Inflammatory Bowel Diseases, 2024, **30**, 2191–2194
https://doi.org/10.1093/ibd/izad314
Advance access publication 30 December 2023
Review Article - Clinical

The Clinical Interpretation of Noninferiority Trials

Pavel S. Roshanov, MD, MSc,[*,†,‡,§,] and Reena Khanna, MD, MSc[†,‡,¶,]

From the [*]Department of Medicine, Division of Nephrology, Western University, London, ON, Canada
[†]Department of Epidemiology and Biostatistics, Western University, London, ON, Canada
[‡]Lawson Health Research Institute, London, ON, Canada
[§]Population Health Research Institute, Hamilton, ON, Canada
[¶]Department of Medicine, Division of Gastroenterology, Western University, London, ON, Canada

Address correspondence to: Pavel S. Roshanov, MD, MSc, Western University, 339 Windermere Road, London, ON, Canada. Tel: 1-519-663-3350; Fax: 1-519-663-3449 (proshano@uwo.ca).

Noninferiority trials are designed to demonstrate that a new treatment is not unacceptably worse than a standard treatment, considering an allowable difference termed the noninferiority margin. We highlight that selection of noninferiority margins at the time of study design can be biased toward wider margins that favor noninferiority claims. We discuss a clinically oriented approach to interpretation of results with a focus on confidence intervals and recommend that readers base their judgments regarding noninferiority on margins reflecting patient values and preferences rather than those set by investigators. We provide examples from trials in inflammatory bowel diseases.

Key Words: noninferiority, trial design, Crohn's disease, inflammatory bowel disease, ulcerative colitis

Most clinical trials aim to demonstrate the efficacy, safety, and superiority of new treatments compared with standard or placebo. However, sometimes the objective is not to prove superiority but rather to determine if a new treatment is not unacceptably worse than an existing one. This approach is especially relevant when the new treatment offers other advantages, such as reduced side effects, costs, or burden of administration. These are noninferiority trials.[1] In situations where comparing a new therapy against placebo instead of active treatment is unethical,[2] proving that it is not unacceptably worse than the standard allows for alternative treatment options without compromising efficacy. In conditions where highly effective treatments already exist, noninferiority trials can prove the value of newer therapies that might have other benefits.

Here, we briefly explain the key distinguishing features of noninferiority trials and provide guidance for their interpretation by the practicing clinician.

Criticisms of Noninferiority Studies

Choice of Noninferiority Margin

Noninferiority trials determine whether a novel treatment is no worse than a reference treatment by a prespecified, clinically acceptable amount. This allowable difference is referred to as the "noninferiority margin." The single most effective way to demonstrate noninferiority is to select a wide noninferiority margin, and investigators and sponsor have much motivation to select margins that will provide that favorable result. This is likely one the major reasons why 96.5% of industry-sponsored noninferiority trials reached favorable conclusions in one review and were substantially more likely to do so than trials designed to test superiority where there is no opportunity to select a margin.[3] For this reason, the choice of noninferiority margin concerns regulators most.[4] Although regulators have created guidance,[5,6] the process of selecting a margin is somewhat subjective, and many noninferiority margins are poorly justified.[7]

Noninferiority trials can be smaller than superiority trials, but this gain of efficiency is directly related to the noninferiority margin. Consider a 1:1 parallel randomized trial design where 45% of patients are expected to achieve remission in the experimental group and 45% to achieve remission in the control group. If the noninferiority margin is set at 5% (ie, 40% is deemed an acceptable remission rate in the experimental group), then 3110 patients are needed to give 80% power to the trial to establish noninferiority. If the noninferiority margin were widened to 10% (ie, 35% is deemed an acceptable remission rate in the experimental group), then 778 patients are needed to achieve the same power for the test of noninferiority. Indeed, it is common that noninferiority margins are selected to match or exceed superiority effect sizes.[7] This may reduce sample size requirements, but the practice is logically inconsistent.

Bias Toward Noninferiority

Attention to the choice and fidelity of intervention delivery in the control group is important: it should reflect the standard of care. Further, defects in trial execution (eg, poor adherence to the intervention, under-reporting, or misclassification of outcome events) typically penalize superiority trials but make noninferiority trials more likely to declare noninferiority. Therefore, for the claim of noninferiority to stand, it requires support from per-protocol (ie, as treated) analyses, which carry lesser weight in the interpretation of superiority trials than intention-to-treat (ie, as randomized) analyses.

Received for publication: November 28, 2023. Editorial Decision: December 7, 2023

Difficult Interpretation of Results

Contrary to what is most familiar to readers, statistical significance (eg, $P < .001$) in a noninferiority hypothesis test does not imply superiority. In the traditional framework for hypothesis testing, the null hypothesis we are trying to reject is one of no difference in effect between the treatment groups. This is not appropriate for assessing noninferiority, because low power or imprecision in a superiority comparison mistakenly biases towards a conclusion that treatments are equivalent. Thus, it is more logical to structure the hypothesis tests such that the base assumption (null hypothesis) is that "the treatments differ," while the counter assumption (alternative hypothesis) posits that "the treatments are identical." This is opposite to the standard approach. Reversing these hypotheses prevents endorsing an inferior treatment merely due to insufficient power to detect a difference. This complicates interpretation of *P* values (ie, a significant *P* value supports smaller, not larger, difference in outcomes between the study groups). Fortunately, there is a much simpler approach to interpreting the results in noninferiority studies: examining the confidence intervals (CIs) around the estimates of treatment effect. Figure 1 depicts this approach.

Result 1 highlights that a noninferiority trial can provide evidence for superiority if the confidence interval entirely excludes "no difference" in the direction of benefit, although this may not always be reported in the article's conclusion.[8] Result 2 shows that the confidence interval does not cross the line of noninferiority. Here, we can declare that there is enough evidence to support noninferiority of the experimental treatment. Result 3 shows the confidence interval entirely inside the region between no difference and inferiority. Some sources will confusingly label this as both "inferior and not noninferior"—the experimental treatment is inferior in the pure sense (ie, because it is less effective than the control) but not so inferior that it would be unacceptable. Readers can do away with this distinction for simplicity: if we have already decided the threshold of acceptable inferiority and the confidence interval does not include or fully exceed that threshold, then we can say simply say that the experimental treatment is noninferior in result 3. In result 4, the range of plausible estimate of treatment effect spans from the region of superiority to the region of inferiority. This is sometimes labeled as indeterminate. For the clinical reader, the most direct interpretation is that there is not enough evidence to exclude inferiority of the novel treatment. Result 5 shows the upper bound of the confidence interval exceeding the noninferiority threshold, so inferiority cannot be excluded. Result 6 shows the confidence interval entirely beyond the inferiority threshold, and the novel treatment is inferior.

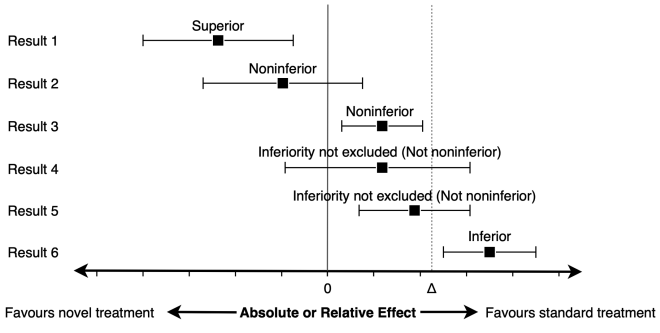

Figure 1. Interpreting noninferiority study results. Error bars indicate 2-sided confidence intervals. The dashed line labeled Δ represents the noninferiority threshold or the maximum allowable excess of outcome events in the experimental treatment group compared with the standard treatment group.

Example: The LADI Trial

The LADI trial which compared persistent flares at week 48 between patients with Crohn's disease in stable remission who were 2:1 randomized to open-label increased compared with conventional adalimumab dose interval.[9] Its results are depicted in Figure 2. The noninferiority margin was 15% on an absolute risk difference scale. At week 48, 3 of 109 patients (3%) in the increased dose interval group had persistent flare compared with 0 of 60 patients in the conventional dose interval group (risk difference adjusted for concomitant medication use 1.86%; 90% CI, –0.35% to 4.07%). Both intention-to-treat and per-protocol populations were identical.[9] Because the upper bound of the confidence interval does not exceed the prespecified noninferiority margin of 15%, the increased dose interval strategy can be said to be noninferior to the conventional dose interval strategy (Figure 2).[9]

Was the Trial Designed and Conducted Sufficiently Well to Provide a Fair Comparison?

Many of the principles of good noninferiority trials are the same as those of superiority trials; discussing them in detail is beyond the scope of this article. The LADI trial was an open-label trial, meaning that everyone involved was aware of patients' group assignment. The trial could have been blinded by including placebo injections of adalimumab in place of active drug every other week. One could argue, however, that the awareness of decreased dosing could bias toward higher incidence of persistent flares in the group receiving drug less frequently.

Opposite to superiority hypotheses, noninferiority and equivalence hypotheses are tested more conservatively in the per-protocol (or "as-treated") analysis than in the intention-to-treat (or "as-randomized") analysis. This is because poor adherence to allocated treatments—as can occur in poorly conducted trials—biases toward equivalence of outcomes and, therefore, risks finding a useless treatment noninferior. We cannot fully rely on the per-protocol analysis, however, because it still carries the same risks of bias that make intention-to-treat the preferred choice in superiority trials. Ideally, the intention-to-treat and per-protocol analyses support the same conclusion (as they do in the LADI trial). Discrepancies between the 2 methods demand scrutiny.

Did Patients in the Control Group Receive the Control Therapy at the Expected Standard?

In the LADI trial, the control group received 40 mg of adalimumab every 2 weeks, as would be expected in patients maintained in stable remission.

Is the Primary Outcome Important to Patients?

The primary outcome on which noninferiority is being declared should be sufficiently important to patients. These

outcomes usually reflect patient symptoms or some burdensome escalation in management. Persistent flare—the primary outcome in the LADI trial—is likely to be important to patients. Demonstrating noninferiority on a surrogate outcome (eg, C-reactive protein) leaves the possibility that the treatment is still inferior on a patient-important outcome.

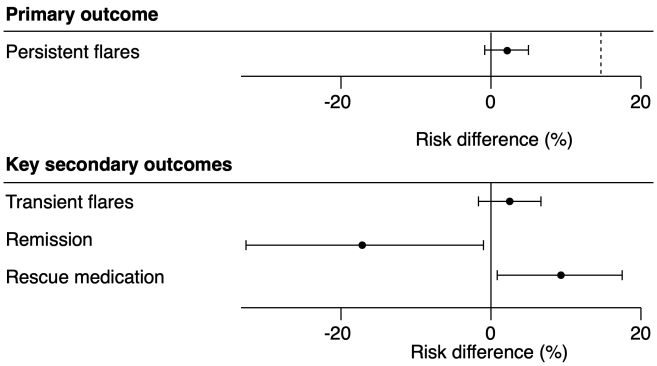

Figure 2. Primary and key secondary outcomes in the LADI trial (adapted from Linschoten et al., 2023).[9] Error bars depict 90% confidence intervals.

Does the Confidence Interval Cross an Inferiority Margin that Reflects my Patients' Values and Preferences?

Noninferior does not mean equivalent; it means inferior to a degree no greater than what people are willing to accept. The noninferiority margin selected by the investigators—together with other assumptions—affected the sample size calculation at the time of study design. It should not affect interpretation of the results. There are many motivations that may lead investigators, drug manufacturers, and regulators to select a noninferiority margin that is more permissive than clinicians or patients might deem acceptable. Once the data are collected, the only noninferiority margin important for interpretation is that which reflects patients' values and preferences.

