$, $s_a^2 = \sum_j \hat{p}_j(1-\hat{p}_j)/k_j$. Here a $100(1-\alpha)\%$ CI for $\mu = \Delta$ is given by

$$D - z_{\alpha/2}S_a \le \Delta \le D + z_{\alpha/2}S_a \tag{2.5}$$

and the statistic by

$$Z_\Delta = (D-\Delta)\Big/\sqrt{s^{*2}/N} \tag{2.6}$$

(Dunnett & Gent, 1977; Zhao, 2003). The denominator of (2.6) can be replaced by $\sqrt{N^{-1}\bar{p}(1-\bar{p})(1/k_1 + 1/k_2)}$ with trivial divergence in $n_1 \approx n_2$ (Hirotsu, 2007).

When $\Delta = 0$, (2.4), (2.5) and (2.6) reduce to their classical counterparts (2.1), (2.2) and (2.3) respectively.

2.3 One-Sided Confidence Intervals

The CI (2.5) is in two-sided (2-sided) form and serves a 2-sided alternative. With non-null hypothesis, however, it is seldom to use 2-sided alternative, where one-sided (1-sided) alternatives are used. Hence we most often need 1-sided CIs.

In a superiority trial, a 95% R-sided CI for $\mu = 0$ (C1) is used for statistical superiority testing, a 95% R-sided CI for $\mu = \Delta$ (C2) for clinical superiority testing and a 97.5% L-sided CI for $\mu = \Delta$ (C3) for non-superiority testing. A better view is shown in the array as below. Notice that CIs are distinct in three aspects such as 95 and 97.5%, L-, R- and

	Clinical trials $\mu \in [-1,1]$ $\alpha = 0.05$ $H_0 : \mu = 0$ $H_1 : \mu \ne 0$ $D - z_{\alpha/2}S_1 \le 0 \le D + z_{\alpha/2}S_1$ (2.2)		
Statistical inferiority $\mu \in [-1,0]$ $\alpha = 0.05$ $H_0 : \mu = 0$ $H_1 : \mu < 0$ $-1 \le 0 \le D + z_\alpha S_1$ (C4)		Statistical superiority $\mu \in [0,1]$ $\alpha = 0.05$ $H_0 : \mu = 0$ $H_1 : \mu > 0$ $D - z_\alpha S_1 \le 0 \le 1$ (C1)	
Clinical inferiority $\mu \in [-1,\Delta]$ $\alpha = 0.05$ $H : \mu \ge \Delta$ $H_a : \mu < \Delta$ $-1 \le \Delta \le D + z_\alpha S_a$ (C5)	Non-inferiority $\mu \in [\Delta,0]$ $\alpha = 0.025$ $H : \mu \le \Delta$ $H_a : \mu > \Delta$ $D - z_\alpha S_a \le \Delta \le 0$ (C6)	Non-superiority $\mu \in [0,\Delta]$ $\alpha = 0.025$ $H : \mu \ge \Delta$ $H_a : \mu < \Delta$ $0 \le \Delta \le D + z_\alpha S_a$ (C3)	Clinical superiority $\mu \in [\Delta,1]$ $\alpha = 0.05$ $H : \mu \le \Delta$ $H_a : \mu > \Delta$ $D - z_\alpha S_a \le \Delta \le 1$ (C2)

2-sided and $\mu = 0$ (null hypothesis) and $\mu = \Delta$ (non-null hypothesis). The choice of CIs depends upon the alternatives.

According to FDA (2010) or International Guideline ICH E9 Hirotsu (2007), we take $\alpha = 0.025$ for non-superiority, i.e., the confidence level 97.5% for (C3). However, 95% is more often used in other situations. One must carefully select a CI according to the correspondence in the array. It is not widely appreciated that using a 2-sided CI in place of two 1-sided CIs is prone errors, which will be discussed with an example in Subsection 6.3.

Analogously, the array also lists CIs for a inferiority trial. As for a significance trial, it is similar to superiority trials except that the absolute value $|\mu| \in [0,1]$ with $|\Delta| \ge 0$ is used instead, which is given in another array. The null hypothesis

	Statistical significance $\lvert\mu\rvert \in [0,1]$ $\alpha = 0.05$ $H_0 : \lvert\mu\rvert = 0$ $H_1 : \lvert\mu\rvert > 0$ $D - z_{\alpha/2}S_1 \le 0 \le D + z_{\alpha/2}S_1$ (2.2)	
Equivalence $\lvert\mu\rvert \in [0,\lvert\Delta\rvert]$ $\alpha = 0.025$ $H : \lvert\mu\rvert \ge \lvert\Delta\rvert$ $H_a : \lvert\mu\rvert < \lvert\Delta\rvert$ $0 \le \lvert\Delta\rvert \le \lvert D\rvert + z_\alpha S_a$ (C8)		Clinical significance $\lvert\mu\rvert \in [\lvert\Delta\rvert,1]$ $\alpha = 0.05$ $H : \lvert\mu\rvert \le \lvert\Delta\rvert$ $H_a : \lvert\mu\rvert > \lvert\Delta\rvert$ $\lvert D\rvert - z_\alpha S_a \le \lvert\Delta\rvert \le 1$ (C7)

is $H_0 : |\mu| = 0$ and its alternative $H_1 : |\mu| > 0$ ($|\mu| \in (0,1]$) for statistical significance testing. The alternative here looks

1-sided but it is identical to the 2-sided form $H_1 : \mu \neq 0$ ($\mu \in [-1, 0) \cup (0, 1]$) so that (2.2) is used. The array also gives CIs for clinical significance and equivalence.

Relations among trials are self-explanatory in the arrays. Superiority and inferiority trials are the two symmetric profiles of significance trials. Superiority and inferiority trials convert to each other as long as treatment and control groups reverse roles. If regarding μ- and Δ-values as absolute, either superiority or inferiority trials convert to significance trials. There is a one-to-one correspondence among the results. Specifically, non-superiority and non-inferiority are the two symmetric profiles of equivalence. Both non-superiority and non-inferiority can be used to interpret equivalence. One may easily find the position of the usual non-inferiority and equivalence trials in the arrays.

3. The Analysis of Clinical Data

The analysis with considering both statistical and clinical significance is usually conducted on a larger data set with a pre-stated Δ-value. If there is no such a value, one may negotiate with investigators to appoint one.

In a superiority trial with $\mu \in [0, 1]$, the favorite outcome is $D \geq 0$. As a desirable course, this is the main scene in this article since such a trial gives an interpretation straightforward. One may unexpectedly see a discordant case that outcome is contrary to plan. For example, a superiority trial with positive μ-value generates an opposite result with negative D-value, which was elaborated by Blackwelder (2004), among others. Discordant cases will not be included in simulations (Section 4) and will be discussed in Subsection 6.1.

