Statistical Model and Clinical Trial Analysis

Xi Chen*

PharmClint Co. Ardsley, USA

***Corresponding author:** Xi Chen, PharmClint Co. Ardsley, New York, USA

Abstract

The different setup of the clinical trial goals may lead to different statistical analysis methodologies and different trial conclusions. Some statistical methodologies used in the conventional clinical trial design are against the common acceptable probability axiomatic system. In the sense, the term superiority and non-inferiority are not only misleading, they are even not well established. To clarify these concepts and to find a feasible solution for clinical trial analysis, it is necessary to review the foundation of the statistical analysis and assumptions in common practice.

Keywords: Confidence interval; Distribution; Measure; Model

Abbreviations: SUP: Superiority; NI: Non-Inferiority; SM: Study Medication; AC: Comparator; PDF: Probability Density Function; Str: strict; EQ: Equivalence; MCID: Minimum Clinical Important Difference; Con: Conservative; Agr.: Aggressive; INF: Inferiority; MDD: Minimum Detectable Difference; LRT: Likelihood Ratio Test; PMF: Probability Mass Function

Introduction

The goal of the clinical trial analysis can be setup in different ways, they may lead to different set of statistical analysis methodologies, and even different statistical conclusions. The conventional clinical trial is to determine the superiority (SUP) or non-inferiority (NI) of study medication (SM) against a comparator (AC). This kind of the trial design used the confidence interval (CI) approach, and the trial design depends upon a prespecified SUP or NI margin. For simplicity, it is named CI based trial. A more objective way to present the trial goal is to determine the statistical model characterizing the treatment difference of SM against AC. The trial is named model based trial. The purpose of this manuscript is to investigate the impact of the trial goals on the statistical analysis inference and the conclusions [1].

Within a clinical trial project, if the random sample follows a Bernoulli distribution, the trial is generally named a Bernoulli trial. Although in practice, the statistical models such as the binomial distribution, the Poisson distribution and even the normal approximation are widely used, they are built up on a common Bernoulli foundation. For this reason, the following discussion is applicable to all the statistical setups [2].

Model

Probability axiomatic system

In Kolmogorov's probability theory (6) the probability P of some event E, denoted by $P(E)$, is usually defined such that P satisfies the Kolmogorov axioms: Let (Ω, F, P) be a measure space with $P(\Omega) = 1$, the (Ω, F, P) is a probability space, with sample space Ω, event space F and probability measure P.

The Kolmogorov axioms can be further divided into the following three axioms: First axiom, $P(E) > 0$ for any $\phi \neq E \subset \Omega$. Second axiom, $P(\Omega) = 1$. Third axiom, any countable sequence of disjoint events $E_1, E_2, \dots$ satisfies $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$.

The Kolmogorov axioms led to several implications in the statistical analysis. For any statistical analysis, the sample space (Ω) and the probability measure (P) has to be kept unique and unchanged. Any element in event space (F) has to be properly defined. Any relaxation from these conditions may lead to the conclusion not on the common acceptable foundation, they will be beyond the scope of the statistical analysis.

A map qualified to be a measure

In statistical analysis, a measure on a set is a systematic way to assign a number to each suitable subset of that set. A properly defined probability measure should have the following properties:

I. Non-negativity. For all E in Ω, $P(E) \geq 0$.

II. Null empty set. $P(\phi) = 0$.

III. Monotonicity. If E_1 and E_2 are measurable sets with $E_1 \subseteq E_2$, then $P(E_1) \leq P(E_2)$.

IV. Subadditivity. For any countable sequence $E_1, E_2, E_3, \dots$ of sets E_n we have $P\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} P(E_i)$

V. Continuity from below. If $E_1, E_2, E_3, \dots$ are measurable sets and E_n is a subset of E_{n+1} for all n, then the union of the sets E_n is measurable, and $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \lim_{i \to \infty} P(E_i)$

VI. Continuity from above. If $E_1, E_2, E_3, \ldots$ are measurable sets and for all n, $E_{n+1} \subset E_n$, then the intersection of E_n is measurable. Furthermore, if at least one of the E_n has finite measure, then $P\left(\bigcap_{i=1}^{\infty} E_i\right) = \lim_{i \to \infty} P(E_i)$.

