Primary Question and Hypothesis Testing in Randomized Controlled Clinical Trials

Jiangtao LUO

Summary: We briefly reviewed and provided cautions about some of the fundamental concepts used in the design of medical and public studies, especially primary question, hypothesis testing and sample size in this short note. We also talked about some of the extensions and development in the recent years.

Keywords: hypothesis, randomize, clinical trial, primary question

1. Primary and secondary questions

Medical research is always about answering scientific questions related to treatments or preventions of diseases in special population of patients, who are defined by inclusion and exclusion criteria and treatments or preventions are performed among them. These questions are often divided into primary and secondary questions. Usually we have only one primary question in a study and it is the key and central question that we want to answer. We should state our primary question in advance rather than define it during our trial process later. Our study should be designed to answer the primary question and our sample size calculation should be based on it. It is unacceptable if a proposal lacks clear statement about primary question. Some research questions are important in public health and clinical practice, but we cannot answer them all due to limited technology and resources. So answerable is another criterion in choosing primary question. Therefore, at the design stage we must choose a primary question that we are able to answer via clinical trials. For example, in a study on ‘Ibrutinib as Initial Therapy for Patients with Chronic Lymphocytic Leukemia’ [1], the primary question is to find the treatment difference between ibrutinib and chlorambucil for ‘previously untreated older patients with CLL or small lymphocytic lymphoma’. All the other research questions are secondary.

Secondary questions are closely related to the primary question. A study may have several secondary questions, which must be stated in advance or at design stage. We do not recommend mining data after the trial although data mining has become increasingly important in other settings. The number of secondary questions should be limited as well. Otherwise we may not have sufficient power to answer truly important questions and effectively control false positive rate. In Physicians’ Health Study [2], the primary question is total mortality rate between aspirin and placebo groups. The secondary question is fatal and nonfatal myocardial infarction.

Once the primary question has been determined, the next step is to define the primary outcome that is used to describe the primary question. We must pay special attention to primary outcome and it must be evaluated in each of all study participants in the same way. Also the evaluation must be unbiased. Our recommendation is to use double blind, hard endpoint, and independent assessment. All study participants should have primary outcomes when the study ends [3, 4].

2. Hypothesis testing

The methods for answering scientific questions from data collected in clinical trials belong to statistical inference. An important part of statistical inference
is hypothesis testing, the foundation of which was laid by Fisher, Neyman, and Pearson among others \cite{9}. Hypotheses consist of null hypothesis (H\textsubscript{0}) and alternative hypothesis (H\textsubscript{1}). The H\textsubscript{1} is our scientific hypothesis, which is what we want to collect data for. To test if H\textsubscript{1} is true, we start from H\textsubscript{0}, our straw man, since we usually already know from our pilot data, animal model or other approaches that H\textsubscript{1} is true. So we assume that H\textsubscript{0} is true and show that the probability that the observed data satisfies the null H\textsubscript{0} is very small (<0.05, usually). Thus, the null H\textsubscript{0} is opposite to hypothesis (H\textsubscript{1}): µ\textsubscript{1} ≠ µ\textsubscript{2}.

The H\textsubscript{1} is our scientific hypothesis, which is what we want to collect data for. For the standard therapy vs. H\textsubscript{1}: cure rate for the experimental intervention is different from that of the experimental intervention.

If we use µ\textsubscript{1} and µ\textsubscript{2} to represent the cure rates of the standard therapy and the experimental intervention, the H\textsubscript{0} and H\textsubscript{1} can be quantitatively expressed as: Null hypothesis (H\textsubscript{0}): µ\textsubscript{1} - µ\textsubscript{2} = 0 (no difference) vs. Alternative hypothesis (H\textsubscript{1}): µ\textsubscript{1} - µ\textsubscript{2} ≠ 0 (difference)

Of course, our alternative hypothesis should be stated as µ\textsubscript{1}<µ\textsubscript{2}, if we are sure that µ\textsubscript{1}>µ\textsubscript{2} will not happen. Such one-sided alternative increases power for the same sample size. If we already have data, we can easily test the above hypotheses \cite{6}. Hypothesis testing in clinical trials is analogous to trials by juries in court cases in which the null hypothesis is that a defendant is presumed innocent, while the alternative hypothesis is that the defendant is proven guilty. Our inference of rejecting the null and accepting the alternative hypothesis at type I error α = 0.05 corresponds to a judicial decision that finds the defendant guilty beyond reasonable doubt. Likewise, a non-conviction decision frees the defendant, it does not imply that the defendant truly does not commit the prosecuted crime, but rather only indicates that there is not enough evidence for the conviction. Therefore, we have two types of errors in making the decision. The type I error or false positive alpha (α) refers to the error in convicting the innocent or, in a clinical trial, conclude the treatments differ, when, in fact, they are the same. The type II error or false negative beta (β) is the error in freeing the guilty or failing to conclude the treatments differ, when, in fact, they are different. In statistics, we often use 1-β, or statistical power, which is the probability for statistically detecting difference when true difference exists. We must control or minimize both types of errors or have small false positive rate and high power for a credible trial. High proportion of early published negative trials did not have adequate power due to small sample sizes and Freiman et al. \cite{7} showed that 50 of 71 negative trials that they had surveyed could miss a 50% benefit. We need a sample size large enough to have small false positive rate and high power in detecting the difference for the primary outcome.