3.1 Patterns of the Combination

A total of six possible patterns (not including discordant cases) are built up on the combinations of a CI under null hypothesis with a CI under non-null hypothesis as shown Figure 1. The abscissa represents the difference μ and the ordinate crosses the abscissa at the origin $\mu = 0$. The graph is thus divided into two halves by the ordinate. The right half belongs to the scope of superiority trials.

In superiority trials, Pattern 1 shows that (C1) does not include zero and (C2) lies outside Δ. Both statistical and clinical superiority are demonstrated. The result is represented by clinical superiority, which implies statistical superiority yet. In Pattern 2, (C2) crosses Δ and only statistical superiority is demonstrated.

Pattern 3 presents that lower bound of (C1) is greater than zero so that statistical superiority is demonstrated. Nonetheless, upper bound of (C3) is less than Δ so that non-superiority is also demonstrated for the margin Δ. It seems to present interpretive problems (FDA, 2010). The result is categorized as non-superiority after a consideration that the difference less than the margin is not clinically significant in spite of the statistical significance (Blackwelder, 2004). Pattern 4 shows that (C3) crosses Δ and only statistical superiority is demonstrated.

In Pattern 5, (C1) crosses zero but (C3) lies outside Δ and only non-superiority is demonstrated. Pattern 6 gives that (C1) crosses zero and (C3) crosses Δ. The result is indeterminate.

Finally, the six patterns refine into the four possible results: Result 1 clinical superiority, Result 2 statistical superiority, Result 3 non-superiority and Result 4 indeterminate.

The patterns in inferiority trials are listed in the left half of Figure 1. As for significance trials, the patterns are similar to those in superiority trials.

3.2 Calculation Process

A set of clinical data is summarized as the difference D with the expectation μ and the margin Δ. First of all, one may see $D \geq 0$ or $D < 0$. In a superiority trial ($\mu \in [0, 1]$ with $\Delta \geq 0$), $D < 0$ is a discordant case.

The analysis begins with testing $H_0 : \mu = 0$ versus $H_1 : \mu > 0$ at the confidence level 95% using (C1) for $\mu = 0$. If zero lies outside (C1), H_0 is rejected. If $D \geq \Delta$, the analysis will go to testing $H : \mu \leq \Delta$ versus $H_a : \mu > \Delta$ at 95% using (C2) for $\mu = \Delta$. If (C2) does not include Δ, H is rejected too. It comes to Result 1 based on Pattern 1, which is in favor of such an interpretation that test drug has clinical superiority over control. If (C2) crosses Δ, H is not rejected. We get Result 2 from Pattern 2. The interpretation is that test drug is statistically superior to control.

If H_0 is not rejected, the analysis will turn to testing $H : \mu \geq \Delta$ versus $H_a : \mu < \Delta$ at 97.5% using (C3). If H is rejected with $D < \Delta$, we reach Result 3 from Pattern 5. The interpretation is that test drug is non-superior to control. If H is not rejected, we get Result 4 from Pattern 6 and the interpretation indeterminate. A flowchart of the four possible results is depicted in Figure 2. The diagram with dashed lines indicated by $D < \Delta$ in Figure 2, which derives from Pattern 3 and 4 in Figure 1, represents a rare situation. It implies a very small D-value since Δ-value is limited as its name states. Such a D-value with statistical significance comes definitely from a very large data set. Accordingly, it will not be included in simulations (Section 4).

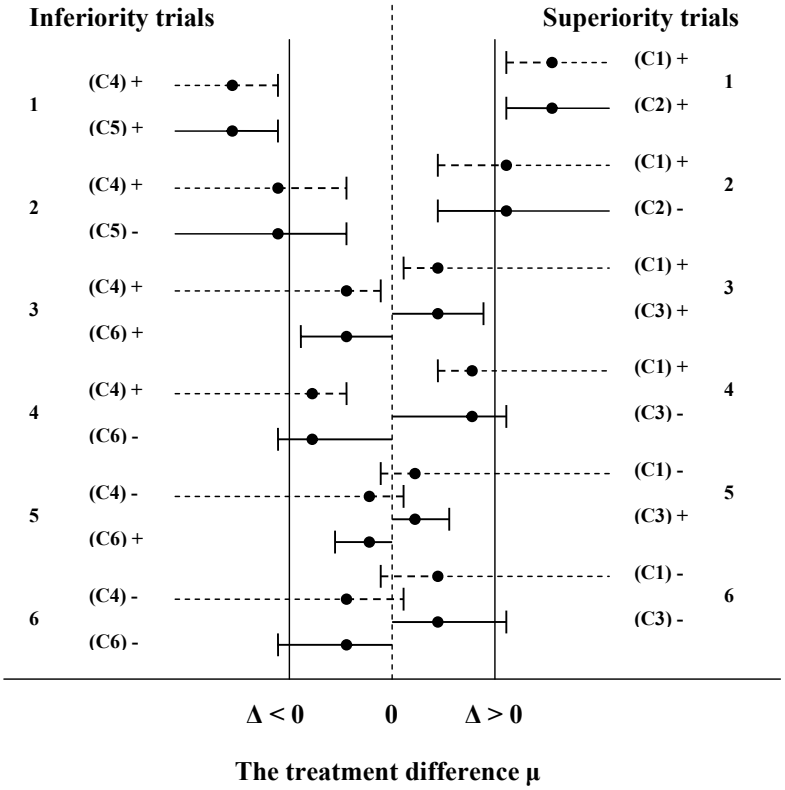

Figure 1. Patterns of combining a confidence interval under null hypothesis with that under non-null hypothesis in superiority or inferiority trials

Dots indicate the point estimate *D*, dashed lines the confidence interval for $\mu = 0$ (null hypothesis) and solid lines that for $\mu = \Delta$ (non-null hypothesis). Symbol "-" denotes that a confidence interval contains the value of parameter and "+" does not.

