Chapter 7
Non-inferiority and Equivalence Trials

Domenic J. Reda

Overview

In the "early" era of randomized clinical trials, which spans from 1946 to the late 1970s, many, if not most trials, sought to establish whether a new treatment has greater efficacy than the existing standard of care. For medical conditions where there was no known treatment to be effective, the natural comparator was no treatment or placebo. For medical conditions where there was an accepted treatment, the comparator was an "active control." With rapid development of new treatment modalities in many medical conditions, new treatments showed preliminary evidence that they could be superior to the existing standard of care, and thus, the traditional parallel group approach was appropriate.

As more effective treatments became available, the likelihood that a new treatment was more effective decreased. However, the new treatment might have other benefits compared with the standard of care, such as greater tolerability or improved side effect profile, or a more convenient treatment regimen, such as once daily dosing versus twice a day.

Thus increasingly, the intent of these trials shifted toward establishing that the effectiveness or efficacy of the new treatment was as good as, or similar to, that of an existing treatment. The earliest trials seeking to establish similarity used the traditional parallel group approach where if the null hypothesis were not rejected, then similarity was established.

However, this approach was inconsistent with the theoretical underpinnings of hypothesis testing and the role of the null and alternative hypotheses. The classical approach is to assume the null hypothesis unless the data collected in the trial indicate strong support for the alternative hypothesis. Thus, the null hypothesis can

D.J. Reda (✉)
Department of Veterans Affairs, Cooperative Studies Program Coordinating Center (151K), Hines VA Hospital, Building 1, Room B240, Hines, IL 60141, USA
e-mail: Domenic.Reda@va.gov

K.M.F. Itani and D.J. Reda (eds.), *Clinical Trials Design in Operative and Non Operative Invasive Procedures*, DOI 10.1007/978-3-319-53877-8_7

be rejected in favor of the alternative. However, when the null hypothesis is not rejected, this does not establish that the data prove the null hypothesis, rather the data are insufficient to reject the null [1]. Thus, the traditional parallel group approach cannot be used to show that an experimental treatment is similar to a control treatment.

Hypothesis Testing

For a traditional 2 parallel group trial, the null and alternative hypotheses for a two-sided and a one-sided null hypothesis are shown in Fig. 7.1.

The corresponding hypothesis testing framework for non-inferiority and equivalence trials is shown in Fig. 7.2.

For an equivalence trial where the primary outcome measure is the 30-day hospitalization rate, let us assume that the equivalence margin is 8%, group A receives the treatment and group B the control. The null hypothesis states that the 30-day hospitalization rate differs by at least 8% in the two groups, in either

Fig. 7.1 Null and alternative hypotheses for traditional parallel group trials

Traditional parallel group trial (two - sided hypothesis)

Null hypothesis $\rightarrow H_o : \Theta_A = \Theta_B$

Alternative Hypothesis $\rightarrow H_A : \Theta_A \neq \Theta_B$

Traditional (2) parallel group trial (one - sided hypothesis)

Null hypothesis $\rightarrow H_o : \Theta_A \geq \Theta_B$

Alternative Hypothesis $\rightarrow H_A : \Theta_A < \Theta_B$

where Θ_A and Θ_B are the population parameters for groups A and B, such as the mean, proportion or hazard rate for each group

Fig. 7.2 Null and alternative hypotheses for equivalence and non-inferiority trials

Equivalence trial (two - sided)

Null Hypothesis $\rightarrow H_o : |\Theta_A - \Theta_B| > \delta$

Alternative Hypothesis $\rightarrow H_A : |\Theta_A - \Theta_B| \leq \delta$

δ is the equivalence margin

Non - inferiority trial (one - sided)

Null Hypothesis $\rightarrow H_o : \Theta_A < \Theta_B - \delta$

Alternative Hypothesis $\rightarrow H_A : \Theta_A \geq \Theta_B - \delta$

δ is the non - inferiority margin

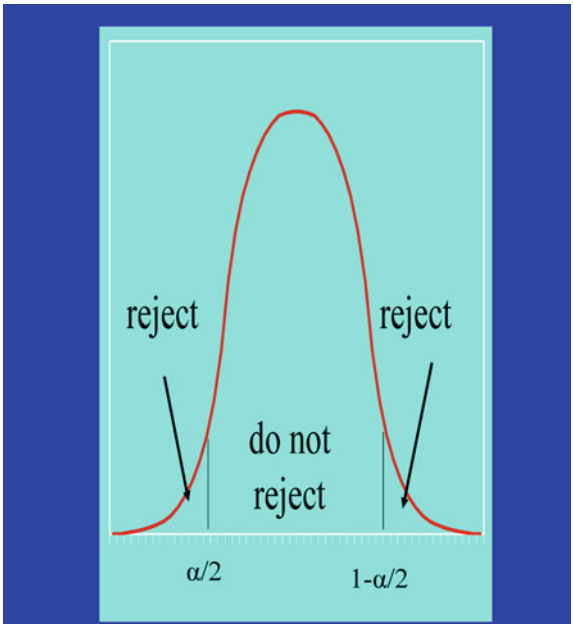

Fig. 7.3 Reject and do not reject regions for traditional two-sided hypothesis test

