Biostatistics BootCamp Lecture 2:

Statistical Considerations in Protocol Development: Study Design

August 16, 2023

Wonsuk Yoo, PhD

Dept. of Translational Neuroscience
wonsuk.yoo@barrowneuro.org

Ivy Brain Tumor Center | Barrow Neurological Institute

Topics in Study Design & Methods (Lec.2):

- Conditions for a good study design
- Hypothesis: Superior tests .vs. Non-inferior tests
- Hypothesis: Two-sided tests .vs. One-sided tests
- Design issues with primary endpoint(s)
- Design issues with comparable groups
- Design issues with single-ram and two-arms trials
- Multiplicity design issue
- Final comments and questions

Quick Review of Lecture 1:

- Let’s assume we are planning a clinical research about a new treatment affecting a biomarker on Parkinson’s disease (PD). We are interested in the level of the biomarker
- Since it is not usually feasible to inspect a total population, we do not the means of two groups from population.
- Statistics uses sample data to determine population parameters. Therefore, experimental design and statistical inference processes are required.
- The uncertainty is denoted by the statistical errors and quantified by p-value and confidence intervals in the hypothesis testing procedure.
- A good design through hypothesis tests is related to the statistical uncertainty, the two types of errors, α and β.

Conditions for A Good Study Design:

- During hypothesis testing, we set $\alpha \leq= 0.05$ (fixed) and minimize β in practice. Thus, conditions for a good experimental design are the study designs that (1) **controlling $\boldsymbol{\alpha}$ up to 5%** and (2) **maximizing the study power (1-$\boldsymbol{\beta}$)**.
- Now what scenarios should we avoid in the hypothesis testing?
  - ➡ [1] Study designs inflating the type 1 error: $\alpha > 0.05$
  - ➡ [2] Study designs leading to low study power
- FDA Guideline (2017) on Multiplicity

> A focus of this guidance is **control of the Type I error rate** for the **prespecified** set of **endpoints** (i.e., primary and secondary endpoints) of a clinical trial **to ensure that the major findings** of a clinical trial are well supported.

Hypothesis: Superior .vs. Non-inferior test (I):

- A **superior test** has the primary objective to show that the investigational efficacy is superior to a comparative efficacy – ICH E9 1998
  Let $\boldsymbol{\mu}$ and $\boldsymbol{X}$ be population mean and sample mean respectively.

  Q1. Which one is right?

  $[1]\ H_0\colon \mu_T = \mu_c$ vs $H_1\colon\ \mu_T \neq \mu_c$ $\quad [2]\ H_0\colon\ X_T = X_c$ vs $H_1\colon\ X_T \neq X_c$

  Q2. Can we use one directional hypothesis, $H_1\colon\ \mu_T > \mu_c$?

- Most of intervention or clinical trials belong to superior test.

Hypothesis: Superior .vs. Non-inferior test (II):

- A **non-inferior test** wants to show that a new treatment is no less effective than an existing treatment (Δ: equivalent margin) – ICH E9 1998

$$H_0\colon \mu_T \le \mu_c - \Delta \quad \text{vs} \quad H_1\colon \mu_T > \mu_c - \Delta$$

  ➡ FDA approved 18 of the 43 NDAs based on non-inferior trials as of Dec. 2009.

- When can we consider the non-inferiority tests?

  ➡ Similar efficacy but some other advantages over existing treatment.

- What is the issue of non-inferior tests? How can we determine the equivalent margin?