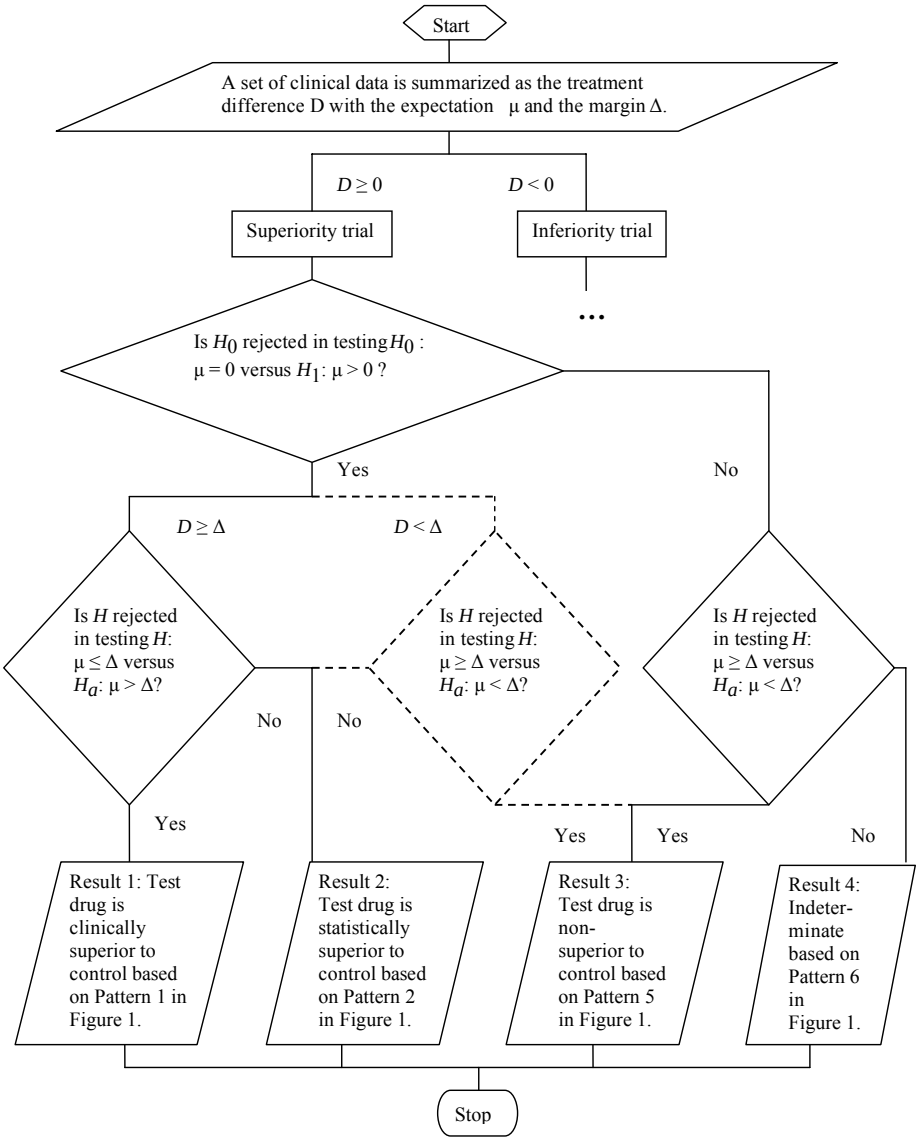

Figure 2. The process in the analysis of clinical data

The diagram with dashed lines indicated by $D < \Delta$, which derives from Pattern 3 and 4 in Figure 1, represents a rare situation. It implies a very small *D*-value since Δ-value is limited as its name states. Such a *D*-value with statistical significance comes definitely from a very large data set.

It is heuristic that a superiority trial may result in non-superiority when considering both statistical and clinical significance as shown in Figure 2. As another profile of equivalence, non-superiority can also be used to interpret equivalence.

A similar flowchart is developed in inferiority trials with the opposite interpretation, which is omitted. A significance trial is defined on $|\mu| \in [0, 1]$ with $|\Delta| \geq 0$. The interpretation for Result 1 - 3 is that the difference between test drug and control is clinically significant and statistically significant and that test drug is equivalent to control respectively. Its flow diagram is omitted.

3.3 Type I Error Rates and Power

The process has an influence on Type I error rates and further on power. For convenience, statistic (2.3) is used in place of (C1) and (2.6) in place of (C2) and (C3). Let α be valuated as α_1, α_2 and α_3 in statistical superiority, clinical superiority and non-superiority testing respectively. With a pre-chosen Δ-value, we have either $D \geq \Delta$ or $D < \Delta$. Let $\xi = I\{D \geq \Delta\}$ be an indicator. We then have $\xi = 1$ when $D \geq \Delta$ and $\xi = 0$ when $D < \Delta$.

Figure 2 indicates that a trial yields only one of the four possible results so that they are mutually exclusive and that each of them comes from a series connection of two tests. As far as Result 1 is concerned, the α-values for the two tests are α_1 and α_2 respectively. Then the Type I error rate is

$$P(1) = \alpha_1 \xi \alpha_2. \tag{3.1}$$

Likewise, we have

$$P(2) = \alpha_1 \xi (1 - \alpha_2) + \alpha_1 (1 - \xi)(1 - \alpha_3), \tag{3.2}$$

$$P(3) = \alpha_1 (1 - \xi)\alpha_3 + (1 - \alpha_1)\alpha_3 \tag{3.3}$$

and

$$P(4) = (1 - \alpha_1)(1 - \alpha_3) \tag{3.4}$$

for Result 2, 3 and 4 respectively. Finally, they come down to a perfect set

$$P_{sum} = P(1) + P(2) + P(3) + P(4) = 1. \tag{3.5}$$

As usual, the indicator shows $\xi = 1$ and we have $P(1) = \alpha_1\alpha_2$, $P(2) = \alpha_1(1-\alpha_2)$, $P(3) = (1-\alpha_1)\alpha_3$, $P(4) = (1-\alpha_1)(1-\alpha_3)$ and $P_{sum} = 1$. We then find a deflated Type I error rate marked in Result 1 and trivial in 2 and 3. Setting $\alpha_1 = \alpha_2 = 0.05$ and $\alpha_3 = 0.025$ yields $P(1) = 0.0025$, $P(2) = 0.0475$, $P(3) = 0.02375$, $P(4) = 0.92625$ and $P_{sum} = 1$.

Substituting the deflated Type I error rate into (2.1) and (2.4) will show a decrease of power and an increase of the required sample size.

If by any chance the indicator shows $\xi = 0$, i.e. the rare case with $D < \Delta$, we have $P(1) = 0$, $P(2) = \alpha_1(1-\alpha_3)$, $P(3) = \alpha_3$, $P(4) = (1 - \alpha_1)(1 - \alpha_3)$ and $P_{sum} = 1$. Also, we may observe the deflated Type I error rate, the decreased power and the increased sample size in this case.

4. Monte Carlo Simulations

A Monte Carlo simulation study was undertaken to compare the performance of the two-proportion CI under null hypothesis with that under non-null hypothesis.