Theorem 1

A properly defined probability measure should be unique.

Proof: Let Ω be decomposed as $\Omega = E \cup E^c$, where E^c is the complement of E. For a properly defined measure P, it is known that $P(E \cup E^c) = P(\Omega) = 1$ due to the second axiom, and $P(E) + P(E^c) = 1$ due to the third axiom. If there is a pair of probability measures, $P_1 \neq P_2$, exists with $P(E) + P(E^c) = 1$ and $P_2(E) + P_2(E^c) = 1$. Let $E = \Omega$ and $E^c = \phi$, we have $P_1(E) = P_2(E) = 1$, it conflicts to the known inequality $P_1(E) \neq P_2(E)$.

Some additional notes regarding axiomatic system

Within a statistical analysis project, there are some additional points should draw plenty of attentions. First, if a continuous random variable X is considered, along with a probability density function (PDF) $f(x)$, the probability measure P is defined as $f(x)dx$. The uniqueness of the probability measure basically requires $f(x)$ to be unique, it is the key argument of this manuscript. Second, the axiomatic system is a commonly believed truth, it serves as a bible which does not need a proof. Any misunderstanding, expansion or relaxation will lead to theories beyond the discussion foundation. For example, Euclidean geometry follows five axioms, a minor change in the parallel axiom led to a complete different set of the geometric theory named the non Euclidean geometry. The communication between the Euclidean geometricians and non Euclidean geometricians are basically speaking different languages without a convincing translation, the debate needs to clarify which geometry to be the discussion topic, without such a foundation, the discussion is meaningless [3,4].

Definition of superiority and non-inferiority

In a Bernoulli clinical trial, assume the cure rate of SM to be p_1 and the cure rate of AC to be p_0, the both arms of SM and AC are assessed by the same statistical model. The concepts of strict (Str.) SUP, NI and equivalence (EQ) are defined as follows:

$$\text{Str.SUP} : \lim_{n \to \infty} \Pr(\hat{p}_1 - \hat{p}_1 > 0) = 1$$

$$\text{Str.NI} : \lim_{n \to \infty} \Pr(\hat{p}_1 - \hat{p}_0 \geq 0) = 1$$

$$\text{Str.EQ} : \lim_{n \to \infty} \Pr(\hat{p}_1 - \hat{p}_0 = 0) = 1$$

Where the hat represents the estimation. Without specified otherwise, all the estimations are maximum likelihood estimates.

Some additional notes may not be redundant, following (7), a statistical model is a set of probability distributions on the sample space Ω. The same statistical model implies that both SM and AC arms are evaluated with same class of distributions, for example, a Bernoulli distribution with parameter p_1 or p_0 respectively.

Generalized outcomes

Moreover, as it is shown in (2), along with a constant minimum clinical important difference (MCID), (D_c), the conservative (Con.) and aggressive (Agr.) SUP, EQ and inferiority (INF) are defined as follows:

$$\text{Con.SUP} : \lim_{n \to \infty} \Pr(\hat{p}_1 - \hat{p}_0 > D_c) = 1$$

$$\text{Agr.SUP} : \lim_{n \to \infty} \Pr(\hat{p}_1 - \hat{p}_0 > -D_c) = 1$$

$$\text{Con.EQ} : \lim_{n \to \infty} \Pr(\hat{p}_1 - \hat{p}_0 = D_c) = 1$$

$$\text{Agr.EQ} : \lim_{n \to \infty} \Pr(\hat{p}_1 - \hat{p}_0 = -D_c) = 1$$

$$\text{Con.INF} : \lim_{n \to \infty} \Pr(\hat{p}_1 - \hat{p}_0 < D_c) = 1$$

$$\text{Agr.INF} : \lim_{n \to \infty} \Pr(\hat{p}_1 - \hat{p}_0 < -D_c) = 1$$

For convenience, the strict outcomes will generally be named as outcomes, the conservative and aggressive outcomes will generally be referred as the generalized outcomes. All of these properties are asymptotic properties. It is necessary to show that when the sample size is beyond a certain level, the statistical conclusion will not change along with the sample size and the trial conduction. Besides, since $p_1 - p_0 \geq 0 = (p_1 - p_0 > 0) \cup (p_1 - p_0 = 0)$, NI can be considered as the union of SUP and EQ. Besides, $NI \cup INF = \Omega$, the generalized NI can be derived from the generalized SUP and EQ, or from the generalized INF [5].

