Words: 1532

Equivalence and non-inferiority testing in psychotherapy research

Falk Leichsenring[1], DSc, Allan Abbass[2], MD FRCPC, Ellen Driessen[3], PhD, Mark Hilsenroth[4], PhD, Patrick Luyten[5], PhD, Sven Rabung[6], PhD, Christiane Steinert[1,7], PhD

[1] Department of Psychosomatics and Psychotherapy, Justus-Liebig-University Giessen, Ludwigstr. 76, D-35392 Giessen, Germany

[2] Department of Psychiatry, Dalhousie University; Centre for Emotions and Health, Halifax, 8203 5909 Veterans Memorial Lane, Halifax, NS, Canada, B3H 2E2

[3] Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health research institute, Vrije Universiteit Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands

[4] The Derner Institute of Advanced Psychological Studies, Adelphi University, Hy Weinberg Center, 1 South Avenue, Garden City, NY 11530-0701, USA

[5] Faculty of Psychology and Educational Sciences, University of Leuven, Klinische Psychologie (OE), Tiensestraat 102 - bus 3722, 3000 Leuven, Belgium, and Research Department of Clinical, Educational and Health Psychology, University College London, Gower Street, London WC1E 6BT, UK

[6] Department of Psychology, Alpen-Adria-Universität Klagenfurt, Universitätsstr. 65-67, A-9020 Klagenfurt, Austria

[7] MSB Medical School Berlin, Department of Psychology, Calandrellistr. 1-9, 12247 Berlin, Germany

Corresponding Author:
Prof. Dr. Falk Leichsenring
University of Giessen
Department of Psychosomatics and Psychotherapy
Ludwigstr. 76, 35392 Giessen, Germany
Fon: +49-641-99 45660
Fax: +49-641-99 45664
Mail: Falk.Leichsenring@psycho.med.uni-giessen.de

With more than 100 non-inferiority or equivalence trials published per year in many areas of research (Piaggio *et al.*, 2012), statistical and methodological issues involved in these trials become increasingly important. A recent article by Rief and Hofmann (2018) suggests, however, that some of these issues are not sufficiently clear. For this reason, central issues will be discussed here and some misunderstandings will be addressed.

Equivalence and non-inferiority margins

For defining a non-inferiority or equivalence margin (i.e. the minimum difference important enough to make treatments nonequivalent), no generally accepted standards exist. In 332 equivalence or non-inferiority medical trials a median margin of 0.50 standard deviations was found (Lange and Freitag, 2005), corresponding quite well to the value of 0.42 reported by Gladstone and Vach (2014). Only five studies used margins < 0.25 (Gladstone and Vach, 2014) and only 12% of studies margins ≤ 0.25 (Lange and Freitag, 2005).

In psychotherapy research, margins ranging, for example, from 0.24 to 0.60 have been proposed (e.g. Steinert *et al.*, 2017, p. 944). For a meta-analysis of psychodynamic therapy (PDT) including different mental disorders, Steinert *et al.* (2017) chose a margin of g=0.25, which is among the smallest margins used in psychotherapy and medical research (Gladstone and Vach, 2014, Figure 2, Steinert *et al.*, 2017, p. 944). This margin is very close to both (a) the threshold for a minimally important difference specifically suggested for depression (0.24, Cuijpers *et al.*, 2014b), and (b) the margin recommended by Gladstone and Vach (2014) to protect against degradation of treatment effects in non-inferiority trials (d= - 0.23).

In their recent correspondence article, Rief and Hofmann (2018) make a quite different proposal, recommending margins not to fall below 90% of the uncontrolled

effect size of the established treatment. This proposal, however, is associated with several problems described in more detail in Table 1, particularly regarding the clinical significance of the suggested margin and its implications for sample size determination, rendering non-inferiority trials in psychotherapy research virtually impossible (Table 1).