4.1 Superiority Trials

Experiment 1: The empirical coverage of the two-proportion CIs under null and non-null hypothesis in superiority trials. For comparison, we took equal confidence level $1 - \alpha = 0.95$ for all 1-sided CIs. The sample size was prescribed as $N = 160$ with the sample fraction $\{k_j\} = (0.6, 0.4)$. In practice, π_2 the proportion of control group is known beforehand and a general approach is to appoint $\{\mu\}$ an predicted increase in treatment group. Here we took $\pi_2 = 0.5$ and $\{\mu\} = (0, 0.04, \cdots, 0.36)$. As for Δ, it relies on the $\{\mu\}$-value and is valuated by clinicians. Here we appointed $\Delta = 0.18$. It follows that $\pi_1 = \pi_2 + \mu$. Crossing combination of these quantities formed 10 patterns. The expected power $1 - \beta$ was calculated by referring to (2.1) and (2.4). The frequency Y_j was generated from the Bernoulli random variable in the S language and then we had the proportions $\hat{p}_j = y_j/n_j$. Pseudo-random numbers were generated from the same initial seed for all sets of simulations. A R-sided CI for $\mu = 0$ was calculated using (C1). With $\mu = \Delta = 0.18$, (C2) and (C3) were used. Also Z-values were calculated following (2.3) and (2.6). Such a process was repeated 1000 times. Then we could calculate the empirical coverage of (C1), i.e. the number of times $\mu = 0$ lies in the interval divided by 1000. Similarly, we got the empirical coverage of (C2) and (C3) for $\mu = \Delta = 0.18$. The observed power of statistic Z was $1-\hat{\beta} = \sum_1^{1000} I\{P \leq \alpha\}/1000$, where $I\{P \leq \alpha\}$ is indicator function. A 95% CI for the power is $1 - \beta \pm 1.96(\beta(1 - \beta)/1000)^{1/2}$. Table 1 lists the values of EC^c the complement of empirical coverage. The observed average difference $\bar{D}$ is close to μ the prescribed difference.

Table 1. The empirical coverage of two-proportion confidence intervals under null and non-null hypothesis in superiority trials

μ	$\bar{D}$	$\bar{p}_1$	$\bar{p}_2$	Null hypothesis			Non-null hypothesis		
				EC_0^c	$1-\hat{\beta}_0$	$1-\beta_0$	EC_Δ^c	$1-\hat{\beta}_\Delta$	$1-\beta_\Delta$
0	0.001	0.498	0.497	0.110	0.109	0.050	0.730	0.730	0.731
0.04	0.044	0.542	0.497	0.148	0.148	0.125	0.525	0.525	0.546
0.08	0.083	0.580	0.497	0.262	0.262	0.258	0.331	0.331	0.352
0.12	0.122	0.619	0.497	0.452	0.452	0.443	0.194	0.192	0.189
0.16	0.162	0.659	0.497	0.661	0.663	0.467	0.104	0.101	0.083
0.20	0.202	0.699	0.497	0.818	0.819	0.819	0.118	0.108	0.082
0.24	0.241	0.739	0.497	0.930	0.931	0.929	0.206	0.194	0.190
0.28	0.281	0.779	0.497	0.982	0.983	0.980	0.378	0.368	0.366
0.32	0.321	0.819	0.497	0.996	0.996	0.996	0.597	0.595	0.586
0.36	0.362	0.859	0.497	1	1	1	0.809	0.809	0.792

$\alpha = 0.05$ (1-sided), $N = 160$, $\{k_j\} = (0.6, 0.4)$, $\pi_2 = 0.5$, μ = the prescribed difference and $\Delta = 0.18$ the difference margin.

$\bar{D}$ = the average difference produced in simulations, $\bar{p}_1$ = the average proportion of treatment group, $\bar{p}_2$ = that of control group, EC^c = the complement of empirical coverage, $1-\hat{\beta}$ = the observed power of statistic Z produced in simulations and $1-\beta$ = the expected power.

The observed average proportion in the jth group $\bar{p}_j$ coincides with the pre-specified value of parameter π_j. As far as the null hypothesis is concerned, the value of EC_0^c increases as μ increases. With respect to the non-null hypothesis, the value of EC_Δ^c decreases as μ increases and approaches its minimum at $\mu = 0.18$ and then increases as μ increases further. The value of the observed power $1-\hat{\beta}$ is almost the same as that of EC^c. Both them are around the left or right side of the corresponding value of the expected power $1-\beta$ and lies within its 95% CI in most experimental sets. Figure 3 shows a comparison of the observed (broken lines) and expected power (smooth curves) of the two-proportion Z-test of null and non-null hypothesis. The observed power coincides with the expected power.

4.2 Inferiority Trials

Experiment 2: The empirical coverage of the two-proportion CIs under null and non-null hypothesis in inferiority trials. The method was the same as that in Experiment 1 except that the μ- and Δ-values were taken to be negative. A similar result holds for Experiment 2 as listed in Table 2. Of these, the value of $1-\beta$ is identical to that in Experiment 1 while

Table 2. The empirical coverage of two-proportion confidence intervals under null and non-null hypothesis in inferiority trials

μ	$\bar{D}$	$\bar{p}_1$	$\bar{p}_2$	Null hypothesis			Non-null hypothesis		
				EC_0^c	$1-\hat{\beta}_0$	$1-\beta_0$	EC_Δ^c	$1-\hat{\beta}_\Delta$	$1-\beta_\Delta$
0	0.001	0.498	0.497	0.110	0.109	0.050	0.733	0.733	0.731
-0.04	-0.039	0.458	0.497	0.141	0.141	0.125	0.569	0.569	0.546
-0.08	-0.078	0.420	0.497	0.253	0.252	0.258	0.378	0.376	0.352
-0.12	-0.116	0.381	0.497	0.416	0.415	0.443	0.209	0.206	0.189
-0.16	-0.157	0.341	0.497	0.616	0.617	0.467	0.122	0.119	0.083
-0.20	-0.197	0.301	0.497	0.810	0.810	0.819	0.126	0.115	0.082
-0.24	-0.236	0.261	0.497	0.914	0.917	0.929	0.205	0.193	0.190
-0.28	-0.276	0.221	0.497	0.983	0.983	0.980	0.346	0.337	0.366
-0.32	-0.316	0.181	0.497	1	1	0.996	0.560	0.556	0.586
-0.36	-0.357	0.141	0.497	1	1	1	0.785	0.784	0.792

See Table 1 for definitions but $\Delta = -0.18$.

the values of EC^c and $1-\hat{\beta}$ have slight variations. They can be shown in a graph symmetric to Figure 3 with respect to the ordinate, which is omitted here.