direction. The null hypothesis is rejected in favor of the alternative hypothesis if the data indicate a high likelihood that the 30-day hospitalization rates are within 8% of each other.

If we decided to conduct a non-inferiority trial with a non-inferiority margin of 8%, then the null hypothesis states that the 30-day complication rate for the experimental treatment is more than 8% worse than that for the control treatment. The null hypothesis is rejected in favor of the alternative hypothesis if the data indicate a high likelihood that the 30-day hospitalization rate for the experimental treatment is no more than 8% worse than that for the control treatment.

This reversal of the intent of the alternative hypothesis from establishing a difference in traditional parallel group trials to establishing similarity (either equivalence or non-inferiority depending on the structure of the alternative hypothesis) impacts how statistical tests are done.

Figure 7.3 shows the null hypothesis rejection region for a traditional two-sided null hypothesis.

This contrasts with an equivalence trial design where the reversal of the rejection region results in two one-sided tests, both of which require rejection in order to reject the equivalence null hypothesis [2]. This is shown in Fig. 7.4.

For a non-inferiority trial, a standard one-sided hypothesis test can be done. The need for two one-sided tests for an equivalence trial is specific to the two-sided hypothesis test situation. However, for a non-inferiority trial, one must still appropriately choose the (one-sided) rejection region so that it is consistent with the alternative hypothesis.

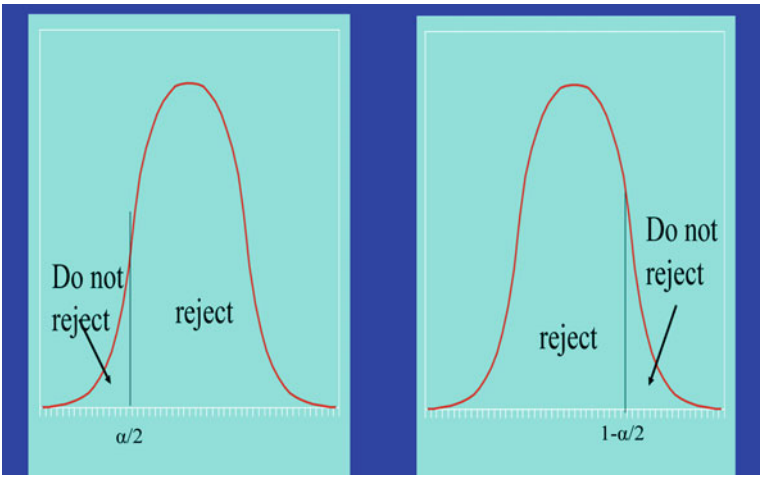

Fig. 7.4 Two one-sided tests for an equivalence trial

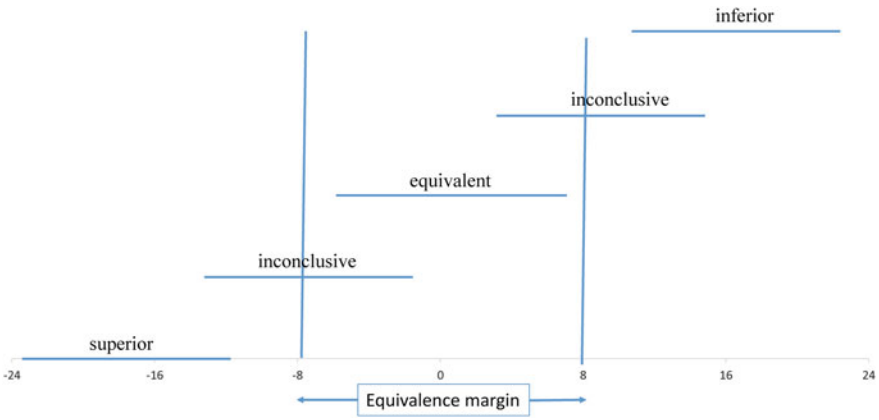

Fig. 7.5 Possible equivalence trial results

Confidence Interval Approach for Equivalence and Non-inferiority Trials

The results of equivalence and non-inferiority trials are more typically shown in a figure that displays the equivalence (or non-inferiority) margin and the confidence interval for the test statistic comparing the results for the two groups [3]. Figures 7.5 and 7.6 show possible trial outcomes for an equivalence trial with an equivalence margin of 8% and a non-inferiority trial with a non-inferiority margin of 8%.