Under normal approximation, here is a key point to be mentioned, assuming σ^2 is known, the strict outcomes are assessed by a model $N(0, \sigma^2)$; the conservative outcomes are assessed by a model $N(\Delta, \sigma^2)$; and the aggressive outcomes are assessed by a model $N(-\Delta, \sigma^2)$.

Conventional clinical trial practice

FDA guidance (4) refers to the CI based clinical trial design. Within a Bernoulli trial, assume the cure rate of the study medication (SM) be p_1 and the cure rate of the comparator to be p_0, for a NI trial, the study hypothesis is

$$H_0 : p_1 - p_0 < -\Delta$$

$$H_a : p_1 - p_0 \geq -\Delta$$

Following the FDA guidance, if H_0 is true, SM is inferior to AC by Δ or more; if H_a is true, SM is not inferior to AC by Δ. In most cases, the test for a treatment effect corresponds to showing that the lower bound of the two-sided 95% confidence interval (equivalent to the lower bound of a one-sided 97.5% confidence interval) for $p_1 - p_0$ is $-\Delta$. This result shows that the effect of SM is not inferior to AC by Δ.

Carefully review the description of the CI based NI trial, some natural questions may arise. At first, the name of noninferiority implies $p_1 - p_0 \geq 0$, to show $p_1 - p_0 \geq -\Delta$ is in fact a limited inferiority other than the non-inferiority. Secondly, how does the CI approach link with the hypothesis test is not quite clear.

It is noted that the 95% confidence interval is formulated as $CI = (\hat{p}_1 - \hat{p}_0 - z_{\alpha/2}\hat{\sigma}, \hat{p}_1 - \hat{p}_0 + z_{\alpha/2}\hat{\sigma},)$, which was assessed by the distribution $N(\hat{p}_1 - \hat{p}_0, \hat{\sigma}^2)$, it is different from either distributions $N(-\Delta, \hat{\sigma}^2)$ or $N(0, \hat{\sigma}^2)$.

Uniqueness of the probability measure

The CI based trial analysis applied some operations in violation to the axiomatic system. The CI based trial is built upon the foundation of two distinct statistical distributions $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_0 \sim N(\mu_0, \sigma_0^2)$, the treatment difference is defined as $dY = Y_1 - Y_0$. Not losing the generality, dY follows a normal distribution $dY \sim N(\mu_1 - \mu_0, \sigma^2)$, where σ^2 is a proper mapping from σ_1^2 and σ_2^2.

It is easy to see along with a $\Delta > 0$, $\omega = \Pr(dY > 0 \| \mu_1 - \mu_0) = 1/2$ and $\theta = \Pr(dY > 0 \| \mu_1 - \mu_0 - \Delta) > 1/2$. In the notation, $\Pr(dY > 0 \| \mu_1 = \mu_0)$ implies to use $N(\mu_1 - \mu_2 = 0, \sigma^2)$ to assess the event $dY > 0$; $\Pr(dY \le 0 \| \mu_1 = \mu_0 - \Delta)$ implies to use $N(\mu_1 - \mu_0 = \Delta, \sigma^2)$ to assess the event $dY \le 0$. It should be noted that both ω and θ are not conditional probabilities in traditional sense, they are still unconditional distributions. $dY > 0$ and $dY \le 0$ are disjoint and $(dY > 0) \cup (dY \le 0) = \Omega$. Based on the second and third axioms, $\Pr(dY > 0 \cup dY \le 0) = P(\Omega) = 1$, but $\omega + \theta \ne 1$. It shows that the two distributions, $N(\mu_1 = \mu_2, \sigma^2)$ and $N(\mu_1 - \mu_2 = -\Delta, \sigma^2)$, do not meet the requirements to construct a probability space, but they were actively used as constraints in the trial design. More specifically, such a constraint is reflected in the well known relationship of $z_{\alpha/2}\sigma + z_{\beta}\sigma = \Delta$ or the similar equation. This equation is widely used to determine the trial sample size, and the arbitrary increase of the trial sample size is not allowed in the current clinical trial practice [6].