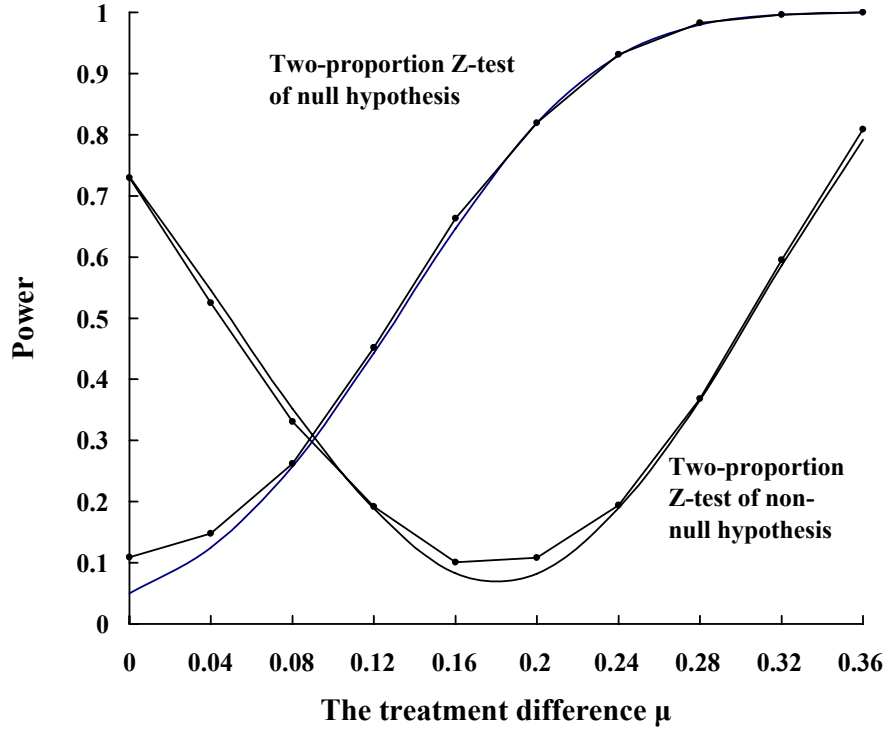

Figure 3. The observed and expected power of the two-proportion Z-tests of null and non-null hypothesis in superiority trials

$\alpha = 0.05$ (one-sided), $N = 160$, $\{k_j\} = (0.6, 0.4)$, $\pi_2 = 0.5$, the difference $\{\mu\} = (0, 0.04, \ldots, 0.36)$ and the margin of difference $\Delta = 0.18$ in the non-null hypothesis test. Broken lines indicate the observed power and smooth curves the expected power.

5. An Example from a Real Trial

A real trial (Liu et al, 2006) has been taken as an example. This trial was intended to assess the efficacy of arbidol hydrochloride tablet with ribovirin tablet as control in the therapy of naturally acquired influenza. A total of 213 patients were randomly allocated into the treatment and control groups with 108 and 105 respectively. The two drugs were received in the same dosage as 0.2 gram three times a day for six days. The two groups had 89 and 77 patients cured respectively. There were two patients in the control group lost to follow-up in the trial period.

5.1 Original Analysis

In the intent-to-treat analysis (ITT), the proportions of clinical cure were estimated as 0.824 and 0.733 and the difference was not statistically significant ($P > 0.05$). The authors concluded that "Arbidol was effective in the treatment of early naturally acquired influenza."

The conclusion was intended to favor the test drug but it did not mention the control. With the control, it may be rephrased as Conclusion 1 or 2. Now the authors' p-value does not support Conclusion 1. As for 2, no evidence can be set up with considering statistical significance alone.

5.2 Analysis for Clinical Significance

Then we had to consider clinical significance in terms of the CI under non-null hypothesis. This is a superiority trial. The confidence level was appointed at 95% (R-sided) for (C1) and (C2) and at 97.5% left-sided (L-sided) for (C3). The two proportions yielded the difference $D = 0.091$. A R-sided 95% CI for $\mu = 0$ was $[-0.002, 1]$ according to (C1). It crosses zero and then H_0 is not rejected.

The authors did not show Δ-value and we had to take an arbitrary value $\Delta = 0.14$ for further calculations. A L-sided 97.5% CI for $\mu = \Delta$ was $[0, 0.202]$ on the basis of (C3). It includes $\Delta = 0.14$ and then H is not rejected as well. This is just the case of Pattern 6 in Figure 1 and then we obtain Result 4 with the interpretation indeterminate in light of Figure 2. Neither Conclusion 1 nor 2 is consistent with the data. The same result was obtained using the test instead. The test (2.3) resulted in $Z_0 = 1.597$ ($P = 0.055$)(R-sided) and (2.6) $Z_\Delta = -0.876$ ($P = 0.190$)(L-sided). These are listed in the fourth row of the first part in Table 3.

Table 3. Results of the trial on arbidol hydrochloride in the therapy of naturally acquired influenza

y_1	y_2	Null hypothesis					Non-null hypothesis					Result
		$D-0$	Confidence interval		Z_0	P_0	$D-\Delta$	Confidence interval		Z_Δ	P_Δ	
1. Superiority trials ($\Delta = 0.14$)												
97	69	0.241	[0.151,1]	(C1)	4.241	0	0.101	[0.151,1]	(C2)	1.797	0.036	1
92	74	0.147	[0.055,1]	(C1)	2.588	0.005	0.007	[0.055,1]	(C2)	0.126	0.450	2
85	81	0.016	[-0.078,1]	(C1)	0.275	0.392	-0.124	[0, 0.127]	(C3)	-2.213	0.013	3
89	77	0.091	[-0.002,1]	(C1)	1.597	0.055	-0.049	[0, 0.202]	(C3)	-0.876	0.190	4
2. Inferiority trials ($\Delta = -0.14$)												
71	95	-0.247	[-1, -0.159]	(C4)	-4.352	0	-0.107	[-1, -0.159]	(C5)	-1.923	0.027	1
76	90	-0.153	[-1, -0.062]	(C4)	-2.700	0.003	-0.013	[-1, -0.062]	(C5)	-0.241	0.405	2
83	83	-0.022	[-1, 0.071]	(C4)	-0.386	0.350	0.118	[-0.133, 0]	(C6)	2.114	0.017	3
80	86	-0.078	[-1, 0.015]	(C4)	-1.378	0.084	0.062	[-0.189, 0]	(C6)	1.105	0.135	4
3. Significance trials ($\|\Delta\| = 0.14$)												
97	69	0.241	[0.134, 0.348]	(2.2)	4.241	0	0.101	[0.151,1]	(C7)	1.797	0.036	1
92	74	0.147	[0.037, 0.257]	(2.2)	2.588	0.010	0.007	[0.055,1]	(C7)	0.126	0.450	2
85	81	0.016	[-0.096, 0.127]	(2.2)	0.275	0.784	-0.124	[0, 0.127]	(C8)	-2.213	0.013	3
89	77	0.091	[-0.020, 0.202]	(2.2)	1.597	0.110	-0.049	[0, 0.202]	(C8)	-0.876	0.190	4

The fourth row of the first part shows the results of the trial, where the size and the frequency are $n_1 = 108$ and $y_1 = 89$ in the treatment group and $n_2 = 105$ and $y_2 = 77$ in the control group. The data in the remaining rows are imaginary for pursuing the other results.