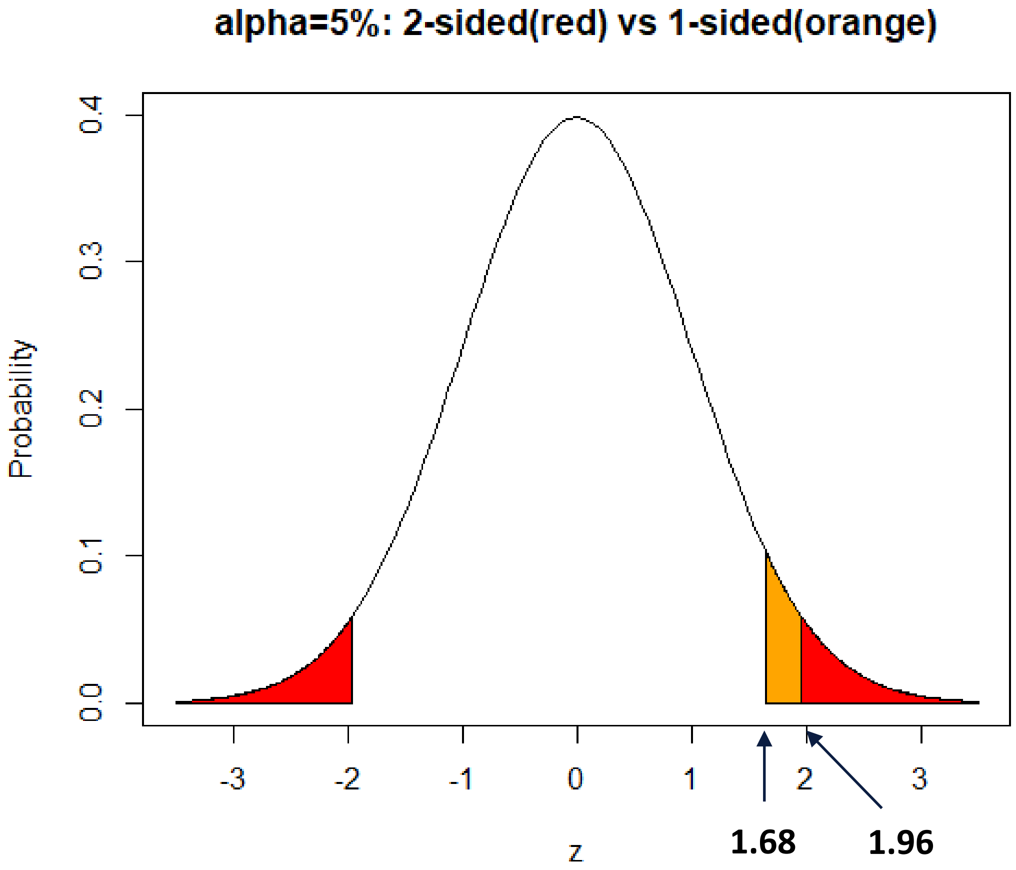
Hypothesis: Two-sided .vs. One-sided test (I)

- Alternative hypotheses:

Two-sided — One-sided

$H_1: \mu_T \neq \mu_c$ .vs. $H_1: \mu_T > \mu_c$

- One-sided test has 5% significance level for upper direction (orange). Two-sided test has 2.5% for each direction (red).
- Which test is more advantage in hypothesis test to reject the null hypothesis?



Hypothesis: Two-sided .vs. One-sided test (II)

- The selection of one-sided or two-sided inference is controversial, and a diversity of views can be found in the statistical literature.
- In superiority trials, it is recommended to use two-sided tests with 5% significance level without specific reasons.
- In non-inferiority, the recommended approach is to pre-specify a margin of non-inferiority in the protocol, and 2.5% significance level should be applied.
- Dubey (1991) indicated that the FDA prefers a two-sided test over a one-sided test procedure in clinical research

Primary endpoint (I): Why Important?

- The primary endpoint should be the variable of the most clinically relevant and directly related to the primary objective of the trial.
- Why does the primary endpoint be important? Because it affect the study design and analysis methods.
- Power analysis (and sample size calculation) should be performed based on the primary endpoint.
- The number of the primary outcome affects the error rate of hypothesis testing procedure to test the study aim.
- The data types for the primary outcome also affect the analysis method (regression models).

Primary endpoint (II): Protection against “α”

- A primary outcome needs to be defined a priori to protect against the risk of a false-positive error.
- There should generally be only one primary variable. If you have multiple primary outcomes in your protocol, what issues/problems occur?

  ➡ Multiple outcomes means multiple tests, which inflate the type 1 error rate. How much inflated? The inflated type 1 error (α) for k statistical tests is $(1 – 0.95^{k})$.

  So, α = 0.098 for k=2, α = 0.19 for k=4, and α = 0.40 for k=10.
- When the trial has multiple outcomes, we says the multiplicity issue occurs. Therefore, the study should be designed to control the type 1 error up to 5%.