4. Equivalent test and other extensions

Another popular type of hypothesis testing is equivalence. The new treatment may not be better than the standard, but offers other benefits, such as cost saving, fewer side effects, or easier to administer, etc. In such a study, the hypotheses are set as following:

Null hypothesis: Cure rate for new treatment is worse than that for standard therapy.

Alternative hypothesis: Cure rate for new treatment is “clinically equivalent” to that for standard therapy:

Null hypothesis: |µ\textsubscript{1} - µ\textsubscript{2}| > Δ vs Alternative hypothesis: |µ\textsubscript{1} - µ\textsubscript{2}| ≤ Δ

where Δ is a minimum clinically-meaningful difference. Note here that Δ is defined by clinicians according to their experience and not by any statistical methods. Please see \cite{10} for a recent example on equivalent test of thrombolytic therapy for acute ischemic stroke.

Readers are referred to \cite{11} for methods of sample size calculations for testing different kinds of hypotheses. Also see \cite{11} for advanced topics on testing statistical hypotheses.
In many studies, we often continuously monitor the trials for important differences in toxicity among treatments, for “definitive” differences in outcome among treatments, and/or for unlikely important differences. The goal of interim analyses is to assess the need to stop study early for severe side effects, efficacy, or futility, or to modify protocol (which belongs to adaptive design). Theoretically, we may have as many interim analyses as we want. But the more interim analyses we perform at say α = 0.05, the more likely we declare “statistical significance”, increasing the type I error. So we need to adjust our significance level according to the number of interim analyses performed. Please see [12] for statistical techniques on such adjustments. In practice, interim analyses are quite costly, with 2 or 3 interim analyses more popular than 5 or 6. Interim analyses are common for studies with high risk treatments such as cancer, especially in Children’s Oncology Group (COG) studies [13, 14].

In this report, we have discussed components of traditional clinical trials, in which primary question and hypotheses do not change during the process of the trial. But the scenario is different in modern adaptive designs. ‘An adaptive design is referred to as a clinical trial design that uses accumulating data to decide on how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial’ [15]. The FDA’s definition is ‘An adaptive design for a medical device clinical study is defined as a clinical trial design that allows for prospectively planned modifications based on accumulating study data without undermining the trial’s integrity and validity’ [16]. An adaptation is defined as a change or modification made to a clinical trial before and during the conduct of the study. Examples include relax inclusion/exclusion criteria, change study endpoints, modify dose and treatment duration, change primary question, change hypotheses etc. [17]. For Bayesian approach of adaptive design, please see [18].

Funding
Not available.

Conflict of interest statement
The authors report no conflict of interest

Acknowledgement
Thanks are given to the two anonymous editors for their valuable comments.
13. Malempati S, Hawkins DS. Rhabdomyosarcoma: Review of the Children’s Oncology Group (COG) Soft-Tissue Sarcoma Committee Experience and Rationale for Current COG Studies. *Pediatri Blood Cancer*. 2012; 59(1): 5–10. doi: http://dx.doi.org/10.1002/pbc.24118

14. Rudzinski ER, Anderson JR, Hawkins DS, Skapek SX, Parham DM, Teot LA. The World Health Organization Classification of Skeletal Muscle Tumors in Pediatric Rhabdomyosarcoma: A Report from the Children’s Oncology Group. *Arch Pathol Lab Med.* 2015; 139(10): 1281–1287. doi: http://dx.doi.org/10.5858/arpa.2014-0475-OA

15. Gallo P, Chow CS, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive designs in clinical drug development—an executive summary of the PhRMA working group. *J Biopharm Stat.* 2006; 16(3): 275-283

16. FDA. Adaptive Designs for Medical Device Clinical Studies: Draft Guidance. FDA; 2015 [cited 2016]. Available from: http://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm446729.pdf

17. Chow SC, Chang M. *Adaptive Design Methods in Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC; 2007

18. Berry SM, Carlin BP, Lee JJ, Muller P. *Bayesian Adaptive Design*. Boca Raton, FL: Chapman & Hall/CRC; 2010

---

Dr. Jiangtao Luo obtained his PhD degree from the University of Florida in 2009. He started to work at the University of Nebraska Medical Center College of Public Health, USA in 2010, and is currently an assistant professor in the Department of Biostatistics. His research interests include design of medical studies, statistical genetics, Bayesian method, numerical optimization, and big data.