From the probability point of view, the statistics are not appropriately defined. ω is defined with $f(x) = N(\mu_1 - \mu_0 = 0, \sigma^2)$ and θ is defined with $dY_0 \sim N(\mu_1 - \mu_0 = -\Delta, \sigma^2)$, although they are properly defined respectively, they are associated with different statistical measures. Any analysis associated with different measures will be meaningless.

The control of the statistical error

Any CI based clinical trial design is always associated with the type I and type II error control, but the two types of the error control were defined with respect to the different statistical distributions. In a NI trial, the distribution associated with the null hypothesis is $dY_0 \sim N(\mu_1 - \mu_0 = -\Delta, \sigma^2)$, and for the alternative hypothesis, the associated distribution is $dY_a \sim N(\mu_1 - \mu_0 = 0, \sigma^2)$. Similarly, in a SUP trial, the distribution associated with the null hypothesis is $dY_0 \sim N(\mu_1 - \mu_0 = 0, \sigma^2)$, and the distribution associated with the alternative hypothesis is $dY_a \sim N(\mu_1 - \mu_0 = \Delta_1, \sigma^2)$.

Obviously, the type I and type II error controls are with respect to the different probability measures, and combination of the two type of the errors is against the basic probability axiomatic system. More specifically, for a CI based NI trial, the type I error is defined as $\alpha/2 = \Pr(dY > z_{\alpha/2}\sigma - \Delta \| \mu_1 = \mu_0 - \Delta)$ which is assessed by $N(\mu_1 - \mu_0 = -\Delta, \sigma^2)$ and the power is defined as $1 - \beta = \Pr(dY > z_{\beta}\sigma \| \mu_1 = \mu_0)$ which is assessed by $N(\mu_1 - \mu_0 = 0, \sigma^2)$.

Needless to say, the power equals $1 - \beta$. The control of the type I and type II error could not be accurate at the same time. More likely, neither control is accurate.

The proper definition of the statistical distribution

In a CI based NI trial, define event $E = (-\Delta < dY < 0)$, $P(E)$ is not properly defined. Similarly, for a CI based SUP trial, define event $E = (0 < dY < \Delta_1)$, $P(E)$ is also not properly defined. The key is that for any $dY \in E$, it is not clear which P is used to evaluate $P(E)$, $N(0, \sigma^2)$ or $N(-\Delta, \sigma^2)$. The width of the interval $-\Delta < dY < \Delta_1$ can be controlled through the choice of $\Delta > 0$ and $\Delta_1 > 0$, but the crucial debate concentrates in this interval.

The flaw in definition

Parallel to a NI trial, the hypothesis for a SUP trial is set as

$H_0 : p_1 - p_0 \le 0$

$H_a : p_1 - p_0 > 0$

To test the hypothesis, "in most cases, the test for a treatment effect corresponds to showing that the lower bound of the one-sided 97.5% confidence interval" for $p_1 - p_0$ is > 0.

It is noted that it is not enough to provide an accurate definition for superiority. In fact, $p_1 - p_0 > 0$ is a logical statement. In logic, a statement is either a meaningful declarative sentence that is either true or false, or that which a true or false declarative sentence asserts. The treatment effect shown through a confidence interval approach does not meet the requirement for an assertion. The width of the CI depends upon the sample size, and the lower bound of the CI less than 0 does not mean that there is no treatment effect. On the other hand, following the FDA convention in testing the NI hypothesis, when the null hypothesis is rejected, the conservative way to make a claim is that the SM is superior to AC by 0. In fact, superior by 0 include the possibility that SM is not superior to AC at all [7].