Statistical hypotheses in equivalence and non-inferiority testing

In equivalence testing the null and alternative hypothesis of superiority testing are reversed and the statistical *alternative* hypothesis is consistent with the assumption of equivalence (Lesaffre, 2008, Walker and Nowacki, 2011). To test for equivalence, two one-sided tests are performed determining whether the upper and the lower boundary of the CI are included in the margin, whereas, for testing non-inferiority, one one-sided test inspecting the lower boundary is used (Lesaffre, 2008; Walker and Nowacki, 2011). A statistically significant result implies here that the effect size and its CI are within the margin, demonstrating equivalence or non-inferiority (Walker and Nowacki, 2011). A recent meta-analysis testing equivalence of PDT to other approaches established in efficacy reported a significant result indicating that the effect sizes and their CIs were completely included in the margin (Steinert et al., 2017). Thus, the recently given interpretation by Rief and Hofmann (2018, p. 2) that Steinert et al. (2017) '… found a significant disadvantage of PDT [psychodynamic therapy] compared with other treatments (including CBT)' is simply wrong (Lesaffre, 2008; Walker and Nowacki, 2011).

Equivalence vs. non-inferiority testing

Equivalence and non-inferiority testing need to be differentiated (Treadwell *et al.*, 2012). In non-inferiority testing, for example, the test treatment is expected to be

superior to the standard treatment in measures not related to efficacy such as side effects or costs (Treadwell *et al.*, 2012). Rief and Hofmann do not make this differentiation. In fact, the meta-analysis by Steinert *et al.* (2017), for example, was a test of equivalence, not of non-inferiority as suggested by Rief and Hofmann (2018).

Assay sensitivity and constancy of study conditions

Equivalence and non-inferiority testing require that the efficacy of the comparator is ensured and that the study conditions are comparable with in which the efficacy of the comparator was established (Treadwell et al., 2012). In those context, Rief and Hofmann (2018) claim that specific issues of (low) study quality favour non-inferiority results, e.g. low response rates found in specific studies or low treatment integrity. Again, however, these claims are not supported by evidence (Table 1). This applies to several further issues put forward by Rief and Hofmann (2018) which are briefly discussed in Table 1, for example to the relationship between equivalence testing and the number of studies available for a specific treatment (Table 1).

Conclusions

Equivalence and non-inferiority testing pose specific methodological problems (Piaggio et al., 2012; Treadwell et al., 2012), for example, in defining a margin, statistical testing, and ensuring the efficacy of the comparator or comparability of study conditions (Table 1). Conclusions about equivalence and noninferiority testing differing from Rief and Hofmann's (2018) are presented which are more consistent with the available evidence and usual standards across a range of scientific disciplines.

References

Connolly Gibbons, M. B., Gallop, R., Thompson, D., 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*.

Cuijpers, P., Cristea, I. A., Karyotak, E., et al. (2016). How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence *World Psychiatry* **15**, 245-258.

Cuijpers, P., Turner, E. H., Koole, S. L., et al. (2014). What is the threshold for a clinically relevant effect? The case of major depressive disorders. *Depression and Anxiety* **31**, 374-8.

Driessen, E., Van, H. L., Don, F. J., et al. (2013). The efficacy of cognitive-behavioural therapy and psychodynamic therapy in the outpatient treatment of major depression: a randomized clinical trial. *American Journal of Psychiatry* **170**, 1041-50.

Gladstone, B. P. & 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.

Lange, S. & Freitag, G. (2005). Therapeutic Equivalence – Clinical Issues and Statistical Methodology in Noninferiority Trials Choice of Delta: Requirements and Reality – Results of a Systematic Review. *Biometrical Journal* **47**, 12-27.

Lesaffre, E. (2008). Superiority,equivalence, and non-inveriority trials. *Bulletin of the NYU Hospital for Joint Diseases* **66**, 150-4.

McGlothlin, A. E. & Lewis, R. J. (2014). Minimal clinically important difference: defining what really matters to patients. *JAMA* **312**, 1342-3.

Munder, T., Brutsch, O., Leonhart, R., et al. (2013). Researcher allegiance in psychotherapy outcome research: an overview of reviews. *Clinical Psychology Review* **33**, 501-11.

Persons, J. B., Bostrom, A. & Bertagnolli, A. (1999). Results of randomized controlled trials of cognitive therapy for depression generalize to private practice. *Cognitve Therapy abd Research* **23**, 535-548.

Piaggio, G., Elbourne, D. R., Pocock, S. J., et al. (2012). Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA* **308**, 2594-604.

Rief, W. & Hofmann, S. G. (2018). Some problems with non-inferiority tests in psychotherapy research: psychodynamic therapies as an example. *Psychological Medicine*, 1-3.

Steinert, C., Munder, T., Rabung, S., et al. (2017). Psychodynamic Therapy: As Efficacious as Other Empirically Supported Treatments? A Meta-Analysis Testing Equivalence of Outcomes. *American Journal of Psychiatry* **174**, 943-953.

Thoma, N. C., McKay, D., Gerber, A. J., et al. (2012). A quality-based review of randomized controlled trials of cognitive-behavioral therapy for depression: an assessment and metaregression. *American Journal of Psychiatry* **169**, 22-30.

Treadwell, J. R., Uhl, S., Tipton, K., et al. (2012). Assessing equivalence and noninferiority. *Journal of Clinical Epidemiology* **65**, 1144-9.

Walker, E. & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine* **26**, 192-6.

Webb, C. A., deRubeis, R. J. & Barber, J. (2010). Therapist adherence/competence and treatment outcome: a meta-analytic review. *Journal of Consulting and Clinical Psychology* **78**, 200-211.

Table 1. Further methodological issues of equivalence and non-inferiority testing

1. Defining a margin and sample sizes	
(a) Minimal important difference	A margin needs to reflect a minimal important difference, and some small difference may not be clinically meaningful (McGlothlin and Lewis, 2014). Rief and Hofmann (2018) recommend margins not to fall below 90% of the uncontrolled effect size of the established treatment. For an uncontrolled effect size of d=0.5, for example, Rief and Hofmann´s proposal implies a very small margin of 0.05. This margin corresponds to differences, for example, in the Hamilton Rating Scale for Depression and the Hamilton Anxiety Rating Scale of 0.28 and 0.35 scale points which can hardly be considered clinically relevant. As shown above, most researchers agree on larger margins (Lange and Freitag, 2005; Gladstone and Vach, 2014; Steinert et al., 2017).
(b) Sample size	Furthermore, Rief and Hofmann´s proposal would have far-reaching consequences. For demonstrating non-inferiority with a power of 0.80 using a margin of d=0.05 and applying one one-sided test at α=0.025 (Lesaffre, 2008) would require 2 x 6281 subjects. In psychotherapy research, sample sizes like this can hardly be realized, rendering non-inferiority trials in this field virtually impossible.
2. Study quality	
Rief and Hofmann (2018) claim that specific issues of (low) study quality favour non-inferiority results. Again, as shown in the following, several claims are not supported by evidence.	
(a) Treatment integrity	Rief and Hofmann (2018) argue that adherence and competence are key in non-inferiority testing. However, a comprehensive meta-analysis did not show a relationship between adherence/competence and outcome (Webb et al., 2010).
(b) Concurrent drug treatments	Rief and Hofmann's (2018) claim that concurrent drug treatments in both treatment arms reduce the differences between treatments in favour of non-inferiority is presently open to further research.
(c) Intent-to-treat analyses	Whether intent-to treat analyses compensating for missing data carry the risk of diluting treatment differences in non-inferiority trials (Rief and Hofmann, 2018, p. 2) is open to further research (Lesaffre, 2008, Walker and Nowacki, 2011).
(d) Efficacy of the comparator	From the relatively low response rates reported by two studies (Connolly Gibbons et al., 2016, Driessen et al., 2013), Rief and Hofmann conclude that the comparator (CBT) may not have been adequately implemented to reach its typical therapeutic effects. However, this claim is not supported by evidence for several reasons. • Concluding from a study´s result which does not meet the researcher´s expectation that its quality was low is scientifically questionable. Results contradicting the researcher's hypothesis may provide important information. Poor study quality needs to be demonstrated independently of study results. • In fact, the studies by Connolly Gibbons et al. (2016) and Driessen et al. (2013) included CBT supervisors to ensure adequate implementation of CBT. The adequate treatment fidelity ratings of both studies support the notion that the comparator was adequately implemented. • The study by Connolly Gibbons et al. (2016) was a community study, for which lower response rates are common. For instance, Persons et al. (1999) found that only 17% of patients receiving CBT in primary care showed both reliable change and clinically significant change. In the Connolly Gibbons trial, 28% of CBT patients met criteria for both,[1] indicating that CBT delivered in this study was effective. • Thus, there is no evidence that in these studies low-effective versions of CBT were implemented favoring non-inferiority. The meta-analysis by Steinert et al. (2017) discussed by Rief and Hofmann in this regard included only studies in which the efficacy of the comparator was established in previous studies and in which the treatments and the comparators were adequately implemented by use of