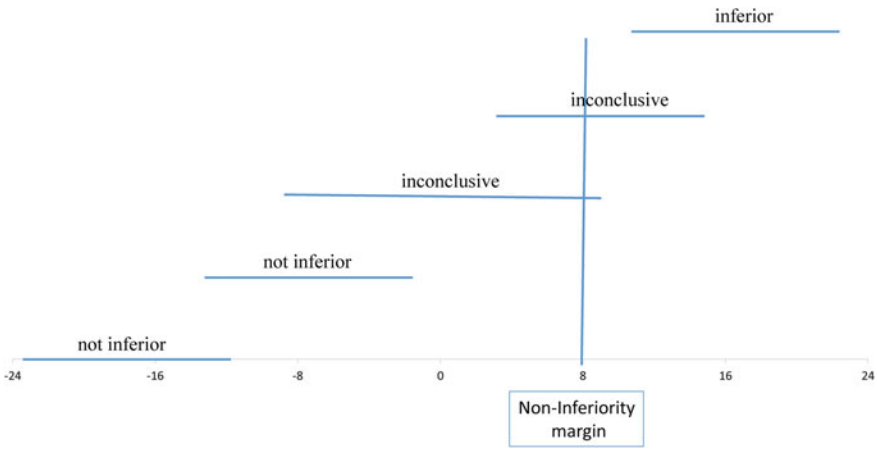

Fig. 7.6 Possible non-inferiority trial results

Example: Oral Versus Intratympanic Steroids for Treatment of Idiopathic Sudden Sensorineural Hearing Loss

Rauch et al. [4] conducted a multicenter unblinded randomized clinical trial to compare the efficacy of intratympanic steroid administration to oral steroids for treatment of idiopathic sudden sensorineural hearing loss (SSNHL). At the time the trial was initiated, standard therapy for SSNHL was a 14-day course of oral prednisolone. More recently, otolaryngologic surgeons had begun administering methylprednisolone as a series of injections into the ear canal, which was expected to produce results at least as good as oral steroid therapy, if not better, due to local concentration of the steroid into the affected area. In addition, the investigators thought intratympanic administration may have some inherent advantages because the likelihood of systemic effects would be much lower. Preliminary data from two very small studies indicated that intratympanic injection was likely as effective as oral steroid but did not appear to be more effective.

Thus, the investigators designed the trial using a non-inferiority design. Eligibility criteria included unilateral sensorineural hearing loss that developed within 72 h and was present for 14 days or less. The pure tone average (PTA), which is calculated as the arithmetic mean of the hearing thresholds at 500, 1000, 2000, and 4000 Hz in the affected ear, must have been 50 dB or higher, and the affected ear must have been at least 30 dB worse than the contralateral ear in at least 1 of the 4 PTA frequencies. Hearing must have been symmetric prior to onset of sensorineural hearing loss based on participant recall, and the hearing loss must have been deemed idiopathic following a suitable otolaryngologic evaluation.

Because oral steroid treatment has long been the standard of care for sudden hearing loss, many patients screened for enrollment in the study had referring physicians that already had initiated this treatment. Therefore, pre-enrollment steroid usage of less than 10 days was acceptable as long as audiometric criteria were met on the day of enrollment.

One hundred twenty-one patients received 60 mg/d of oral prednisone for 14 days with a 5-day taper and 129 patients received 4 doses over 14 days of 40 mg/mL of methylprednisolone injected into the middle ear.

The primary end point was the change in hearing at 2 months after treatment. Non-inferiority was defined as less than a 10 dB difference in hearing outcome between treatments. In the oral prednisone group, PTA improved by 30.7 dB compared with a 28.7 dB improvement in the intratympanic treatment group. Recovery of hearing on oral treatment at 2 months by intention-to-treat analysis was 2.0 dB greater than on intratympanic treatment (95.21% upper confidence interval, 6.6 dB). Thus, the null hypothesis of inferiority of intratympanic methylprednisolone to oral prednisone for primary treatment of sudden sensorineural hearing loss was rejected.