A confidence interval provides two pieces of information, a fact and the confidence. The two are correlated, but they are independent. In a NI trial, the low end of the confidence interval above the non-inferiority margin $(-\Delta)$ does not provide information of the fact, in the other words, it does not mean that there is no any inferiority. The claimed noninferiority is in fact a limited inferiority. $p_1 - p_0 > -\Delta$ provides a confidence only to the fact of this limited inferiority. It is a central debating point on the current practice.

Model Based Clinical Trial Design

The model based clinical trial design is summarized in (1) and (2), where (1) discusses the strict outcomes and (2) discusses the generalized outcomes.

Study hypothesis setup

Within a model basend clinical trial, for either a NI study or a SUP study, the study hypothesis has the same form

$$H_0 : M(p_1) = M(p_0)$$

$$H_a : M(p_1) \neq M(p_0)$$

W here M represents the statistical distribution to characterize the efficacy of SM or AC. M is used in both sides implies that both SM and AC are assessed by the same distribution. The only difference is the mapping between the trial outcome and the trial conclusions. For each trial, there are three possible outcomes, EQ implies H_0 is confirmed, or say H_a is rejected. SUP implies H_a is confirmed and $p_1 - p_0$. INF implies H_a is confirmed and $p_1 < p_0$. Naturally, NI implies EQ or SUP conclusion.

The formulation of the model

We will consider the strict outcomes first. With Bernoulli model for each subject, the probability mass function (p.m.f.) can be written as

$$f(m, p) = p^m (1-p)^{1-m} \text{ for } m \in (0,1)$$

Where $m = 1$ represents a cure, and $m = 0$ represents a failure, for n subjects, the likelihood function can be written as

$$L(n, p) = \prod_{i=1}^{n} p^{m_i} (1-p)^{(1-m_i)}$$

$$\log L(n, p) = np \log(p) + n(1-p) \log(1-p)$$

If SM performs same as AC, $p_1 = p_0 = p$. $M_E = M(p, p) = M(p)M(p)$, $\log L_E = 2 \log L(n, p)$. If there is a treatment difference, let $p_1 = p + \varepsilon/2$ and $p_0 = p - \varepsilon/2$, $M_U = M(p_1, p_0) = M(p_1)M(p_0)$, $\log(L_U) = \log L(n, p_1) + \log L(n, p_0)$.

Combine the results, we get

$$\log(L_U) = \log L(n, p_1) + \log L(n, p_0).$$

$$\log(L_E) = 2n \log(p) + 2n(1-p) + \log(1-p)$$

$$+ np_0 \log(p_0) + n(1-p_0) \log(1-p_0)$$

The generalized outcomes

When the generalized outcomes are considered, there are two sets of the trial hypotheses under consideration. Along with a pre-specified D_c, based on M_+ model, the trial hypothesis is

$$H_{0+} : M(p_1) = M(p_0 + D_c)$$

$$H_{a+} : M(p_1) \neq M(p_0 + D_c)$$

Similarly, based on M_- model, the trial hypothesis is

$$H_{0-} : M(p_1) = M(p_0 - D_c)$$

$$H_{a-} : M(p_1) \neq M(p_0 - D_c)$$

Compared with the strict model, the only difference is the definition of p_1 and p_0. When M_+ model is considered, $p_1 = p + D_c + \varepsilon/2$ is considered to be the correct model, otherwise, the correct model is M_U. Compared with the conventional analysis, the χ^2 distribution is the only statistical model involved in the statistical inference, there is no second distribution used.

The control of the statistical error

The model based clinical trial does not consider the type I and type II statistical errors, since they are defined with different statistical distributions. But, any statistical analysis is always associated with two types of statistical error, the system error and the random error. The conventional clinical trial practice does not distinguish the two. The model based trial (1) introduced two important limitations for the error control, constant MCID ($D_c \geq 0$) and statistical MCID ($D_s > 0$). Between the two, D_c controls the system error and D_s controls the random error.