	treatment manuals and treatment fidelity measures (e.g. training, supervision, and/or treatment integrity checks). Further, the mean quality rating for the studies included in this meta-analysis was relatively high (35.5) compared to that reported for CBT of depression (25.8) (Thoma et al., 2012).
(e) Researcher allegiance	Researcher allegiance has a major impact on comparative psychotherapy outcome research (Munder et al., 2013). It is highly relevant for both superiority and equivalence testing. For this reason, Steinert et al. (2017, p. 945, 947) controlled for researcher allegiance both at a statistical and at an experimental level by including representatives of both PDT and CBT (adversarial collaboration). In spite of these careful procedures Rief and Hofmann (2018, p. 2) suggest that the interpretation of study results was influenced by the financial sponsor of the study. Steinert et al. (2017, p. 951), however, clearly stated that the sponsor did not have any influence on the design, the evaluation, and the interpretation of this meta-analysis. Furthermore, an adversarial collaboration was established precisely to prevent allegiance effects. This is true for the studies by Driessen et al. (2013) and Connolly Gibbons et al. (2016)[2], too.
(f) Equivalence testing vs. number of studies	Rief and Hofmann (2018, p. 2) state that due to the larger number of studies fr a specific therapeutic approach (CBT) the CIs of the effect sizes for CBT are smaller than those for other approaches. They use this point to argue that success would be more reliably achieved with CBT, even if equivalence had been demonstrated. This approach is questionable for several reasons. • Equivalence testing is confused here with issues of reliability. • Only the CIs of randomized head-to-head comparisons may be directly compared, otherwise study conditions may differ. Taking risk of bias into account, the large number of CBT studies shrinks to 11 low-bias studies for depression and 21 studies for anxiety disorders (Cuijpers et al., 2016). For this reason Cuijpers et al. (2016, p. 245) concluded from their meta-analysis that the effects of CBT are "uncertain and should be considered with caution", implying less confidence in the results of CBT than suggested by Rief and Hofmann (2018)..
(g) Deductive hypothesis testing vs. inductive conclusions	Steinert et al. (2017) tested the hypothesis that PDT is as efficacious as treatments with established efficacy. This hypothesis was corroborated in a strict test that included a small margin, a control of researcher allegiance, and adequately implemented comparators established in efficacy (Steinert et al., 2017). Steinert et al. (2017) have never claimed that their results may be inductively generalized to conditions for which no studies of PDT exist, such as insomnia as suggested by Rief and Hofmann (2018).

[1] Paul Crits-Christoph, personal communication, 16 February 2018.

[2] Paul Crits-Christoph, personal communication, 26 February 2018.