Contemporary Clinical Trials Communications 7 (2017) 23–27

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Design of phase II non-inferiority trials

Sin-Ho Jung

Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

ARTICLE INFO

Keywords:
Binary outcome
Non-inferiority margin
Phase II/III trials
Randomized phase II trials
Sample size calculation
Survival outcome

ABSTRACT

With the development of inexpensive treatment regimens and less invasive surgical procedures, we are confronted with non-inferiority study objectives. A non-inferiority phase III trial requires a roughly four times larger sample size than that of a similar standard superiority trial. Because of the large required sample size, we often face feasibility issues to open a non-inferiority trial. Furthermore, due to lack of phase II non-inferiority trial design methods, we do not have an opportunity to investigate the efficacy of the experimental therapy through a phase II trial. As a result, we often fail to open a non-inferiority phase III trial and a large number of non-inferiority clinical questions still remain unanswered. In this paper, we want to develop some designs for non-inferiority randomized phase II trials with feasible sample sizes. At first, we review a design method for non-inferiority phase III trials. Subsequently, we propose three different designs for non-inferiority phase II trials that can be used under different settings. Each method is demonstrated with examples. Each of the proposed design methods is shown to require a reasonable sample size for non-inferiority phase II trials. The three different non-inferiority phase II trial designs are used under different settings, but require similar sample sizes that are typical for phase II trials.

1. Introduction

In a standard superiority trial, we usually want to show that an experimental therapy is more efficacious than a standard therapy, also called a control therapy. In this case, the null hypothesis is that the two therapies have equal efficacy and the alternative hypothesis is that the experimental therapy has higher efficacy than the standard therapy.

Sometimes, we may want to prove that an experimental therapy may not be better than, but not inferior to, a standard therapy. In this case, the experimental therapy is less extensive, less toxic or less expensive than the standard therapy, so that the former may be acceptable as far as there is evidence that its efficacy is not much worse than the latter. These types of studies are generally referred to as 'non-inferiority trials'. Due to the directional nature of the hypotheses involved, the statistical tests used for non-inferiority trials are one-sided. This aspect of testing distinguishes non-inferiority trials from 'equivalence trials', where the objective is to find evidence against "no difference". Sometimes, however, equivalence and non-inferiority have been used interchangeably. For example, Dunnet and Gent [1] and Mehta, Patel and Tsiatis [2] use the term of 'equivalence' for the 'non-inferiority' type problem. Durrleman and Simon [3] and Whitehead [4] review broad examples of non-inferiority trials and discuss sequential monitoring of such trials.

Analysis of a non-inferiority trials requires specification of a non-inferiority margin. A key component of the design for such a trial is to calculate a sample size required for a reasonable power, e.g. 80%–90%, to conclude that the regimens differ in efficacy by less than a specified non-inferiority margin Δ_0 when two arms have an identical efficacy. A typical non-inferiority margin is very small. Rothman et al. [5] propose to use a non-inferiority margin of about half of the superiority margin associated with the benefit of the standard therapy over no therapy. As a result, the required sample size for a non-inferiority trial becomes so big that a typical phase III non-inferiority trial requires a sample size that is often considered infeasible. Hence, in such cases it might be useful to conduct a smaller phase II trial to collect some evidence for potential non-inferiority of an experimental therapy before proceeding to a non-inferiority phase III trial.

Phase II trials are designed to screen out inefficacious experimental regimens before evaluating them through larger size phase III trials. Accordingly, phase II trials should be done quickly with small sample sizes. The sample size of a non-inferiority trial is determined by type I error rate, power and non-inferiority margin. So, we will have to compromise the level of these design components for a small non-inferiority phase II trial. In this paper, we propose design and analysis methods for non-inferiority phase II trials. We focus on randomized phase II trial designs here, but we could also consider corresponding single-arm phase II trial designs.

