EDITORIAL

Challenges in the Design, Registration, and Reporting of Noninferiority Trials

Rolf H. H. Groenwold, MD, PhD; Eibert R. Heerdink, PhD; Olaf H. Klungel, PharmD, PhD

In superiority trials, usually researchers expect that the treatment of interest is more effective (superior) compared with placebo or an active comparator treatment. To demonstrate superiority, a null hypothesis is defined that usually states that the outcomes are the same in the treatment arms. Convincing evidence against that hypothesis (often based on a *P* value that is less than a predefined threshold, eg, .05) would lead researchers to reject the null hypothesis and thus conclude that one treatment is better than the other. Therefore, in a superiority trial that provides convincing evidence that an experimental treatment is superior to placebo (or active comparator), conclusions about (relative) efficacy can be drawn.

Related article page 193

In noninferiority trials, researchers aim to demonstrate that a new treatment is not "unacceptably worse" than the comparator treatment, where "unacceptably worse" is quantified by a so-called noninferiority margin Δ.[1] In this case, the null hypothesis would be that the new treatment is in fact inferior by at least this margin Δ. To reject the null hypothesis requires to show that the difference between the 2 treatments is not larger than this margin. This is commonly done by calculating a 95% confidence interval and when the margin is not included in this interval, the null hypothesis of inferiority is rejected: we accept the alternative hypothesis that states that the new treatment is noninferior to the comparator.

In the study by Foa and colleagues,[2] choosing a noninferiority design instead of a superiority design is completely reasonable, because tapered discontinuation of serotonin reuptake inhibitor (SRI) treatment is not expected to be superior with regard to the primary outcome measure of wellness (Yale-Brown Obsessive-Compulsive Scale [Y-BOCS]), but neither is it expected to be unacceptably inferior. Nevertheless, tapering SRI treatment may have desired effects such as reduction of adverse effects and health care costs. Thus, it is a clinically relevant question to ask whether tapering SRIs in patients with obsessive-compulsive disorder after successful exposure/response prevention therapy is noninferior compared with continuing SRIs.

Defining the Noninferiority Margin

Clearly, a critical aspect of a noninferiority trial is its prespecified margin. If too lenient (ie, too large), the researchers may claim that one treatment is noninferior compared with the other, while actually effects may differ by a clinically relevant amount. Yet, a margin that is too strict would require a large sample size to have sufficient power to be able to show noninferiority (should it exist). It is therefore important to prespecify a margin that is informed by statistical and clinical reasoning in order to convince the clinical research community on the conclusions of a noninferiority study. The basis of this reasoning is formed by results (point estimate or its confidence interval) of the effect of the active comparator (eg, against placebo) found in previous studies. It is then a clinical judgment to decide which fraction of that effect should be maintained by the treatment under investigation in order to be noninferior.[3]

In the article by Foa et al, they indicate that the noninferiority margin was based on the observed baseline variation in the outcome variables of the study (Y-BOCS, Hamilton Depression Rating Scale [HDRS], and quality of life), namely half of the standard deviation (SD) of those measures. Interestingly, those SDs do not correspond to the SDs reported in Table 1 in the Foa et al article, ie, 3.6 for Y-BOCS and 3.1 for HDRS. Using those instead, the margin for the primary outcome Y-BOCS (ie, Δ = 3.0) corresponds to more than 80% (3.0/3.6) of an SD. Based on a margin of half of the observed SD, ie, Δ = 1.8, the conclusion of this study would not be that tapering is noninferior; in that case the hypothesis of inferiority would not be rejected. However, in the protocol, which is available as a Supplement to the Foa et al article, the authors indicate that the predefined margin of Δ = 3.0 for Y-BOCS is based on a clinically minimal difference. Given the consequences and thus the relevance of the noninferiority margin, it is critical that it is defined prior to data analysis, and made publicly accessible in a repository such as ClinicalTrials.gov.

Significance Threshold

The researchers used a 1-sided 95% confidence interval of which the upper bound corresponds to a 2-sided 90% confidence interval. More conventionally, a 1-sided 97.5% confidence interval would be chosen, corresponding to the upper bound of a 2-sided 95% confidence interval. For the primary outcome, Y-BOCS, this would not make a difference, because the upper bound of a 1-sided 97.5% confidence interval would be approximately 2.5, still below the noninferiority margin Δ = 3.0. However, for the secondary end point HDRS, the upper bound of a 1-sided 97.5% confidence interval would be approximately 2.6, which is greater than its margin Δ = 2.5, and therefore would not have led to the conclusion that tapering is noninferior to continuing SRI treatment. This again illustrates the importance of prespecifying statistical choices, not only in terms of the noninferiority margin, but also in terms of the statistical significance threshold. Again, ClinicalTrials.gov could be a place where such information is recorded, yet no information could be found.

Intention to Treat vs Per Protocol

Data analysis included intention-to-treat analysis as well as per-protocol analysis. While in superiority studies the former is the preferred way of analysis, in noninferiority studies an intention-to-treat analysis is generally considered to be in favor of rejecting the null hypothesis, ie, in favor of claiming noninferiority. Therefore, the per-protocol analysis should indicate noninferiority too, in order to reject the null hypothesis of inferiority. Particularly when the number of crossovers (or dropouts) in a study is substantial, intention to treat can lead to an underestimation of differences compared with the situation in which all participants had adhered to the study protocol. In the study by Foa et al, both analyses are described in the Methods section. Dropouts are reported in eTable 1 in the Supplement, showing that the number of dropouts and participants removed because of clinical worsening is indeed substantial and much higher in the tapering group compared with the continuation group. Tapering in this study was done in 4 weeks with a 25% dose reduction per week, which may have resulted in antidepressant withdrawal symptoms.[4] It is possible that a longer tapering period with smaller dose reductions supported by more frequent contacts with a physician could have reduced dropout and removal due to clinical worsening, resulting in more valid outcome measurements.

In conclusion, the study by Foa et al aimed to answer the question whether SRI treatment can be discontinued in patients with obsessive-compulsive disorder who are successfully treated with exposure/response therapy. This is clinically relevant because it may prevent unnecessary use of medication with known adverse effects. A noninferiority trial is the appropriate design to answer this question. However, it is also methodologically challenging because it requires to prove that patients who discontinue treatment do not have worse outcomes than patients who continue SRI use.

In all randomized clinical trials, many choices need to be made before results are analyzed and reported. Transparent reporting of the choices made is essential and can be done by preregistration of these decisions in a publicly available protocol, or on repositories such as ClinicalTrials.gov. For noninferiority trials, transparency is crucial regarding the noninferiority margin, because that ultimately decides the outcome of the trial. Uncertainty about this will not change the conclusions of those noninferiority studies per se, but may prevent discussions about those studies to focus on methodological aspects, instead of the clinical implications a study may have.

ARTICLE INFORMATION

Author Affiliations: Department of Clinical Epidemiology, Leiden University Medical Center (LUMC), Leiden, the Netherlands (Groenwold); Division of Pharmacoepidemiology & Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands (Heerdink, Klungel).

Corresponding Author: Olaf H. Klungel, PharmD, PhD, Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, PO Box 80082, 3508 TB Utrecht, the Netherlands (o.h.klungel@uu.nl).

Published Online: January 26, 2022. doi:10.1001/jamapsychiatry.2021.3932

Conflict of Interest Disclosures: None reported.

REFERENCES

1. Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. *Trials*. 2011;12:106. doi:10.1186/1745-6215-12-106

2. Foa EB, Simpson HB, Gallagher T, et al. Maintenance of wellness in patients with obsessive-compulsive disorder who discontinue medication after exposure/response prevention augmentation: a randomized clinical trial. *JAMA Psychiatry*. Published online January 26, 2022. doi:10.1001/jamapsychiatry.2021.3997

3. Althunian TA, de Boer A, Groenwold RHH, Klungel OH. Defining the noninferiority margin and analysing noninferiority: an overview. *Br J Clin Pharmacol*. 2017;83(8):1636-1642. doi:10.1111/bcp.13280

4. Fava GA, Cosci F. Understanding and managing withdrawal syndromes after discontinuation of antidepressant drugs. *J Clin Psychiatry*. 2019;80(6):19com12794. doi:10.4088/JCP.19com12794