Example: ACOSOG Z6051—Laparoscopic-Assisted Resection Versus Open Resection of Stage II or III Rectal Cancer

The Alliance for Clinical Trials in Oncology published the results of a trial in 2015 comparing laparoscopic-assisted resection to open resection in participants with stage II or III rectal cancer [5]. This was designed as a non-inferiority trial. A total of 486 patients with clinical stage II or III rectal cancer within 12 cm of the anal verge were randomized after completion of neoadjuvant therapy to laparoscopic or open resection. The primary efficacy outcome measure was a composite of circumferential radial margin greater than 1 mm, distal margin without tumor, and completeness of total mesorectal excision.

Assuming a baseline rate of 90% oncologic success (circumferential radial margin results negative, distal margin results negative, and total mesorectal excision complete or nearly complete) for the open resection arm, the sample size of 480 patients (240 per arm) provided 80% power to declare noninferiority if oncologic success rates were truly identical, using a 1-sided z score with $\alpha = 0.10$ for falsely declaring noninferiority when the true oncologic success rate for laparoscopic resection was 84%.

Two hundred forty patients with laparoscopic resection and 222 with open resection were evaluable for analysis. Successful resection occurred in 81.7% of laparoscopic resection cases and 86.9% of open resection cases and did not support non-inferiority (difference, −5.3%; 1-sided 95% CI, −10.8% to ∞; P for

non-inferiority = 0.41). The investigators concluded that the findings do not support the use of laparoscopic resection in these patients.

Choosing a Non-inferiority Design Versus a Traditional Parallel Group Design

As noted earlier, the intent of a traditional parallel group design is to determine whether there is a difference between treatment and control while for a non-inferiority (or an equivalence) design, the intent is to establish similarity. When deciding between these two approaches, the following questions should be considered: (1) If a traditional design is considered, does the experimental treatment show preliminary evidence of superiority or is there a theoretical basis that supports an expectation of superiority? (2) If a non-inferiority design is considered, does the experimental treatment offer other possible advantages if its efficacy is shown to be similar to that of the control. If so, then these should be considered as possible secondary questions in the trial, e.g., safety and tolerability, quality of life, treatment compliance. (3) Even if superiority of the experimental treatment is possible, is it sufficient to establish similarity?

Choice of the Non-inferiority (or Equivalence) Margin

In a traditional parallel group design, before the required sample size can be calculated, the investigators must decide how large a difference between experimental and control groups would warrant a conclusion that the experimental treatment is more effective. Thus, choice of δ for a parallel group trial is often thought of as the minimum difference necessary to establish superiority.

In trials that are intended to establish similarity, δ is considered to be the maximum difference allowed to establish similarity. Originally, it was thought that establishment of similarity would warrant a smaller δ than would be used for establishment of superiority. Thus, sample size requirements were often larger for these trials than for traditional trials. However, this thinking has evolved over time and sample size requirements for the two types of designs tend to be the same. Mulla et al. [6] provide good insights on how to consider the non-inferiority margin.

In addition, equivalence designs are in the minority and many trials intending to establish similarity use non-inferiority designs. Note that because non-inferiority trials use a one-sided hypothesis test, the overall alpha for such a trial is generally 0.025, rather than 0.05. This avoids the problem of using less conservative criteria to establish similarity than would be used to establish superiority.

Non-inferiority Is Not Transitive

As long as the result for the experimental treatment is no more than δ worse than that for the control treatment, the former will be shown not inferior to the latter even if it is a little less effective. One could imagine doing a series of trials, each with a new experimental treatment in comparison with the experimental treatment from the previous trial, with a little slippage in efficacy each time. Table 7.1 summarizes this series of studies.

As a result of this potential problem, regulatory agencies additionally require that an experimental treatment be shown to be more effective than placebo when non-inferiority is demonstrated with an existing treatment known to be effective.

Poor Study Conduct Makes Establishing Non-inferiority Easier

In a traditional trial, the primary foci of study conduct include recruitment, study dropouts, completeness of data, adherence to protocol, and precision of data. All of these have a negative impact on study power, either by reducing the actual sample size achieved or by increasing the variability of the measurements. Since power is the likelihood of rejecting the null hypothesis if the null is false, then reduced power makes it more difficult to reject the null and increases the likelihood of missing an important difference in outcomes between the experimental and control treatments.