The confidence level is appointed at $1 - \alpha = 0.95$ (2-sided) in (2.2), at 0.975 (1-sided) in (C3), (C6) and (C8) and at 0.95 (1-sided) in remaining confidence intervals.

5.3 An Extended Analysis

Table 3 also exhibits other possible results under certain conditions with both margins fixed and D-value varied. One may

pursue Pattern 1 on condition that the cure proportions change to 0.898 (97/108) and 0.657 (69/105) respectively. Then the difference was $D = 0.241$. The R-sided confidence interval for $\mu = 0$ became $[0.151, 1]$ based on (C1). It does not include zero and then H_0 is rejected. The process entered the non-null hypothesis with the result the R-sided confidence interval for $\mu = \Delta$ was $[0.151, 1]$ based on (C2). It does not include $\Delta = 0.14$ and then H is rejected as well and we get Result 1. It gives the interpretation that the efficacy of arbidol has clinical superiority over that of ribovirin. Pattern 2 may be obtained on condition of cure proportions 0.852 (92/108) and 0.705 (74/105). Hence the confidence interval became [0.055, 1] for $\mu = 0$ and [0.055, 1] for $\mu = \Delta$ and then H_0 is rejected but H not. We get Result 2 with the interpretation that the efficacy of test drug is statistically superior to that of control. Both the two results convey Conclusion 1 at different degrees.

Conclusion 2 comes from Pattern 5 and may be realized on condition that the cure proportions are 0.787 (85/108) and 0.771 (81/105). A R-sided 95% confidence interval for $\mu = 0$ was [-0.078, 1] and a L-sided 97.5% confidence interval for $\mu = \Delta$ [0, 0.127]. Then H_0 is not rejected and H is rejected. We thus reach Result 3 with the interpretation that the test drug is non-superior to the control. The result conveys Conclusion 2. These are listed in the first part of Table 3.

One may put the case into an inferiority trial as shown in the second part. Also it may be put into a significance trial as listed in the third part. Here the data were taken from the first part with a simple alteration that the values of D, μ and Δ were regarded to be absolute. And we took (2.2) directly to calculate a 2-sided 95% confidence interval for $\mu = 0$.

6. Discussion

Considering both statistical and clinical significance may broaden horizons in the analysis of clinical data. The analysis is practiced by combining a CI under null hypothesis with a CI under non-null hypothesis. The combination exhibits six possible patterns (see Figure 1). This is an intermediate step prior to the four possible results (see Figure 2).

6.1 Results of the Analysis

In superiority trials, Result 2 statistical superiority is the most common. In the example of Section 5, it requires the difference $D = 0.147$ as shown in the second row of the first part in Table 3. Result 3 non-superiority can be seen in the situation with very small difference and large margin. It requires the difference $D = 0.016$ as shown in the third row, which is smaller than $D = 0.091$ calculated from the data as shown in the fourth row. Result 1 clinical superiority is the most ideal but less common because it requires a larger difference, a smaller margin and/or a larger data set. The first row gives the required difference $D = 0.241$, which is much greater than $D = 0.091$. As to Result 4 with the interpretation indeterminate, it is usually due to a small data set.

The four results constitute a quadripartite procedure and may helpful for a meticulous evaluation of drugs. Of these, Result 1 and 2 convey Conclusion 1 at different degrees and Result 3 conveys Conclusion 2. Specifically, as another profile of equivalence, non-superiority can also be used to interpret equivalence. One may select a drug with the efficacy of clinical superiority, statistical superiority, or non-superiority in clinical therapies.

This approach prepares a convenience for interpreting discordant cases. In a superiority trial, we have $\mu \in [0, 1]$ with $\Delta \geq 0$ as mentioned in Subsection 2.2. It is therefore not difficult to claim non-superiority according to the pre-stated positive Δ-value. This is indicated by Pattern 5 in the right half of Figure 1, which will not alter even if the point estimate shifts to the left of $\mu = 0$ and becomes a discordant case. By contrast, it is inappropriate to claim non-inferiority post hoc from a superiority trial (Piaggio et al, 2006) because there is no a pre-stated negative Δ-value.

6.2 A Few Remarks

There are a few remarks when putting the combination into practice:

(i) The null does not refer necessarily to zero but any null hypothesis can be rewritten in the form $H_0 : \mu = 0$. For example, $H_0 : \mu = 0.618$ can be rewritten $H_0 : \mu = 0$, where $\mu = \pi - \pi_0$ and $\pi_0 = 0.618$. It therefore has a corresponding non-null hypothesis $H : \mu = \Delta$ with a pre-specified Δ-value. In the context with both $H_0 : \mu = 0$ and $H : \mu = \Delta$, it is necessary to use the term non-null hypothesis so as to avoid any confusion with the term null hypothesis .

(ii) A prescribed value is needed for the margin Δ. Its meaning has a corresponding relation and can be cross-referenced in different trials. FDA (2010) gives an advice in detail for valuating the margin. The following recommendations are provided in some other references (Blackwelder, 1982, Huque et al, 1989). It is generally set at 20% of the control effect, defined as the difference between the control response and the non-existing placebo response or set at half the lower limit of the 95% CI for the estimate of the control effect. If so, Δ-value will rely on the control effect and then the power will be difficult to control. We recommend that $\Delta = \eta\mu$, where $\eta = 0.1, \cdots, 0.5$ for testing $H : \mu \leq \Delta$ versus $H_a : \mu > \Delta$ and that $\Delta = \eta^{-1}\mu$ for testing $H : \mu \geq \Delta$ versus $H_a : \mu < \Delta$. The μ-value is specified first and then Δ-value is settled in design. With regard to analysis, D is used in place of μ-value. The power can be controlled since Δ-value depends upon μ-value.

If necessary, one may even negotiate with the FDA (Rockville, Maryland, United States) or the EMEA (European Medicines Agency, London, United Kingdom) about an acceptable boundary value.

(iii) Considering both statistical and clinical significance affects Type I and II errors and further the size of sample. Special attention is paid to the deflated Type I error rate marked in Result 1 and trivial in 2 and 3. It suggests that the sample size needed be larger for Result 1 than for 2 or 3. The expected and observed power have certain changes as shown in Section 4 and Figure 3.

