Psychological Medicine

cambridge.org/psm

Some problems with non-inferiority tests in psychotherapy research: psychodynamic therapies as an example

Winfried Rief[1] and Stefan G. Hofmann[2]

[1]University of Marburg, Marburg, Germany and [2]Boston University, Boston, MA, USA

Correspondence

Cite this article: Rief W, Hofmann SG (2018). Some problems with non-inferiority tests in psychotherapy research: psychodynamic therapies as an example. *Psychological Medicine* **48**, 1392–1394. https://doi.org/10.1017/S0033291718000247

Received: 15 November 2017
Revised: 15 January 2018
Accepted: 16 January 2018
First published online: 14 February 2018

Author for correspondence:
Dr. W. Rief, E-mail: rief@uni-marburg.de

Abstract

In virtually every field of medicine, non-inferiority trials and meta-analyses with non-inferiority conclusions are increasingly common. This non-inferiority approach has been frequently used by a group of authors favoring psychodynamic therapies (PDTs), concluding that PDTs are just as effective as cognitive-behavioral therapies (CBT). We focus on these examples to exemplify some problems associated with non-inferiority tests of psychological treatments, although the problems also apply to psychopharmacotherapy research, CBT research, and others. We conclude that non-inferiority trials have specific risks of different types of validity problems, usually favoring an (erroneous) non-inferiority conclusion. Non-inferiority trials require the definition of non-inferiority margins, and currently used thresholds have a tendency to be inflationary, not protecting sufficiently against degradation. The use of non-inferiority approaches can lead to the astonishing result that one single analysis can suggest both, superiority of the comparator (here: CBT) *and* non-inferiority of the other treatment (here PDT) at the same time. We provide recommendations how to improve the quality of non-inferiority trials, and we recommend to consider them among other criteria when evaluating manuscripts examining non-inferiority trials. If psychotherapeutic families (such as PDT and CBT) differ on the number of investigating trials, and in the fields of clinical applications, and in other validity aspects mentioned above, conclusions about their general non-inferiority are no more than a best guess, typically expressing the favored approach of the lead author.

After many attempts to investigate the superiority of one psychological treatment over another have failed, more and more studies use non-inferiority approaches when investigating psychotherapies. Here, we will discuss characteristics, but also limitations of non-inferiority comparisons between psychological treatments. This approach has been most frequently used by a group of authors favoring psychodynamic therapies (PDTs), concluding that PDTs are just as effective as cognitive-behavioral therapies (CBT). We will focus on these examples to exemplify some problems associated with non-inferiority tests of psychological treatments. However, the same problems with non-inferiority tests are evident in virtually every field of medicine, because non-inferiority trials are increasingly common in pharmacological research [e.g. on antidepressants (Szegedi *et al.* 2005; Jeong *et al.* 2015)], in CBT research (e.g. cognitive therapy *v.* behavioral activation (Richards *et al.* 2016)), or when comparing face-to-face psychotherapy with internet-based psychotherapy (Lappalainen *et al.* 2014).

The traditional clinical trial uses a superiority test to examine whether the difference of average improvements between two or more treatments is significantly different from zero (also known as the Null Hypothesis Significance Testing). In contrast, non-inferiority trials define a non-inferiority threshold (e.g. an effect size of Cohen's $d = 0.20$) to test for non-inferiority if this threshold is not included in the confidence interval (CIs) of the average difference of improvements between treatments. This approach is problematic for a number of conceptual and methodological reasons.

First, so far there is no general agreement about the non-inferiority threshold. A recently published trial on PDT used a criterion of standardized mean difference (SMD) = 0.25, and other PDT trials used similar thresholds (Steinert *et al.* 2017). However, in contrast to the Null Hypothesis of superiority trials, such a non-inferiority criterion cannot be evaluated without reference to expected effects. The expected effects depend on the efficacy of the intervention, but also on the method applied. Leucht *et al.* (2012) found median SMDs compared with placebo for the most-frequently used medical drugs of SMD = 0.37 (SMD = 0.41 for psychopharmacological drugs). Using this as a guide, the lowest threshold of considering PDT as non-inferior to other treatments would be treatment benefits that correspond just to 32% of the effectiveness of other medical treatments [(0.37–0.25)*100%/0.37]. Compared with the average effectiveness of psychological therapies for depression (Cuijpers *et al.* 2014), PDT would have been considered as non-inferior even if it is not sure whether it exceeds just

53% of the effectiveness of psychological comparison treatments. Clearly, such a threshold is inflationary, hiding clinically meaningful differences that might exist.

Second, a major difference between superiority and non-inferiority tests is the influence of the low quality of the trials and the low efficacy of the interventions. Low quality can impede the detection of superiority but makes it more likely to observe non-inferiority. Two of the most popular non-inferiority trials comparing PDT with CBT (Driessen *et al.* 2013; Connolly Gibbons *et al.* 2016) reported response rates (reduction of 50%) as low as 16% for PDT and 22% for CBT (Connolly Gibbons *et al.* 2016). However, according to meta-analyses, typical response rates for CBT are 53% (Cuijpers *et al.* 2014), indicating that the comparison treatments of these trials may not have been adequately conducted to reach its typical therapeutic effects. Factors such as low adherence (including patients attending none or only one treatment session), or low treatment fidelity can contribute to these flaws. Thus, non-inferiority trials must first show that all treatments (especially the comparator) have been implemented according to the standards before testing for non-inferiority.

Third, the low quality of trials can easily and incorrectly lead investigators to conclude non-inferiority of the two treatments. For example, the more data that are missing in a trial, the more likely it is that the two treatments are considered non-inferior; this effect is even more pronounced if the missing values are estimated using the dataset of the comparison treatment arm, as is occasionally done (Zipfel *et al.* 2014). Similarly, trials that permit concurrent treatments (e.g. concurrent drug treatments of patients receiving psychotherapy), comparable treatment elements in two different treatment arms, poor training of the psychotherapists in the specific treatments, etc. further blur the differences between psychological treatments. This increases the likelihood that one treatment is judged to be non-inferior over another even though it is not. The aforementioned anorexia trial (Zipfel *et al.* 2014), for instance, reports similar BMI results for PDT and CBT (and enhanced medical care) when the clinical groups not only included anorexia patients but also patients with subthreshold anorexia. However, when restricting BMI analyses to only anorexia patients (meeting the underweight criteria BMI <17.5), CBT was superior to the other treatments. Thus, including patients with normal primary outcome scores further supports non-inferiority conclusions. These examples show that the reasons for non-inferiority can be manifold and may not be related to the efficacy of the interventions.

The use of non-inferiority approaches can lead to the astonishing result that one single trial can suggest both, superiority of the comparator *and* non-inferiority of the target treatment at the same time. A recently published article used the per protocol approach of non-inferiority of PDT *v.* CBT, and the authors found their assumptions confirmed. However, they also found a statistically significant disadvantage of PDT compared with other treatments (including CBT) (Steinert *et al.* 2017). Does this support the (per protocol) non-inferiority of PDT, or does it support the statistically more robust inferiority of PDT *v.* other treatments? It should be noted that in such an equivocal situation, the financial sponsor of this study (the German Association for Psychoanalysis, Psychotherapy, Psychosomatics, and Psychodynamic Psychology) might favor one interpretation over the other.

These meta-analyses (many of them published in high-ranking journals) follow the same aim: show that one treatment (here PDT) is as effective as the current best-evidence treatment (here CBT). However, attempting to answer this question using these broad descriptors of treatments seems highly questionable. Both CBT and PDT are not specific interventions, but classes of psychotherapies that offer a wide variety of treatments. Some of these interventions may be potentially advantageous in one condition, but not in other clinical conditions.

At present, the trials supporting CBT far outnumber the trials supporting PDT. Therefore, even if all direct comparisons resulted in similar improvements from both interventions (which they do not), the CIs for CBT are much smaller due to a greater amount of evidence; therefore success would be achieved more reliably with CBT interventions. Again using the example of depression (Cuijpers *et al.* 2014), CBT results in a mean BDI improvement of 14.4 points (CI 13.1–15.8), while PDI resulted in improvements of 10.7 (CI 5.3–16.1). Thus, the interval for PDT indicates that its real efficacy can be slightly above CBT, or in the range of placebo interventions, or somewhere in between. Very broad CIs in meta-analyses are typical for rarely investigated interventions, and this should not be interpreted as evidence for non-inferiority. Only if the CIs are based on stable evidence (and preferably if CIs are comparable between populations/treatments) should they be used to evaluate comparability between treatments.

Moreover, for several clinical conditions, there is a lack of any convincing evidence for PDTs, while strong evidence for CBT approaches exits (e.g. OCD, psychosis, hypochondria, and insomnia). However, two types of psychological interventions cannot be concluded to be similarly effective in general, if one treatment shows evidence in fields where the other treatment does not. Thus, the question whether PDT is as efficacious as CBT can only be answered by examining whether a specific psychodynamic approach is non-inferior to a specific CBT approach when applied to a specific clinical condition, but not in a generalized way.

To conclude, non-inferiority trials have specific risks of different types of validity problems, such as construct validity (e.g. definition of treatment type), statistical validity (e.g. biases when imputing missing values), internal validity (e.g. permission of concurrent treatments), and external validity (e.g. inclusion of subthreshold disorders). They require the definition of non-inferiority thresholds, and current approaches do not protect sufficiently against degradation (Gladstone & Vach, 2014). We recommend that this threshold should not fall below 90% of the expected effects of the first-line treatments (e.g. threshold SMD of ±0.05, if the uncontrolled effect size is expected as SMD = 0.50), and minimum clinically acceptable differences must be defined more restrictively. Non-inferiority trials need to consider the many factors that could (erroneously) contribute to non-inferiority results. In particular, they must demonstrate the adequate use of the comparison treatment. Significant superiority results are typically more robust than non-inferiority results because non-inferiority trials can have a systematic bias toward non-inferiority [(FDA, 2016) p. 10]. These recommendations should be further harmonized with other recommendations for non-inferiority trials (e.g. (Wang *et al.* 2015; FDA, 2016)), and they should be taken into consideration when evaluating manuscripts examining non-inferiority trials.

If psychotherapeutic families (such as PDT and CBT) differ on the number of investigating trials, and in the fields of clinical applications, and in other validity aspects mentioned above, conclusions about their non-inferiority are no more than a best guess, typically expressing the favored approach of the lead author.

https://doi.org/10.1017/S0033291718000247 Published online by Cambridge University Press

Acknowledgement. The authors did not receive any funding that could have influenced the content of this work.

References

Connolly Gibbons M, Gallop R, Thompson D, Luther D, Crits-Christoph K, Jacobs J, *et al.* (2016) Comparative effectiveness of cognitive therapy and dynamic psychotherapy for major depressive disorder in a community mental health setting: a randomized clinical noninferiority trial. *JAMA Psychiatry* **73**, 904–911.

Cuijpers P, Karyotaki E, Weitz E, Andersson G, Hollon SD and van Straten A (2014) The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. *Journal of Affective Disorders* **159**, 118–126.

Driessen E, Van HL, Don FJ, Peen J, Kool S, Westra D *et al.* (2013) The efficacy of cognitive-behavioral therapy and psychodynamic therapy in the outpatient treatment of major depression: a randomized clinical trial. *American Journal of Psychiatry* **170**, 1041–1050.

FDA (2016) Non-Inferiority Clinical Trials to Establish Effectiveness. Guidance for Industry. Available at https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf

Gladstone BP and Vach W (2014) Choice of non-inferiority (NI) margins does not protect against degradation of treatment effects on an average–an observational study of registered and published NI trials. *PLoS ONE* **9**, e103616.

Jeong JH, Bahk WM, Woo YS, Lee KU, Kim DH, Kim MD *et al.* (2015) Efficacy and safety of generic escitalopram (Lexacure (R)) in patients with major depressive disorder: a 6-week multicenter, randomized, rater-blinded, escitalopram-comparative, non-inferiority study. *Neuropsychiatric Disease and Treatment* **11**, 2557–2564.

Lappalainen P, Granlund A, Siltanen S, Ahonen S, Vitikainen M, Tolvanen A *et al.* (2014) ACT Internet-based vs face-to-face? A randomized controlled trial of two ways to deliver acceptance and commitment therapy for depressive symptoms: an 18-month follow-up. *Behaviour Research and Therapy* **61**, 43–54.

Leucht S, Hierl S, Kissling W, Dold M and Davis JM (2012) Putting the efficacy of psychiatric and general medicine medication into perspective: review of meta-analyses. *British Journal of Psychiatry* **200**, 97–106.

Richards DA, Ekers D, McMillan D, Taylor RS, Byford S, Warren FC *et al.* (2016) Cost and outcome of behavioural activation versus cognitive behavioural therapy for depression (COBRA): a randomised, controlled, non-inferiority trial. *Lancet* **388**, 871–880.

Steinert C, Munder T, Rabung S, Hoyer J, and Leichsenring F (2017) Psychodynamic therapy: as efficacious as other empirically supported treatments? A meta-analysis testing equivalence of outcomes. *American Journal of Psychiatry* **174**, 943–953.

Szegedi A, Kohnen R, Dienel A and Kieser M (2005) Acute treatment of moderate to severe depression with hypericum extract WS 5570 (St John's wort): randomised controlled double blind non-inferiority trial versus paroxetine. *Bmj* **330**, 503.

Wang Y, Harigaya Y, Cavaillé-Coll M, Colangelo P and Reynolds KS (2015) Justification of noninferiority margin: methodology considerations in an exposure–response analysis. *Clinical Pharmacology & Therapeutics* **97**, 404–410.

Zipfel S, Wild B, Gross G, Friederich HC, Teufel M, Schellberg D *et al.* (2014) Focal psychodynamic therapy, cognitive behaviour therapy, and optimised treatment as usual in outpatients with anorexia nervosa (ANTOP study): randomised controlled trial. *Lancet* **383**, 127–137.

https://doi.org/10.1017/S0033291718000247 Published online by Cambridge University Press