Psychological Medicine

cambridge.org/psm

Correspondence

Cite this article: Jenkins PE (2019). Treatment of anorexia nervosa: is it lacking power? *Psychological Medicine* **49**, 1055–1056. https://doi.org/10.1017/S0033291718003434

Received: 23 October 2018
Accepted: 25 October 2018
First published online: 20 November 2018

Author for correspondence:
Paul E. Jenkins, E-mail: pej106@gmail.com

CAMBRIDGE UNIVERSITY PRESS

Treatment of anorexia nervosa: is it lacking power?

Paul E. Jenkins

School of Psychology and Clinical Language Sciences, University of Reading, RG6 6AL, UK

Two recent studies in *PSM* (Brockmeyer *et al.*, 2018; Murray *et al.*, 2018) have independently synthesised recent evidence on treatments for anorexia nervosa (AN) in adults and have reached similar, sobering, conclusions. In essence, despite the large amount of time, effort, and money that has been invested in evaluating psychological treatments (one, for example, reviewed studies comprising 2092 patients in 19 trials over 5 years), 'no single psychotherapy has emerged as clearly superior to others in the treatment of adults with AN' (Brockmeyer *et al.*, 2018, p. 1250). Whilst this particular conclusion is largely supported by the evidence, results have often been interpreted as meaning that those psychological therapies that were evaluated are, therefore, equivalent in effectiveness. Although it has been argued that this supports a 'common factors' approach to treatment (e.g. Lose *et al.*, 2014), one shortcoming is often given insufficient attention: low statistical power.

When comparing two treatments of an illness, the researcher's aim is critical. Consider a novel treatment, treatment B, and an existing treatment, treatment A. Is the aim to determine that treatment B is: (1) different from treatment A (either better or worse); (2) at least as effective as treatment A, or (3) equivalent to treatment A (Tamayo-Sarver *et al.*, 2005)? The apparent confusion around equivalence and non-inferiority trials has been raised by other authors (e.g. see Leichsenring *et al.*, 2018), but suffice it to say here that the hypotheses (null and alternative) differ for each of these aims such that use of a traditional comparative test – as might be used in scenario (1) – is likely to be inappropriate in tests of non-inferiority and equivalence (Walker and Nowacki, 2011).

Sticking with perhaps the simplest example: a researcher wishes to find out that treatment B is superior to treatment A. A power calculation is vital in determining the minimum sample size needed to detect an effect. Looking at the two review articles, effect sizes of the difference between two treatments on body mass index (BMI) outcomes rarely exceeded $d = 0.30$, a small-to-medium-sized effect (and were typically smaller than this, particularly for psychological outcomes). In the absence of an agreed non-inferiority margin (or, 'clinically acceptable difference'), this would seem to be a reasonable assumption of magnitude and is less than half of the anticipated effect of the comparator treatment (for BMI, for example, effect sizes are usually around $d = 0.6$–1.00 for established treatments; e.g. Zipfel *et al.*, 2014). Given conventions around acceptable error rates (i.e. $\alpha = 0.05$, $1 - \beta = 0.80$), this would suggest a minimum sample size of 139 in each arm to demonstrate that treatment B is superior to treatment A. None of the studies of outpatient trials in adults reached this level. This issue is emphasised when looking at confidence intervals of treatment comparisons, which are often large and encompass what might be proffered as a clinically acceptable difference.

Even if other study biases are limited, low statistical power will engender a tendency towards showing more false negatives. By illustration, in their review, Brockmeyer *et al.* (2018) considered only studies 'with a minimal sample size of $n = 100$' (p. 1229) and included having a 'sample size $n > 30$ in each condition' (p. 1229) in their quality appraisal. Although this may seem encouraging, it remains possible that any lack of differences found between treatments rests on low statistical power; a sample size of 30 seems arbitrary and, if based on the hypothesised effectiveness (effect size) of one treatment, is unlikely to be sufficient to detect an effect between two treatments. Studies with low statistical power also risk reporting findings that are true when in fact they are not and over-estimating the magnitude of those effects (Button *et al.*, 2013). This can have impacts on later research, whereby sample sizes determined on 'historical precedent rather than through formal power calculation' may hamper attempts at replication (Button *et al.*, 2013, p. 367).