In non-inferiority trials, the effect of poor study conduct is the same insofar as it reduces effective sample size, increases variability and makes the experimental and control treatments appear to be more similar. Thus, problems with study conduct **increase** the likelihood of establishing non-inferiority (or equivalence) erroneously, i.e., make it more likely to reject the null in favor of the alternative hypothesis. As a result, greater attention needs to be paid to these study conduct issues in non-inferiority trials than in parallel group trials. Regulatory agencies may not

Table 7.1 Non-inferiority is not transitive

Response rate		Conclusion
Drug A: 50%	Placebo: 30%	Drug A superior to placebo
Drug B: 45%	Drug A: 50%	Drug B equivalent to Drug A
Drug C: 40%	Drug B: 45%	Drug C equivalent to Drug B
Drug D: 35%	Drug C: 40%	Drug D equivalent to Drug C
		Is Drug D superior to placebo if it has been shown to be equivalent to Drug C (and therefore Drugs A and B)?

accept the results of a non-inferiority trial if the completeness and precision of the collected data differ substantially from that assumed for the power and sample size calculation.

Guidance on Non-inferiority Trials

In November, 2016, the US Food and Drug Administration released its **Guidance for Industry on Non-Inferiority Clinical Trials to Establish Effectiveness** [7]. Many of the concepts in this chapter have been included in this guidance. Of particular interest is the following:

> Reasons for Using a Non-Inferiority Design
> The usual reason for using an NI active control study design instead of a superiority design is an ethical one. Specifically, this design is chosen when it would not be ethical to use a placebo, or a no-treatment control, or a very low dose of an active drug, because there is an effective treatment that provides an important benefit (e.g., life-saving or preventing irreversible injury) available to patients for the condition to be studied in the trial. Whether a placebo control can be used depends on the nature of the benefits provided by available therapy. The International Conference on Harmonisation guidance E10: Choice of Control Group and Related Issues in Clinical Trials (ICH E10) states:
> In cases where an available treatment is known to prevent serious harm, such as death or irreversible morbidity in the study population, it is generally inappropriate to use a placebo control. There are occasional exemptions, however, such as cases in which standard therapy has toxicity so severe that many patients have refused to receive it.
> In other situations, where there is no serious harm, it is generally considered ethical to ask patients to participate in a placebo-controlled trial, even if they may experience discomfort as a result, provided the setting is non-coercive and patients are fully informed about available therapies and the consequences of delaying treatment [ICH E10; pps.1314].
> Aside from this ethical reason, there may be other reasons to include an active control, possibly in conjunction with a placebo control, either to compare treatments or to assess assay sensitivity (see section III.D). Caregivers, third party payers, and some regulatory authorities have increasingly placed an emphasis on the comparative effectiveness of treatments, leading to more studies that compare two treatments. Such studies can provide information about the clinical basis for comparative effectiveness claims, which may be helpful in assessing cost effectiveness of treatments. If a placebo group is included in addition to the active comparator, it becomes possible to judge whether the study could have distinguished treatments that differed substantially, e.g., active drug versus placebo. Such comparative effectiveness studies must be distinguished from NI studies, which are the main focus of this document. The word noninferior is used here in a special sense. The methods described in this document are intended to show that a new treatment that demonstrates non-inferiority is effective, not that it is as effective as the active comparator. A new treatment may meet the standard of effectiveness (that it is superior to placebo) without justifying a conclusion that it is as effective or even nearly as effective as the active comparator.

Summary

With the availability of effective treatments for many medical conditions, the use of placebo-controlled traditional parallel group trials intended to show superiority of the experimental treatment has decreased. Non-inferiority (and equivalence) trial designs were developed to demonstrate similarity between an experimental treatment and an active control. While the outward structure of such trials seems identical to those of traditional parallel group trials, there are important differences in the underlying null and alternative hypothesis, approaches to sample size calculation and analysis of data. In addition, greater attention needs to be paid to good study conduct in these trials to reduce the likelihood of erroneously establishing non-inferiority.

References

1. Blackwelder WC. Proving the null hypothesis in clinical trials. Control Clin Trials. 1982;3 (4):345–53.
2. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. 1987;15 (6):657–80.
3. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ. 1996;313:36–9.
4. Rauch SD, Halpin CF, Antonelli PJ, et al. Oral versus intratympanic corticosteroid therapy for idiopathic sudden sensorineural hearing loss: a randomized trial. JAMA. 2011;305(20):2071–9.
5. Fleshman J, Branda M, Sargent DJ, et al. Effect of laparoscopic-assisted resection versus open resection of stage II or III rectal cancer on pathologic outcomes: the ACOSOG Z6051 randomized clinical trial. JAMA. 2015;314(13):1346–55.
6. Mulla SM, Scott IA, Jackevicius CA, et al. How to use a noninferiority trial. JAMA. 2012;308 (24):2605–11.
7. U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER). Nov 2016. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf.