PharmaSUG 2019 - Paper SA-91

Equivalence, Superiority and Non-inferiority with Classical Statistical Tests: Implementation and Interpretation

Marina Komaroff, Noven Pharmaceuticals, Inc., Jersey City, NJ

ABSTRACT

Determining superiority of an experimental treatment versus a standard of care has been a popular objective of randomized controlled trial in the pharmaceutical industry. Superiority is determined by the statistical and clinical significance of a clinical endpoint. However, researchers often question why non-significant p-values cannot be viewed as evidence that the two treatments were equivalent. They might say: if the p-value < 0.05 we assume the null hypothesis is false, so it must also follow that if the null hypothesis is true, the p-value must be ≥ 0.05. It is not clear whether this reasoning is a poor attempt to salvage a failed study or reflects a misunderstanding of null hypothesis testing. This paper will clarify the meaning of p-values and demonstrate how p-values can be used to conclude equivalence (no difference) in treatment effects.

There are many papers that review statistical methods for analyzing equivalence, superiority and non-inferiority trials utilizing the POWER, TTEST, TOST and FREQ procedures in SAS/STAT® software. Nonetheless, subtle and confusing issues arise in the application and interpretation of such methods. The author will present simulated data to visually demonstrate how changes in boundary margins affect sample size and power calculations. The goal is to help researchers thoughtfully choose boundary parameters, plan the operating characteristics of an equivalence/inferiority clinical trial, and correctly interpret the results. The paper will provide guidance for not only implementing such studies, but also promotes better understanding of the designs for critically reviewing the published research that utilized such methods.

INTRODUCTION

Often the result from clinical trial is not statistically significant due to small sample size. Nevertheless, it does not mean the absence of beneficial effect of the tested treatment. On the other hand, clinical study with a large sample size may result in statistically significant but clinically meaningless result. A careful choice of study design and operational parameters will increase a chance for success in clinical trial.

Clinical **superiority trials** aim to demonstrate that new treatment is better (superior) than control. There are four types of concurrently controlled trials that provide evidence of effectiveness: placebo, no treatment, dose-response and active control [FDA's regulations: 21 CFR 314.126]. There are other designs that should be used in cases: (1) where the use of placebo control is inappropriate [International Conference on Harmonization (ICH) guidance E10: Choice of Control Group and Related Issues in Clinical Trials (ICH E10)], and (2) where information is needed to support comparative effectiveness of the treatments.[1-5]

Clinical **non-inferiority trails** aim to demonstrate that new treatment is not worse (equivalent or better) than the control with an acceptable level to be worse called *margin of non-inferiority* (MNI). It implies that the interest is one-sided, and new treatment could be even better (superior) than control.[1-5]

Clinical **equivalence trials** are active control trials with the objective to demonstrate that new treatment is equivalent to active control with an acceptable level called *margin of equivalence* (ME).[1-5]

APPLICATIONS

For this paper, the research question is to determine if new test (T) treatment for pain differs from active control (C) treatment. Pain is measured by the score on the digital visual analog scale (VAS) from 0 (no pain) to 10 (strongest). Assessments are performed at baseline and week 1, and the endpoint is the reduction in pain from baseline to week 1, adjusted for baseline. The (T) treatment group is considered to be "better" if group mean is greater than the mean of C group: $\mu_t > \mu_c$, or if $\Delta = \mu_t - \mu_c > 0$. The number of

subjects in T and C treatment group is n_t and n_c, respectively. The descriptive statistics for simulated data set named MYDS is presented in Output 1. The reduction in pain units for T and C group on average was 2.4 (standard deviation = 1.2) and 5 (standard deviation = 1.5), respectively.

Output 1: Descriptive statistics for data set MYDS

treatment	N	Mean	Std Dev	Std Err	Minimum	Maximum
Test	82	2.444	1.218	0.135	0.021	7.493
Control	84	5.032	1.464	0.160	0.410	10.533
Diff (1-2)		-2.588	1.348	0.209		

For practice purposes, the first task is to test if experimental treatment is superior to the control. Note, if there is no statistical significance, we cannot conclude that treatments are "similar" or "equivalent". In other words, the statistical demonstration of "no effect" is not the same as demonstrating "the absence of clinically meaningful effect". The dilemma can be resolved by conducting an independent study of non-inferiority and test if new treatment is at least "not worse" than control with pre-defined margin of non-inferiority. Another equivalence study can be conducted to test if treatment groups are "equivalent" with a priori defined margin of equivalence. Non-inferiority and equivalence studies are valuable if new treatment has other preferable qualities like better safety profile, and/or lower cost.

Three following examples explain the hypotheses setting for superiority, non-inferiority and equivalence study designs. The meaning of p-values for each example is explained. The clarification of how these p-values should be used for conclusions is provided.

EXAMPLE #1: Superiority

1A. In **traditional comparative study** the hypotheses are to determine the difference between treatment groups.

H_0: $\mu_t - \mu_c = 0$ (there is no difference between treatment groups)

H_a: $\mu_t - \mu_c \neq 0$ (there is a difference between treatment groups).

The null hypothesis is rejected based on the two group 2-sided t-test (α=0.05) when the p-value < 0.05.

SAS code for two-tailed t-test (α=0.05) for MYDS data set along with the outputs is provided in BOX # 1. T-values fall further away from critical values: t_{lower}= -12.37 < -1.97, and t_{upper} = 12.37 > 1.97 that corresponded to p-value <0.05 (Output 2). Thus, the hull hypothesis is rejected and conclusion is that two treatments are different. However, this test does not tell us if new treatment has any clinically meaningful effect and how strong it is compare to C. Is T superior than C?

1B. In **superiority study**, we refine the alternative hypothesis showing T is better than C, what indicates one-side statistical t-test.

H_0: $\mu_t - \mu_c = 0$

H_a: $\mu_t - \mu_c > 0$ (the beneficial response from T is greater than C).

SAS code for right-tailed t-test (α=0.05) for MYDS data set along with the outputs is provided in BOX # 2. T-value is much smaller than critical value: t = -12.37 < 1.65 that corresponded to p-value >0.05 (Output 4). Thus, the hull hypothesis is not rejected and conclusion is that T is not superior than C.

Note, for our example in 1A and 1B, the 95% CI for difference between means support the conclusions based on p-values by not including "0" (Output 3**)** and including "0" (Output 5), respectively.

If T is not superior to C, does T have any clinical benefit?

BOX # 1: Traditional Comparative

SAS code: Two-Tailed t-test for $\Delta = 0$, where $\Delta = \mu_t - \mu_c$

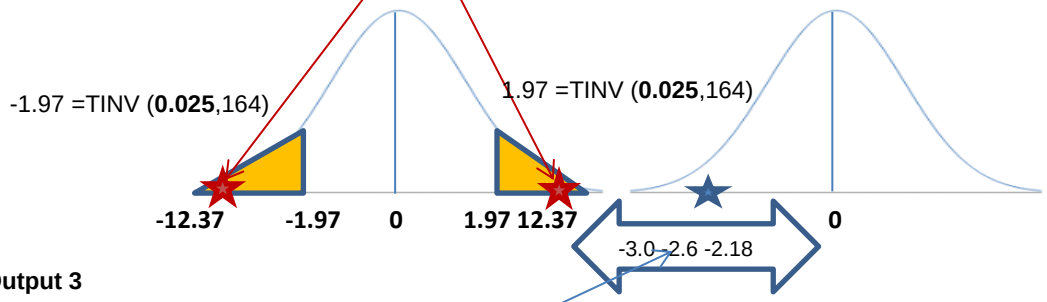
$$t = \frac{\bar{x}_t - \bar{x}_c - (\mu_t - \mu_c)}{\sqrt{\frac{\sigma_t}{n_t} + \frac{\sigma_c}{n_c}}}$$

```
proc ttest data=myds;
class treatment;
var es;
run;
```

Output 2

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	164	-12.37	<.0001

Reject Null, p-value < 0.05



Output 3

treatment	Method	Mean	95% CL Mean	
Diff (1-2)	**Pooled**	-2.588	-3.001	-2.175

Reject Null, CI does not include "0"

BOX # 2: Superiority

SAS code: Right-Tailed t-test for $\Delta = 0$, where $\Delta = \mu_t - \mu_c$

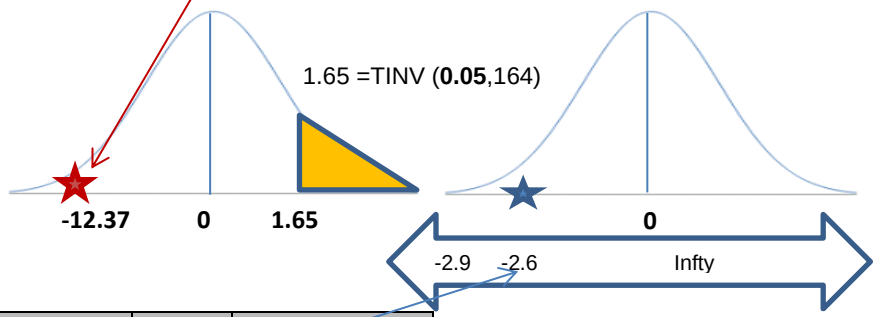
```
proc ttest data=myds CI=None Slides=U;
class treatment;
var es;
run;
```

$$t = \frac{\bar{x}_t - \bar{x}_c - (\mu_t - \mu_c)}{\sqrt{\frac{\sigma_t}{n_t} + \frac{\sigma_c}{n_c}}}$$

Output 4

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	164	-12.37	1.0000

Do not reject Null, p-value ≥ 0.05



Output 5

treatment	Method	Mean	95% CL Mean	
Diff (1-2)	**Pooled**	-2.588	-2.934	Infty

Do not Reject Null, CI includes "0"

EXAMPLE #2: Non-Inferiority

There still might be some clinical benefit if T is worse than C only by some acceptable margin. Let's define non-inferiority margin as 3 points, and test if T is non-inferior than C by 3 points or less.

H_0: $\mu_t - \mu_c = -3 \Rightarrow H_0: \mu_t - \mu_c + 3 = 0$

H_a: $\mu_t - \mu_c + 3 > 0$ (the beneficial response from Tis worse than C by 3 points or less).

The statistical test is similar to the last example of superiority where right-tailed test should be applied.

SAS code for right-tailed t-test (α=0.05) for MYDS data set along with the outputs is provided in BOX # 3. T-value is greater than critical value: t = 1.97 > 1.65 that corresponded to p-value < 0.05 (Output 6). Thus, the hull hypothesis is rejected and conclusion is that T is non-inferior than C with MIN=3.

Non-inferiority means not worse. How about independent study to test if T is equivalent to C at the pre-defined margin of equivalence ME=3?

BOX # 3: Non-inferiority

SAS code: Right-Tailed t-test for Δ_{ni} =0, where $\Delta_{ni} = \mu_t - \mu_c + 3$

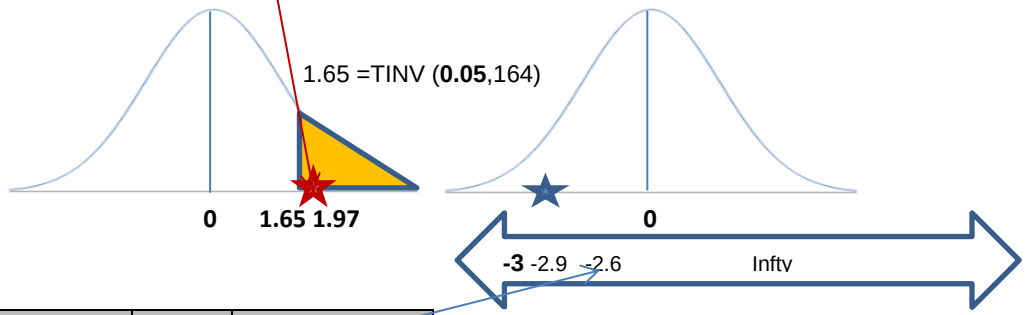
$$t = \frac{\bar{x}_t - \bar{x}_c - (\Delta)}{\sqrt{\frac{\sigma_t}{n_t} + \frac{\sigma_c}{n_c}}}$$

```
proc ttest data=myds CI=None Slides=U Ho=-3;
class treatment;
var es;
run;
```

Output 6

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	164	1.97	0.0253

Reject Null, p-value < 0.05



Output 7

treatment	Method	Mean	95% CL Mean	
Diff (1-2)	**Pooled**	-2.588	-2.934	Infty

Reject Null, Lower 95% CI > -3

EXAMPLE #3: Equivalence

Equivalence testing originates from the testing bioequivalence using pharmacokinetics parameters between new and existing drug. A statistical approach is the "two 1-sided" (TOST) testing [Schuirmann, 1987] for the two composite null hypotheses are tested with pre-specified upper and lower equivalence margins.

In this paper, a pre-defined margin of equivalence ME = 3, where ME_{lower} is the lower bound and ME_{upper} is an upper bound of equivalence set up to -3 and 3, respectively.

H_0: $|\mu_t - \mu_c| = 3$; or H_{o1}: $\mu_t - \mu_c + 3 \leq 0$ and H_{02}: $\mu_t - \mu_c - 3 \geq 0$

H_a: $|\mu_t - \mu_c| < 3$; or H_{a1}: $\mu_t - \mu_c + 3 > 0$ and H_{a2}: $\mu_t - \mu_c - 3 < 0$

When both H_{01} and H_{02} are rejected, then equivalence is concluded which means indicates that the effect falls within the pre-specified margins of equivalence.

BOX # 4: Equivalence

SAS code: Two 1-sided t-tests for $|\Delta_e| = 0$, where 1) $\Delta_{e1} = \mu_t - \mu_c + 3$ and 2) $\Delta_{e2} = \mu_t - \mu_c - 3$

$$t_1 = \frac{\bar{x}_t - \bar{x}_c - (\mu_t - \mu_c - ME_{lower})}{\sqrt{\frac{\sigma_t}{n_t} + \frac{\sigma_c}{n_c}}}$$

```
proc ttest data=myds tost (-3, 3);
class treatment;
var es;
run;
```

Output 8

Method	Variances	Test	Null	DF	t Value	P-Value
Pooled	Equal	Upper	**-3**	164	1.97	**0.0253**
Pooled	Equal	Lower	**3**	164	-26.70	**<.0001**

$$t_2 = \frac{\bar{x}_t - \bar{x}_c - (\mu_t - \mu_c - ME_{upper})}{\sqrt{\frac{\sigma_t}{n_t} + \frac{\sigma_c}{n_c}}}$$

Output 9

treatment	Method	Mean	Lower Bound		90% CL Mean			Upper Bound	Assessment
Diff (1-2)	**Pooled**	-2.588	-3	<	**-2.934**	**-2.242**	<	3	**Equivalent**

SAS code for two 1-sided t-test (TOST, each with α=0.05) for MYDS data set along with the outputs is provided in BOX # 4.

T-value is greater than critical value: $t_{upper} = 1.97 > 1.65$ for the right-tail test, smaller than the critical value $t_{lower} = -26.7 < -1.65$ for the left-tail test, that corresponded to p-value <0.05 for both tests (Output 8). Thus, the hull hypothesis is rejected and conclusion is that T is equivalent to C with ME=3.

These results are supported by the two 1-sided tests for difference between means: (1) the lower level of 95% CI (-2.9, +∞) is greater than a lower bound of equivalence -3 (Output 7), and (2) the upper level of 95% CI (-∞, -2.2) is below the upper bound 3 of equivalence. This is the same as the 90% CI for difference between means falls entirely within the margins of equivalence (Output 9).

There are some relations between the study designs: superiority might be a special case of non-inferiority, and equivalence is the combination of two non-interiority trials.

The important question remains about the rationale of choosing ME and MNI = 3. If the margins are more conservative (smaller) or less conservative (bigger) then results and conclusions for presented examples will be different.

CHOICE OF MARGINS

The choice of margin is very important; margins have clinical meaning and the choice should be supported by the rationale. Daniel Lakens recommends using equivalence margin based on the effect sizes that would make it possible to conclude if an effect is too small to be meaningless for anyone.[2] Meanwhile, pharmaceutical industry follows guidance from the agencies how ME or MNI must be defined.[5]

In this paper, the author demonstrates how choice of ME & MNI influence sample size and power of the study. Sample sizes depend on: (1) true difference in responses Δ (under the null $\Delta = 0$), (2) variances σ_t and σ_c, (3) a level of significance α (type I error), and (4) the power which equals to $1 - \beta$ (type II error).

SAS Code: Non-Inferiority

```
proc power ;
 twosamplemeans test=diff
  nulldiff = -3 -2.9 -2.8 -2.7 -2.6
  meandiff =-2.588
  sides =U
  alpha=0.05
  stddev=1.348
  npergroup=.
  power =0.80 0.90;
run;
```

SAS Code: Equivalence

```
proc power ;
 twosamplemeans test=equiv_diff
  lower=-3
  upper=3
  meandiff =-2.588
  alpha=0.1
  stddev=1.348
  npergroup=.
  power =0.80 0.90;
run;
```

SAS code for power calculation for non-inferiority and equivalence study designs was applied to the information from MYDS data. For both study designs, sample sizes increase when margin gets smaller and/or power get higher (Table 1). However, more subjects will be required for non-inferiority study compared to equivalence.

Table 1: Power and Sample Size in Non-inferiority and Equivalence study

Power	Margin	Non-Inferiority Sample size per group	Equivalence Sample size per group
0.80	3.0	134	97
0.80	2.9	232	169
0.80	2.8	501	365
0.80	2.7	1792	1307
0.80	2.6	156034	113768
0.90	3.0	185	142
0.90	2.9	321	246
0.90	2.8	694	532
0.90	2.7	2482	1904
0.90	2.6	216132	165799

CONCLUSIONS

This paper clarified the hypotheses and the meaning of p-values for superiority, non-inferiority and equivalence studies. The application and interpretation of such studies was presented for comparing means for two treatment groups on simulated data. Nevertheless, the same logic should be used for other types of data, and helpful materials about the SAS code can be found in **[1]**.

The importance of clinically correct choices for non-inferiority margins was underscored. For additional details and examples, the author refers to the recent FDA guidance where appropriate methods and rationale is provided.[5]

REFERENCES

[1] John Castelloe and Donna Watts; Equivalence and Noninferiority Testing Using SAS/STAT® Software SAS Institute Inc. Paper SAS1911-2015; https://support.sas.com/resources/papers/proceedings15/SAS1911-2015.pdf

[2] Danie¨ l Lakens; Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses; Social Psychological and Personality Science; 2017, Vol. 8(4) 355-362

[3] Lesaffre E. Use and misuse of the p-value. Bull NYU Hosp Jt Dis. 2008;66(2):146-9.

[4] Lesaffre E. Superiority, Equivalence, and Non-inferiority Trials; Bulletin of the NYU Hospital for Joint Diseases 2008;66(2):150-4

[5] FDA Guidance for Industry: Non-Inferiority Clinical Trials to Establish Effectiveness; November, 2016

CONTACT INFORMATION

Marina Komaroff, Dr.P.H., M.P.H., M.S.
Director – Biometrics
Product Development/Clinical Opertations/Regulatory
Noven Pharmaceuticals, Inc.
100 Town Squre Place, 5th Floor
Jersey City, NJ 07310
Tel: 551-233-2645
Email: Mkomaroff@noven.com