It is noted that the rejection of the null implies the confirmation of the alternative and vice versa. There are two levels for a trial execution, confirmed or unconfirmed. For a given D_s, if $|dY| < D_s$, the trial result is confirmed, otherwise it is unconfirmed. Only the confirmed trial result can be considered as the trial final conclusion.

The choice of D_c has impact on the statistical model used in inference, more specifically, it is related to the mean of the random variable. The choice of D_s has only impact on the accuracy of the parameter estimation. The width of the CI has integrated both D_c and D_s into one control, it does not distinguish the system and random errors.

The control of the random error

To decide which model between M_E and M_U to be the correct model, it needs to get a solution of the following equation.

$$f(\varepsilon) = -2 \log(L_E) + 2 \log(L_U) - 3.84 = 0$$

The solution to the equation, $\varepsilon*$, named as the minimum detectable difference (MDD), will be the largest magnitude of ε under which M_E is considered to be the correct model. $\varepsilon*$ with different p and n are summarized in the Table 1.

Table 1: MDD for different p and n.

p	$n = 300$	$n = 350$	$n = 400$	$n = 450$	$n = 500$
0.6	0.0783	0.0725	0.0679	0.064	0.0607
0.65	0.0763	0.0706	0.0661	0.0623	0.0591
0.7	0.0733	0.0678	0.0635	0.0598	0.0568
0.75	0.0692	0.0641	0.06	0.0565	0.0536
0.8	0.0639	0.0592	0.0554	0.0522	0.0495
0.85	0.057	0.0528	0.0494	0.0466	0.0442
0.9	0.0478	0.0443	0.0414	\0.0391	0.0371

Table 2: Minimum sample size n per arm for different p and D_s.

p	$D_s = 3.0\%$	$D_s = 4.0\%$	$D_s = 5.0\%$	$D_s = 0.6\%$
0.6	2049	1153	738	512
0.65	1942	1092	699	486
0.7	1793	1008	645	448
0.75	1600	900	576	400
0.8	1365	768	491	341
0.85	1087	611	391	271
0.9	766	430	274	190

Defined the statistical MCID as D_s, to meet the requirement of $|\varepsilon^*| < D_s$. the minimum sample size needed for different p and D_s are summarized in the Table 2.

Seen from the Table 1, the higher the p and the larger the n, the smaller the MDD. Also seen from table 2, the lower the p and the lower the D_s, the higher the n. the smaller the MDD. Also seen from Table 2, the lower the p and the lower the D_s, the higher the n. The linkage among p, n, D_s and ε^* is the key for the adaptive property (1). During the trial design, the actual p could not be accurately predicted. As a safeguard practice, the sample size should be higher than the minimum. Since the algorithm has an asymptotic property, the increase of the sample size will not affect the trial conclusion, it is a key difference of model based trial from the conventional CI based trial.

In the clinical trial practice, the confirmed SUP conclusion requires that $CI = \left(\overline{Y_1} - \overline{Y_0} - z_{\alpha/2}\hat{\sigma}^2, \overline{Y_1} - \overline{Y_0} + z_{\alpha/2}\hat{\sigma}^2 \right)$ and $\varepsilon^* < D_s$. The confirmed conclusion of EQ requires that $|p_1 - p_0| = \varepsilon < \varepsilon^*$ and $\varepsilon^* < D_s$. As discussed before, $0 < p_1 - p_0 = \varepsilon > \varepsilon^*$ or $|p_1 - p_0| = \varepsilon < \varepsilon^*$ provides a statement regarding a fact, and $\varepsilon^* < D_s$ provides the confidence. A unconfirmed conclusion will not be used as the trial final conclusion, but it provides an important guidance for the trial execution such as the interim analysis.

Comparison of the two strategies

Interpretation of NI and SUP

For CI based trial, guidance (4) has claimed that the NI is a misnomer, as guaranteeing that the test drug is not any (even a little) less effective than the control can only be demonstrated by showing that the test drug is superior. What noninferiority trials seek to show is that any difference between the two treatments is small enough to allow a conclusion that the new drug has at least some effect or, in many cases, an effect that is not too much smaller than the active control.

The CI based NI trial concentrated on $dY = Y_1 - Y_0$ with a model $dY \sim N\left(\overline{Y_1} - \overline{Y_0}, \hat{\sigma}^2 \right)$. The $(1-\alpha)\%$ CI for dY can be written as $CI = \left(\overline{Y_1} - \overline{Y_0} - z_{\alpha/2}\hat{\sigma}^2, \overline{Y_1} - \overline{Y_0} + z_{\alpha/2}\hat{\sigma}^2 \right)$. The lower bound of CI above $-\Delta$ and the upper bound of CI above 0 will be the criteria to reject the NI null hypothesis, which implies $\mu_1 - \mu_0 > -\Delta$. In fact, a positive outcome of the NI trial implies that dY came from a distribution $dY \sim N\left(\mu_1 = \mu_0, \hat{\sigma}^2 \right)$, not $dY \sim N\left(\mu_1 = \mu_0 - \Delta, \hat{\sigma}^2 \right)$. The conclusion $\mu_1 = \mu_0$ is not claimed, since it is not the most conservative way to present the trial result. $\mu_1 - \mu_0 \geq -\Delta$ is considered to be a conservative conclusion, but it is based on a rejected statistical model. If a statistical model is considered false, any conclusion derived from the statistical model could not be considered as a truth. As previously discussed, a statement regarding a fact does not need to be either conservative or aggressive, it should stay on a neutral viewpoint. In the other words, to characterize a random variable, the mean of the distribution should be used, not the end of the confidence interval.

The key question the guidance (4) did not answer is how to define "small enough". For example, with $\Delta = 0.10$, if $dY = 0.09$ can be considered small enough? The guidance is not able to provide a definite answer.

Unification of SUP and NI trials

In CI based trial design, SUP and NI trial are defined on different statistical distributions. Although there is a guidance (3) for the switching from NI to SUP trial, the switching is not strictly on the statistical analysis foundation and the switching from SUP to NI is impossible. In the model based trial design, both SUP and NI trial are on exactly same set of statistical analysis based on the same statistical model, and the difference only exist in the final stage of interpreting the trial result.

Statistical model for inference

There is a significant difference in the choice of the statistical model for inference. In the CI based trial design, there are always two distinct statistical models involved, which automatically introduced two statistical measures in the same analysis. In the model based clinical trial design, no matter SM and AC are same or not, all the statistical inference is based on a single χ^2 distribution. All the statistical inference have been unified into the same statistical structure. In a hypothesis test, the alternative hypothesis has to be strictly complementary to the null hypothesis. Then, the rejection of the null hypothesis is considered to be the proof that the alternative hypothesis to be true. If there is a gap between the null hypothesis and alternative hypothesis set up, the foundation of the hypothesis test is destroyed. Return to the CI based NI trial, the success of the trial is to show that $dY \sim N\left(\mu_1 = \mu_0, \hat{\sigma}^2 \right)$, not $dY \sim N\left(\mu_1 = \mu_0 - \Delta, \hat{\sigma}^2 \right)$. In the other words, type I error control only is not enough to show the trial conclusion, since $dY \sim N\left(\mu_1 = \mu_0 - \Delta, \hat{\sigma}^2 \right)$ is a rejected model. To show a successful trial conclusion, it is more important to show the type II error control, but the CI based trial generally lack the information in trial conclusion. In contrast, the model based trial design unified the type I and type II error controls, the statistical inference is based on a sole statistical distribution, or say the same model. In the sense, the statistical hypothesis test is more accurate.

Confidence of statistical inference

In CI based trial design, the confidence of the parameter estimates were controlled by the width of CI. Since the width of CI is the function of the trial sample size, the change of the sample size has a direct impact on the trial conclusion. On the other hand, the CI based trial design does not make difference between the system error and the random error, and it does not clarify the size of the system error. When the model based trial design is considered, if there is an allowed inferiority specified, it will be controlled by the constant MCID (D_c). It is discussed extensively in (3), which is related to a shifted model.