For an experimental regimen, its non-inferiority margin will be determined regardless of the phase of a trial. So, the only design components we can change for a small non-inferiority phase II trial are

E-mail address: sinho.jung@duke.edu.

http://dx.doi.org/10.1016/j.conctc.2017.04.008
Received 12 December 2016; Received in revised form 21 March 2017; Accepted 22 April 2017
Available online 05 May 2017

type I error rate and power. In this paper, we propose design and analysis methods for non-inferiority phase II trials. We focus on randomized phase II trial designs here, but we could also consider corresponding single-arm phase II trial designs.

2. Phase III non-inferiority trials: review (Test 1)

Let θ denote the parameter to measure the difference in efficacy between an experimental arm (arm 2) and a control arm (arm 1), such as log-odds ratio for binary outcomes and log-hazard ratio (log-HR) for survival outcomes. Without loss of generality, we assume that a large θ value means a higher efficacy for the experimental therapy under investigation and $\theta = 0$ means an equal efficacy between control and experimental therapies. Suppose that $\hat{\theta}$ is an efficient estimator of θ which is approximately $N(\theta, \sigma^2/n)$ for large n, the total number of patients. Most commonly used estimators satisfy this condition, e.g. the MLE. For a chosen non-inferiority margin $\delta_0(>0)$, we want to test H_0: $\theta \le -\delta_0$ vs. H_a: $\theta > -\delta_0$. The type I error rate of a non-inferior trial is the probability that we accept the experimental therapy when the measurement of its efficacy is at most θ_0. For a chosen type I error rate α, we reject H_0 if

$$\frac{\sqrt{n}\,(\hat{\theta} + \delta_0)}{\hat{\sigma}} > z_{1-\alpha}, \tag{1}$$

where $z_{1-\alpha}$ denotes the $100(1-\alpha)$ percentile of the standard normal distribution and $\hat{\sigma}^2$ denotes a consistent estimator of σ^2. Due to the directional hypothesis, we usually use a one-sided type I error rate in non-inferiority testing.

In a typical phase III non-inferiority trial, the highest expectation for the experimental therapy is to have an efficacy similar to that of the control therapy. In this sense, we set $\theta = 0$ under the alternative hypothesis for a power or sample size calculation. Then, the required sample size for a non-inferiority phase III trial is given as

$$n = \frac{\sigma^2 (z_{1-\alpha} + z_{1-\beta})^2}{\delta_0^2}. \tag{2}$$

In a phase III trial, we choose $\alpha = 0.025$ or 0.05, $1-\beta \approx 0.9$ and a very small δ_0, so that the resulting sample size is very large. Jung et al. [6] propose a sample size formula of the log-rank test for non-inferiority trials.

By assuming $\hat{\sigma}^2 \approx \sigma^2$ and plugging (2) in (1), the non-inferiority test will reject H_0 if

$$\hat{\theta} > -\frac{z_{1-\beta}}{z_{1-\alpha} + z_{1-\beta}}\delta_0. \tag{3}$$

Note that, with α and β smaller than 0.5, the rejection value on the right hand side of (3) lies in $(-\delta_0, 0)$, closer to $-\delta_0$ if $\alpha > \beta$ and closer to 0 if $\alpha < \beta$. Especially, if $\alpha = \beta$, then this rejection rule is simplified to $\hat{\theta} > -\delta_0/2$.

Example 1. Suppose that the progression-free survival (PFS) of the control treatment has a median of 2 years, which corresponds to an annual hazard rate of $\lambda_1 = 0.35$ under an exponential PFS distribution assumption. A hazard ratio (HR), Δ, smaller than $\Delta_0 = \lambda_1/\lambda_2 = 0.8$, or a log-HR of $-\delta_0 = -0.223$, is considered to be non-inferior, while the PFS of the experimental treatment is not expected to be longer than that of the control at best. We assume an accrual rate of $r = 200$ patients per year, additional follow-up period of $b = 2$ years and 1-to-1 allocation ($p_1 = p_2 = 1/2$). Then, for H_0: $\Delta = 0.8$ (or $\theta = -0.223$) and H_a: $\Delta = 1$ (or $\theta = 0$), the non-inferiority log-rank test of Jung et al. [6] with 1-sided $\alpha = 0.025$ requires $n = 1090$ patients ($D = 845$ events) for $1-\beta = 0.9$ by the sample size calculation method described in the Appendix. This sample size calculation is available from commercial softwares, such as East [7] or PASS [8], possibly using different approximations.