In the LADI trial, many patients enjoying persistent remission of their Crohn's disease on a stable q2 week regimen of adalimumab may consider that 15% higher risk (in absolute terms) of a persistent flare is too high to justify adopting a q4 week regimen. However, the observed absolute risk difference was 1.86% (90% confidence interval –0.35% to 4.07%), and many patients may readily accept a risk increase in that range.

A survey of physicians found that most would consider a margin of 10% or smaller to be appropriate for clinical efficacy in head-to-head trials on the outcome of clinical remission in biologic-experienced patients.[10] The margins deemed

Table 1. Examples of noninferiority randomized controlled trials with clinical outcomes in inflammatory bowel disease.

Study	Comparison	Potential Advantage of Novel Treatment	Primary Outcome	Study NI Margin	Result (95% CI)	Conclusion With Study NI Margin	Conclusion With -10% NI Margin
Ye 2019[11]	Biosimilar CT-P13 vs infliximab in Crohn's disease	Lower cost	Absolute % difference in patients with a decrease of ≥ 70 points in CDAI from baseline to week 6.	–20%	–4·9% favoring infliximab (-16.9 to 7.3)	CT-P13 is noninferior to infliximab.	CT-P13 is *not* noninferior to infliximab.
Dignass 2014[12]	Once vs three times daily dosing of budesonide	Convenience	Absolute % difference in patients with clinical remission defined as CDAI < 150 at week 8.	–15%	–3.9% favoring three times daily dosing (-14.6 to 6.4)	Once daily dosing is noninferior to three times daily dosing.	Once daily dosing is *not* noninferior to three times daily dosing.
Kruis 2009[13]	3 gram daily vs 1 gram three times per day mesalazine granules in ulcerative colitis	Convenience	Absolute % difference in patients achieving clinical remission defined as clinical activity index ≤4 at week 8.	–15%	3.4% favoring 3 gram daily dosing (-5.0 to 11.8)	3 gram daily dosing is noninferior to 1 gram three times per day.	3 gram daily dosing is noninferior to 1 gram three times per day.
Kruis 2022[14]	Budesonide suppository vs budesonide recal foam	Convenience	Absolute % difference in patients achieving each of the following co-primary endpoints: 1) Clinical remission defined as UC-DAI subscore for stool frequency of 0 or 1 and a subscore for rectal bleeding of 0 2) Mucosal healing, defined as a modified UC-DAI subscore for mucosal appearance of 0 or 1	–10%	Clinical remission: 4.5% favoring suppository (–3.0 to 11.9) Mucosal healing: 0% (-6.9 to 6.9)	Suppository is noninferior to foam.	Suppository is noninferior to foam.

Abbreviations: CDAI, Crohn's disease activity index; UC-DAI, Ulcerative colitis disease activity index.

acceptable ranged between 5 and 15% for most outcomes examined.[10]

Were All Important Outcomes Examined? Were There Important Outcomes on Which the Experimental Treatment Performed Unacceptably Worse?

Although effects on the primary outcome, on which the claim of noninferiority is based, support little, if any, increase in the risk of persistent flare, secondary outcomes that may be important to patients may tell a different story. The increased dose interval led to substantially higher risk of rescue medication use (mostly corticosteroids), substantially lower probability that patients would remain in remission at 48 weeks, and a (nonstatistically significant) increase in the risk of transient flares.

Table 1 shows other examples of noninferiority trials and their interpretation.

Conclusion

Noninferiority trials will play an increasingly large role in IBD as numerous effective treatment options are established. These trials have some unique pitfalls that require thoughtful interpretation. In our view, the choice of noninferiority margin made by study investigators should be ignored by clinical readers who should instead replace it with a margin determined by asking what trade-offs are acceptable to the patient in their clinic.