Asymptotic and adaptive properties

During the discussion of the statistical properties of parameter estimation, the asymptotic and self-adaptive properties of the parameter estimation are important properties to the analysis. The asymptotic property refers to the property that the random

error will get smaller along with the increase of the sample size. More importantly, when the sample size is beyond the minimum requirement, the conclusion of the trial will not change along with the increase of the sample size. As discussed in (1), this property can be shown through $\lim_{n\to\infty} \varepsilon^* = 0$.

The self-adaptive property refers to the property that the trial conclusion does not depends upon the change in trial parameters. Any clinical trial has to have certain initial trial parameter assumptions, the trial conclusions can be different due to these assumptions. If the trial conclusion is sensitive to these assumptions, the trial conclusion can be challenged. This property is shown through the function of ε^*, n, and p, it is also discussed in (1). Besides, the likelihood ratio test requires that $\hat{p}$ has to be the maximum likelihood estimate of p, it is also shown in (1).

The CI based clinical trial design does not first two properties, so the trial practice adopted some additional limitations not from the statistical analysis principle. For example, the trial sponsor is not allowed to increase the sample size arbitrarily. The first two properties are advantages of the model based clinical trial compared to the CI based clinical trial design methodology.

The foundation of the model based clinical trial

In this manuscript, the importance of the likelihood ratio test is stressed. In the statistical analysis, the Neyman-Pearson lemma, (5), has shown that the likelihood ratio test is the uniformly most powerful statistical test among all the statistical hypothesis test. On the other hand, it is a natural link between the conventional statistical analysis and the model based statistical analysis. There is a school of statisticians concentrated on the model based statistical analysis, such as information theoretical model selection methodologies. These methodologies are built on the same foundations of the statistical axiom systems, but they were developed in different path. For example, the basic philosophy in model selection is to choose a best model out of a large model selection, the best model does not guarantee to be a "good" model, there is no concept on the confidence. It is the reason that the model selection methodology has not been widely used in the clinical trial practice.

To introduce the model selection methodology into the clinical trial analysis, it is necessary to link the model based analysis and the conventional analysis through a bridge. Strictly speaking, the likelihood ratio test and the associated χ^2 distribution can only be vaguely called the model based statistical analysis, but it provides a natural common ground for both conventional and the model based analyses. In the other words, before the model based analysis is fully explored, the purpose of this manuscript is to built a bridge to link the two. More general model based clinical trial design methodologies will be introduced in a follow-up manuscript.

Conclusion

With CI based trial design, the NI trial is not able to show the NI in common sense. Similarly, the SUP trial is not able show SUP in common sense. When a SUP is claimed, a natural question is by how much the SM is superior to AC. It will lead to the conservative SUP trial design, since the strict SUP can only show that SM is superior to AC by 0. The key is that the definition of NI and SUP have not been accurately addressed, and its statistical inference implied many approximations in violation to the statistical analysis axiom systems. This manuscript introduced the basic idea of model based clinical trial design through the likelihood ratio test (LRT), but LRT is not the only tool for model based clinical trial design. The important of LRT is not only the ideal statistical properties such as the asymptotic and self-adaptive properties as the manuscript described, as well as the Neyman-Pearson lemma pointed out, it serves a key bridge between the general model based statistical analysis with the conventional statistical analysis. It will serve as a foundation for further discussion.

References

1. Chen X (2013) A non-inferiority trial design without need for a conventional margin. Journal of Mathematics and System Science 3: 47-54

2. Chen X (2013) The generalized clinical trial outcomes. Journal of Mathematics and System Science 3: 167-172.

3. EMA (2000) Points to consider on Switching between superiority and non-inferiority.

4. FDA (2016) Non-Inferiority Clinical Trials to Establish Effectiveness.

5. Hogg Craig (2004) Introduction to Mathematical Statistics, Pearson Education.

6. Kolmogorov AN (1933) Grundbegriffe der Wahrscheinlichkeitrechnung, Ergebnisse Der Mathematik; translated as Foundations of Probability. New York: Chelsea Publishing Company, USA

7. McCullagh P (2002) What is a statistical model. The Annals of Statistics 30(5): 1225-1310.