The deflated Type I error rate implies a decrease of power. For example, the power was $1 - \beta = 0.949$ for evaluating statistical significance alone. This was calculated based on (2.1) with $\alpha = 0.05$ (1-sided), $N = 25$, $\{k_j\} = (0.6, 0.4)$, $\pi_2 = 0.2$ and $\mu = 0.6$. Considering clinical significance alone, the power was 0.809 based on (2.4) with $\alpha = 0.025$ and $\Delta = 0.06$. Considering both statistical and clinical significance like Result 3, the power became 0.801 based on (2.4) with $\alpha = P(3) = 0.024$. It implies that the usual non-inferiority trial contain a trivial pseudo increase in power.

For another, we may observe an increase of the required sample size. The required sample size was $N = 86$ and 131 for evaluating statistical and clinical significance respectively. The former was computed by (2.1) with $\alpha = 0.05$ (1-sided), $\beta = 0.1$, $\{k_j\} = (0.6, 0.4)$, $\pi_2 = 0.2$ and $\mu = 0.3$ and the latter by (2.4) with $\alpha = 0.025$ and $\Delta = 0.03$. Considering both of them, using $\alpha = P(3) = 0.024$ yielded $N = 133$.

6.3 Comparisons with Usual Approaches

The arrays are smoothly connected with the three-sided hypothesis testing proposed by Goeman et al, (2010). The first hypothesis, i.e. equivalence, is related to one of the results in a significance trial. The second refers to clinical superiority, a result in a superiority trial. The third belongs to clinical inferiority, a result in an inferiority trial.

The result equivalence from a significance trial is set up under the L-sided alternative $H_a : |\mu| < |\Delta|$ ($|\mu| \in [0, |\Delta|)$) and equivalence holds when $H : |\mu| \geq |\Delta|$ is rejected with $|D| < |\Delta|$. It can be rephrased in the form of 2-sided like $H_a : \mu \neq \Delta$ ($\mu \in [\Delta, 0]$ or $[0, \Delta)$). Clearly, the definition of equivalence in the arrays is the same as that in International Conference on Harmonisation (1998).

The figure in Piaggio et al, (2006), which is similar to Figure 1 in this article, shows eight scenarios in inferiority trials. In the figure, the right half serves inferiority trials and the left superiority trials. Scenarios A, B and E belong to the discordant cases since the point estimates are located in the left half. Scenarios C has its counterpart Pattern 5 in Figure 1 of this article. Scenarios D plays the same role as Pattern 3. Scenarios F corresponds to Pattern 6. Scenarios G and H are the same as Pattern 4 and 1 respectively. No scenarios corresponds to Pattern 2.

Similarly, we find six rules in Figure 2 in FDA (2010, Page 5). The structure of the figure is the same as that in Piaggio et al, (2006). Rule 1 has its counterpart Pattern 5 in Figure 1 of this article. Both Rule 2 and 3 play the same role as Pattern 6. As for Rule 4 and 5, they belong to the discordant cases. Rule 6 corresponds to Pattern 3. No rules correspond to Pattern 1, 2 and 4.

One may want to use a 2-sided CI in place of two 1-sided CIs. This is just the case in the figure (Piaggio et al, 2006), for example, Scenarios C shows a 2-sided CI used for two purposes. One purpose conveys that its lower bound is less than zero and the null hypothesis is not rejected. The other gives that its upper bound is less than Δ so that non-inferiority is demonstrated under non-null hypothesis. The confidence level may differ for the two purposes, e.g., 95% for the former and 97.5% for the latter. This way of using CIs relies on such reasons as below: (i) The lower and upper bound of a CI can be calculated at different confidence levels. (ii) Two-sided CIs can serve 1-sided alternatives. (iii) The lower and upper bound of a CI can serve different hypotheses. No one of the reasons is flawless and such a way of using CIs raises questions. Rule 1 of Figure 2 in FDA (2010, Page 5) is the same as Scenarios C. This can be overcome by using two separate 1-sided CIs in place of a 2-sided CI as shown in the arrays and Figure 1 of this article. The calculation will be reliable as long as following the correspondence in the arrays. Only the 2-sided CI (2.5) is unable to serve so many variations.

In a usual approach, statistical significance is considered alone. The relevant CI is used only under null hypothesis, there is no Type I error rate deflation problem, and the required sample size is smaller. However, one may lose the information of clinical superiority or non-superiority. This is just the case in the example in Section 5.

In another approach, clinical significance might be considered singly. Hence only a CI under non-null hypothesis would be needed and the effect size would be determined under non-null hypothesis with no Type I error rate deflation problem. A trial only considering clinical significance, however, fails to have the information of Result 2 though it is still useful after Result 1 or 3 is not demonstrated.

Statistical significance is logically prior to clinical significance. It is not possible to achieve clinical significance without obtaining statistical significance since $\mu > \Delta$ implies $\mu > 0$. Therefore the parameter μ is valuated first and Δ afterwards

in planning a trial. In a superiority trial, for instance, null hypothesis works on the range $[0, \infty)$ of μ and non-null one on the subrange $[0, \Delta]$ or $[\Delta, \infty)$. This provides an insight into the inference, which should be done on the whole range first and then on one of its subranges.

It is well known that a difference with statistical significance may be no real interest and a difference with clinical importance may be statistically non-significant. Thus clinical significance has been emphasized (Laupacis et al, 1988). A similar idea was mentioned by Kirk (2001). There are three questions concerning the estimated difference. First, is an observed result real or should it be attributed to chance (i.e. statistical significance)? Second, if the result is real, how large is it (i.e. effect size)? Third, is the result large enough to be meaningful and useful (i.e. clinical or practical significance)? Such an idea can be even dated back to Kendal and Stuart (1979). There are two questions concerning the estimated difference. The first is whether there is any true difference and the second is concerning its magnitude.

Active control has been recommended in clinical trials since ethical doubts on placebo were reminded in the fifth edition of Helsinki Declaration (Temple, 1983, Fleming, 1987, Fleming, 1990). The data of active control trials are often analyzed with consideration of both statistical and clinical significance.

With considering both statistical and clinical significance, a superiority trial may result in superiority or non-superiority. Nevertheless, a larger data set is usually needed. By contrast, considering statistical significance alone requires a smaller data set but may gain just superiority. The explanation presented is quite general and, it is hoped, may be a reference for those who wish to consider both statistical and clinical significance in the analysis of clinical data. The arrays and the flowchart should be sufficient for most uses.

Acknowledgements

This research was supported by The Development Plan of Scientific Researches in Henan grant 009017200. I am grateful to Professor Xu Dezhong for valuable comments and to Renee Plumb, MD for her careful revising of the script.

References

Blackwelder, W. C. (2004). Current Issues in clinical Equivalence Trials. *J Dent Res 83* (Spec Iss C), C113-C115. http://dx.doi.org/10.1177/154405910408301s23

Blackwelder, W. C. (1982). Proving the null hypothesis in clinical trials. *Controlled clinical trials 3* (Spec Iss C), 345-353. http://dx.doi.org/10.1016/0197-2456(82)90024-1

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall (ISBN 0412124203 0412124203), Section 7.2 (iii).

D'Agostino Sr, R. B., Massaro, J. M., & Sullivan, L. M. (2003). Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. *Statistics in Medicine 22*, 169-186. http://dx.doi.org/10.1002/sim.1425

Dann, R. S., & Koch, G. G. (2008). Methods for one-sided testing of the difference between proportions and sample size considerations related to non-inferiority clinical trials. *Pharm Stat 7*, 130-141. http://dx.doi.org/10.1002/pst.287

Dunnett, C. W., & Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables. *Biometrics 33*, 593-602. http://dx.doi.org/10.2307/2529457

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

Fleming, T. R. (1990). Evaluation of active control trials in AIDS. *Journal of Acquired Immune Deficiency Syndrome 2*, 82-87.

Fleming, T. R. (1987). Treatment evaluation in active control studies. *Cancer Treatment Reports 71*, 1061-1065.

Goeman, J. J., Solari, A., & Stijnen, T. (2010). Three-sided hypothesis testing: simultaneous testing of superiority, equivalence and inferiority. *Statistics in Medicine 29*, 2117-2125. http://dx.doi.org/10.1002/sim.4002

Good, I. J. (1992). The Bayes / non-Bayes compromise: a brief review. *Journal of the American Statistical Association 87*, 597-606. http://dx.doi.org/10.1080/01621459.1992.10475256

Greenstein, G. (2003). Clinical versus statistical significance as they relate to the efficacy of periodontal therapy. *J Am Dent Assoc 134*, 583-591. http://dx.doi.org/10.14219/jada.archive.2003.0225

Haase, R. F., Ellis, M. V., & Ladany, N. (1989). Multiple criteria for evaluating the magnitude of experimental effects. *Journal of Counseling Psychology 36*, 511-516. http://dx.doi.org/10.1037/0022-0167.36.4.511

Hirotsu, C. (2007). A unifying approach to non-inferiority, equivalence and superiority tests via multiple decision processes. *Pharmaceutical Statistics 6*, 193-203. http:// dx.doi.org/ 10.1002/ pst. 305

Holmgren, E. B. (2001). Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. *Journal of Biopharmaceutical Statistics 9*, 651-659. http://dx.doi.org/10.1081/BIP-100101201

Huque, M. F., Dubey, S., & Fredd, S. (1989). Establishing therapeutic equivalence with clinical endpoints. *Proceedings of the American Statistical Association, the Biopharmaceutical Section*. American Statistical Association: Alexandria, VA, 46-52.

International Conference on Harmonisation. (1998). Guidance E9: statistical principles for clinical trials. *Fed Register 63(179)*, http:// www.ifpma.org/ich1.html.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol. 59*(1), 12-19. http://dx.doi.org/10.1037/0022-006X.59.1.12

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy 15*, 336-352. http://dx.doi.org/10.1016/S0005-7894(84)80002-7

Kendal, S. W., & Stuart, A. (1979). *The Advanced Theory of Statistics. Volume 2.*. London: Charles Griffin and Company Limited, 175.

Killoy, W. J. (2002). The clinical significance of local chemotherapies. *J Clin Periodontol supplement 2*, 22-29. http://dx.doi.org/10.1034/j.1600-051x.29.s2.2.x

Kirk, R. E. (2001). Promoting good statistical practices: some suggestions. *Educ Psychol Meas 61*, 213-218. http://dx.doi.org/10.1177/00131640121971185

Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *J Am Acad Child Adolesc Psychiatry 42*, 1524-1529. http://dx.doi.org/10.1097/01.chi.0000091507.46853.d1

Laupacis, A., Sackett, D. L., & Roberts, R. S. (1988). An assessment of clinically useful measures of the consequences of treatment. *N. Engl J Med 318*, 1728-1733. http://dx.doi.org/10.1056/NEJM198806303182605

Lindgren, B. R., Wielinskyi, C. L., Finkelstein, S. M., & Warwick, W. J. (1993). Contrasting clinical and statistical significance within the research setting. *Pediatr Pulmonol. 16*, 336-340. http://dx.doi.org/10.1002/ppul.1950160603

Liu, H. B., Qu, W. X., Li, S. Q., Sun, J. M., Miao, L. N., Jia, Y., & Zhang, J. (2006). Multiceter randomized double blind parallel clinical trial of arbidol hydrochloride tablet in the treatment of natural acquired influenza. *The Chinese Journal of Clinical Pharmacology 22*, 403-405.

Lu, Y., & Fang, J. Q. (2003). *Advanced Medical Statistics*. New Jersey: World Scientific Publishing Co Pte Ltd, 448.

Marriott, F. H. C. (1990). *A Dictionary of Statistical Terms*. Harlow: Longman Scientific and Technical (ISBN 0-582-01905-2), 5th edition.

Mehta, C. R., Patel, N. R., & Tsiatis, A. A. (1984). Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics 40*, 819-825. http://dx.doi.org/10.2307/2530927

Pearson, E. S. (1939). William Sealy Gosset, 1876-1937 (2) 'Student' as Statistician. *Biometrika 30*, 210-250.

Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., & Evans, S. J. W. (2006). Reporting of non-inferiority and equivalence randomized trials an extension of the consort statement. *JAMA 295*, 1152-1160. http://dx.doi.org/10.1001/jama.295.10.1152

Scott, I. A. (2009). Non-inferiority trials: determining whether alternative treatments are good enough. *The Medical Journal of Australia 190*, 326-330.

Temple, R. (1983). Difficulties in evaluating positive control trials. *Proceedings of the American Statistical Association, the Biopharmaceutical Section.*. American Statistical Association: Alexandria, VA, 1-7.

U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. (2010). Guidance for Industry Non-inferiority Clinical Trials., 3-6.

Zhao, G. (2003). A test of non-null hypothesis on proportions and its applications in clinical trials. *Journal of the Chinese Statistical Association 41*, 21-37.

Zhao, G. (2015). A test of non-null hypothesis for linear trends in proportions. *Communications in Statistics - Theory and Methods 44*(8), 1621-1639. http://dx.doi.org/10.1080/03610926.2013.776687

Zhao, G. (2008). Tests of non-null hypothesis on proportions for stratified data. *Statistics in Medicine 27*(9), 1429-1446. http://dx.doi.org/10.1002/sim.3023