Conflicts of Interest

R.K. reports fees for Consulting/Speaking from: AbbVie, Amgen, BMS, Encycle, Innomar, Gilead, Janssen, Jamp, Lilly, Merck, Pendopharm, Pfizer, Roche/Genetech, Alimentiv (formerly Robarts Clinical Trials), Shire, and Takeda Canada outside the submitted work; research fees from Roche/Genetech.

P.S.R.: none

References

1. Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. *Trials*. 2011;12(1):1-12.
2. Roshanov PS, Khanna R. The use of placebo in clinical trials of inflammatory bowel disease. *Inflamm Bowel Dis*. 2023:izad281. doi:10.1093/ibd/izad281. Online ahead of print.
3. Flacco ME, Manzoli L, Boccia S, et al. Head-to-head randomized trials are mostly industry sponsored and almost always favor the industry sponsor. *J Clin Epidemiol*. 2015;68(7):811-820.
4. Althunian TA, de Boer A, Mantel-Teeuwisse AK, et al. Assessment of the regulatory dialogue between pharmaceutical companies and the european medicines agency on the choice of noninferiority margins. *Clin Ther*. 2020;42(8):1588-1594.
5. Committee for medicinal products for human use (CHMP) of the European Medicines Agency. Guideline on the Choice of the Non-Inferiority Margin. 2006. Accessed December 1. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-choice-non-inferiority-margin_en.pdf
6. Office of New Drugs in the Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER) at the Food and Drug Administration. Non-Inferiority Clinical Trials to Establish Effectiveness Guidance for Industry. 2016. Accessed December 1. https://www.fda.gov/media/78504/download
7. Ofori S, Cafaro T, Devereaux PJ, et al. Noninferiority margins exceed superiority effect estimates for mortality in cardiovascular trials in high-impact journals. *J Clin Epidemiol*. 2023;161:20-27.
8. Stähli BE, Varbella F, Linke A, et al.; MULTISTARS AMI Investigators. Timing of complete revascularization with multivessel PCI for myocardial infarction. *N Engl J Med*. 2023;389(15):1368-1379.
9. Linschoten RCA van, Jansen FM, Pauwels RWM, et al. Increased versus conventional adalimumab dose interval for patients with Crohn's disease in stable remission (LADI): a pragmatic, open-label, non-inferiority, randomised controlled trial. *Lancet Gastroenterol Hepatol*. 2023;8(4):343-355.
10. Olivera P, Sandborn WJ, Panés J, et al. Physicians' perspective on the clinical meaningfulness of inflammatory bowel disease trial results: an International Organization for the Study of Inflammatory Bowel Disease (IOIBD) survey. *Aliment Pharmacol Ther*. 2018;47(6):773-783.
11. Ye BD, Pesegova M, Alexeeva O, et al. Efficacy and safety of biosimilar CT-P13 compared with originator infliximab in patients with active Crohn's disease: an international, randomised, double-blind, phase 3 non-inferiority study. *Lancet*. 2019;393(10182):1699-1707.
12. Dignass A, Stoynov S, Dorofeyev AE, et al.; International Budenofalk Study Group. Once versus three times daily dosing of oral budesonide for active Crohn's disease: a double-blind, double-dummy, randomised trial. *J Crohns Colitis*. 2014;8(9):970-980.
13. Kruis W, Kiudelis G, Rácz I, et al.; International Salofalk OD Study Group. Once daily versus three times daily mesalazine granules in active ulcerative colitis: a double-blind, double-dummy, randomised, non-inferiority trial. *Gut*. 2009;58(2):233-240.
14. Kruis W, Siegmund B, Lesniakowski K, et al. Novel budesonide suppository and standard budesonide rectal foam induce high rates of clinical remission and mucosal healing in active ulcerative proctitis: a randomised, controlled, non-inferiority trial. *J Crohns Colitis*. 2022;16(11):1714-1724.