Primary endpoint (III): Protection against “β”

- A study is conducted to answer the research question, whether a particular treatment is effective for a particular purpose.
- If the sample size is small and the study fails to show that the treatment is effective, then one of two explanation is possible:

  1. The treatment is truly ineffective, or
  2. The treatment is effective, but the study failed to identify a statistically significant advantage because the sample size was too small.
- Studies may fail because of inadequate sample size are wasteful.
- Investigators need to a priori estimate the minimum sample size required to answer a particular research question ➡ at most β i.e., at least power (1- β).

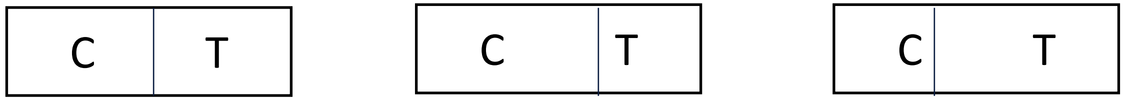
Design: How to select the controls?

- What are control groups? It is not exposed to the treatment and may receive a placebo treatment.
- Why does a control group be important? Even though the control group does not receive treatment, it does play a critical role in the experimental process. How? Since participants have been randomly assigned to either group, it can be assumed that the groups are comparable and the difference is assumed to result from the treatment difference.
- How does it be used in hypothesis testing? It is used to provide a baseline to compare results against the experimental group.

Design issue: Two-arm trials

- What is the ideal ratio in subjects between treatment and control groups in two-arm trials with given sample size?

  ➡ If we have 100 patients, which one has more power of the two-arm:



(a) 50:50 (b) 70:30 (c) 30:70

- What happens if we have more than two treatment groups?

  ➡ If we have four treatment groups, how many tests do we need?

Design issue: Single-arm trials

- Single-arm studies have been traditionally used in Phase II oncology clinical trials with binary endpoints as primary outcome.
- How do the single-arm trials evaluate the efficacy of experimental treatment?

  ➡ The historical control data (H_0: $p_E = p_{hc}$) have been applied.
- When we borrow the historical control information from other published studies, there exist heterogeneities (variabilities among rates). The variabilities substantially inflates the Type 1 error rate or false positive error rate and may lead to erroneous conclusions.
- Guideline: We aim to avoid overly optimistic expectation on their trials.

Determination of Two Errors (α, β):

- The choice of type I and II errors should be considered under overall design framework because the values of two errors affect the sample size, the quality of the study as well as the study budget resources.
- And the values of two types of error must be clearly stated in the protocol.
- <u>Conventional rule</u>: 5% and 20% as their type I and II errors respectively.
- How can we determine the error rates for studies little harm and might be good to public health (benefits)?
- How about studies related toxicity or with serious side effects?.

Multiplicity (I): When we encounter multiplicity?

- We need to perform multiple hypothesis tests for multiple primary endpoints ➡ Multiple primary endpoints

- We need to perform multiple ad-hoc tests for more than 2 treatment groups tiple (more than 2) treatment groups

  ➡ Multiple comparison tests

- We need to test multiple tests If the experiment be scheduled to test the interim efficacy during the experiment ➡ We have an interim (efficacy) test.

- When do we encounter the multiplicity (ICH E9)?

> When multiplicity is present, the usual statistical approach may necessitate an adjustment to the type 1 error. Multiplicity may arise, for example, from multiple primary variables, multiple comparisons of treatments, repeated evaluation over time, and/or interim analyses.

Multiplicity (II): What multiplicity results from?

- The multiplicity (multiple tests) leads to the inflation of type 1 error rate.
  - ➡ For k statistical tests, the inflated type 1 error (α) = ($1 – 0.95^k$).
    So, α = 0.098 for k=2, α = 0.19 for k=4, and α = 0.40 for k=10.
  - ➡ This is a violation of the prespecified significance level.
- FDA Guideline (2017) on Multiplicity:

> The multiplicity of analyses may cause inflation of the Type I error rate. Hence, by inflating the Type I error rate, multiplicity produces uncertainty in interpretation of the study results such that the conclusions about whether effectiveness has been demonstrated in the study become unreliable.

Multiplicity (III): How can we overcome?

- FDA Guideline (2017) on Multiplicity

> There are various approaches that can be planned prospectively and applied to maintain the overall Type I error rate at 2.5% or below.

- Bonferroni correction: α/k

  ➡ Adv: (1) Easy to use, (2) always controlled under 5%

  ➡ Weak: (1) low actual α (conservative), (2) increased β => decreased power

- Various methods invented to increase power rather than Bonferroni.
- For example, Holm method, Hochberg method, …

My suggestions for a good study design:

- We have usually used the conventional rule (5% and 20% for type 1 and 2 error rates respectively) for confirmatory studies. How can we determine the type 1 & 2 error levels for early efficacy trials or exploratory studies?

  ➡ For confirmatory tests, $\alpha = 0.05$ and $\beta = 0.2$. ex. Rubinstein (2005)

  ➡ My suggestion:

- How can we determine the error rates for studies little harm and might be good to public health (benefits)? How about studies related toxicity or with serious side effects?

- What is the better combination(s) of two error rates?

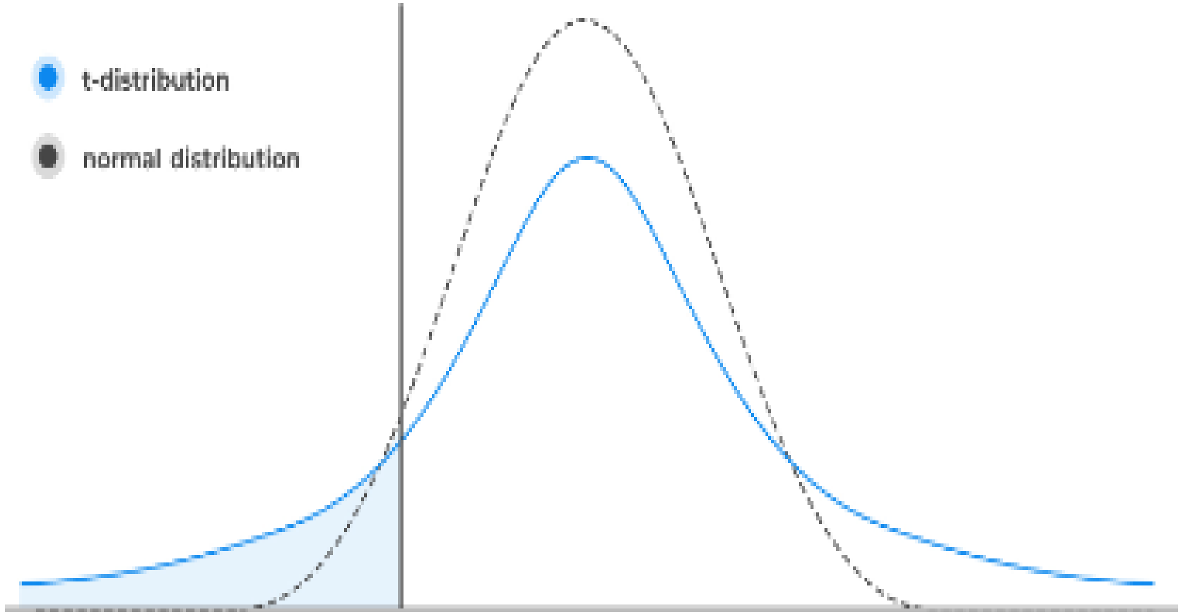
  ➡ α=0.05 & β=0.2 or α=0.1 & β=0.1 ?

My suggestions for a good study design (2):

- The preliminary data on our new treatment have shown superior to those from other existing treatments. Can we apply one-sided hypothesis rather than two-sided hypothesis?
- Let’s consider a two-arm trial for rare diseases and it has small sample size (N=25). Can we increase the statistical power? Is so, how?
- If your research needs more than one primary endpoint, what design components should be considered for a good study design?
- Two primary endpoints .vs. co-primary endpoints

Thought 1: Normal .vs. t-distribution

The critical value at 97.5 percentile in normal distribution is $Z_{0.975}$=**1.96** while the critical value at 97.5 percentile in t-distribution is $t_{0.975}$ = **2.23**. Even though the t-distribution seems to be similar to the normal for more than 30 samples, why do these two numbers be different?



Thought 2:

The confidence interval (CI) conveys more useful information that p-values. The CIs provide the magnitude of the true difference of interest scientifically or clinically. How can we the range/interval in the magnitude of the true effect size (i.e., mean difference of two groups) interpret?

Thought 3:

Our goal in statistical inference is minimize the two types of errors. But it's impossible to control two errors simultaneously because the two errors (α and β) have the reverse relationship. How can we construct a clinical trial with good study designs?

Q&A