3. Design of phase II non-inferiority trials

We investigate design and statistical testing for non-inferiority phase II trials. In Sections 3.1 and 3.2, we consider the standard non-inferiority trials where the experimental regimens can not be more efficacious than the control regimens. In Section 3.3, we consider the case where an experimental therapy is acceptable as long as it is non-inferior to a control therapy, but the experimental therapy can possibly be slightly superior to the control therapy.

3.1. Based on point estimator (Test 2)

In order to lower the sample size of (2), the design parameters we can change are α, β and the non-inferiority margin. As an effort to lower the sample size for a phase II trial, we propose to lower the power $1-\beta$, say from 90% to 80%. Furthermore, we propose to reject H_0 if $\hat{\theta} > -\delta_0$ in this section. In other words, we accept the experimental therapy, or reject H_0: $\theta \le -\delta_0$, for further investigation when the observed efficacy measure $\hat{\theta}$ is larger than the specified non-inferiority margin, $-\delta_0$. This testing rule, called Test 2, corresponds to $\alpha = 0.5$ in (1). Note that, in an interim analysis testing for futility in a superior phase III trial, it is not unusual to reject the experimental therapy and stop the trial early if the p-value from the interim analysis is larger than or equal to 0.5 [9–11]. Regarding a phase II trial as an interim analysis in a large scale experiment including a phase II trial and a consecutive phase III trial, $\alpha = 0.5$ may be a reasonable choice for a phase II trial as an efficacy screening test before a large scale phase III trial.

In this case, it is easy to show that the required sample size of a non-inferiority phase II trial H_0: $\theta = -\delta_0$ against H_a: $\theta = 0$ is given as

$$n = \frac{\sigma^2 z_{1-\beta}^2}{\delta_0^2}. \tag{4}$$

Note that this formula corresponds to (2) with $\alpha = 0.5$ that gives $z_{1-\alpha} = 0$.

In order to investigate how successful our effort is to lower the sample size of a non-inferiority phase II trial, let $\alpha_l, 1-\beta_l$ and n_l denote the type I error rate, power and the required sample size, respectively, for phase $l\,(=II, III)$ trial. Then, from (2) and (4), we have

$$\frac{n_{II}}{n_{III}} = \left(\frac{z_{1-\beta_{II}}}{z_{1-\alpha_{III}} + z_{1-\beta_{III}}}\right)^2.$$

Suppose that we choose $1-\beta_{II} = 0.8$, 1-sided $\alpha_{III} = 0.025$ and $1-\beta_{III} = 0.9$. Then, we have $n_{II}/n_{III} = 0.0674$. That is, before we conduct a non-inferiority phase III trial requiring 1000 patients, we may conduct a non-inferiority phase II trial to collect some evidence on the experimental therapy with less than 70 patients. This comparison is based on the assumption that the phase II trial randomizes patients between the experimental arm and a prospective control arm. If we have reliable historical control data, then we can further reduce the sample size by using a single-arm phase II trial design.

Example 2. For a phase II trial, we assume the same design setting as in Example 1 except that 1-sided $\alpha = 0.5$. Then, Test 2 requires $n = 106$ patients ($D = 58$ events) for $1-\beta = 0.8$ and $n = 227$ ($D = 133$) for $1-\beta = 0.9$, compared to $n = 1090$ ($D = 845$) for a phase III trial as described in Example 1. Considering $\alpha = 0.5$ too liberal, one could choose a different α level. Table 1 lists the sample sizes for different $(\alpha, 1-\beta)$ values under the same design setting. We observe that the sample size easily goes beyond the feasibility for a phase II trial with an α level much smaller than 0.5.

3.2. Based on a futility test (Test 3)

For $\delta_1 (> \delta_0 > 0)$, suppose that δ_1 is a generally accepted superiority margin for the patient population. In a phase II trial, we may want to