This brief summary echoes the conclusions of Rief and Hofmann (2018) that the scientific community needs to consider issues around study design more seriously. Of concern, a number of authors have repeatedly argued for the issue of low statistical power to be addressed, with studies suggesting that the problem is, in fact, endemic (e.g. see Le Hananff *et al.*, 2006; Button *et al.*, 2013; Vankov *et al.*, 2014). Power to detect an effect ought to be an essential element of research design within null-hypothesis significance testing; failing to meet a minimum sample size is likely to render subsequent conclusions questionable at best. Guidance has been published in the reporting non-inferiority and equivalence trials (Piaggio *et al.*, 2012).

That researchers and other stakeholders invest so much in evaluating treatments of AN likely reflects both the severity of the illness and the positive intent of those wishing to eradicate it. Failing to recruit sufficient numbers casts doubt over most conclusions, can influence later meta-analyses (e.g. Peters *et al.*, 2006), and, ultimately, does not do justice to those who volunteer to participate in these studies. If sufficient investment is given (alongside appropriate structural change; see Vankov *et al.*, 2014; Higginson and Munafò, 2016), rewards in terms of treatment outcomes should begin to emerge more clearly.

Conflict of interest. None.

References

Brockmeyer T, Friederich H-C and Schmidt U (2018) Advances in the treatment of anorexia nervosa: a review of established and emerging interventions. *Psychological Medicine* **48**, 1228–1256.

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ and Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 365–376.

Higginson AD and Munafò MR (2016) Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biology* **14**, e2000995.

Le Hananff A, Giraudeau B, Baron G and Ravaud P (2006) Quality of reporting of noninferiority and equivalence randomized trials. *JAMA* **295**, 1147–1151.

Leichsenring F, Abbass A, Driessen E, Hilsenroth M, Luyten P, Rabung S and Steinert C (2018) Equivalence and non-inferiority testing in psychotherapy research. *Psychological Medicine* **48**, 1917–1919.

Lose A, Davies C, Renwick B, Kenyon M, Treasure J and Schmidt U (2014) Process evaluation of the maudsley model for treatment of adults with anorexia nervosa trial. Part II: patient experiences of two psychological therapies for treatment of anorexia nervosa. *European Eating Disorders Review* **22**, 131–139.

Murray SB, Quintana DS, Loeb KL, Griffiths S and Le Grange D (2018) Treatment outcomes for anorexia nervosa: a systematic review and meta-analysis of randomized controlled trials. *Psychological Medicine* 1–10, DOI: 10.1017/S0033291718002088.

Peters JL, Sutton AJ, Jones DR, Abrams KR and Rushton L (2006) Comparison of two methods to detect publication bias in meta-analysis. *JAMA* **295**, 676–680.

Piaggio G, Elbourne DR, Pocock SJ, Evans SJW and Altman DG (2012) Reporting of noninferiority and equivalence randomized trials. Extension of the CONSORT 2010 statement. *JAMA* **308**, 2594–2604.

Rief W and Hofmann SG (2018) Some problems with non-inferiority tests in psychotherapy research: psychodynamic therapies as an example. *Psychological Medicine* **48**, 1392–1394.

Tamayo-Sarver JH, Albert JM, Tamayo-Sarver M and Cydulka RK (2005) How to determine whether your intervention is different, at least as effective as, or equivalent: a basic introduction. *Academic Emergency Medicine* **12**, 536–542.

Vankov I, Bowers J and Munafò MR (2014) On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology* **67**, 1037–1040.

Walker E and Nowacki AS (2011) Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine* **26**, 192–196.

Zipfel S, Wild B, Groß B, Friederich H-C, Teufel M, Schellberg D, Giel KE, de Zwaan M, Dinkel A, Herpertz S, Burgmer M, Löwe B, Tagay S, von Wietersheim J, Zeeck A, Schade-Brittinger C, Schauenburg H and Herzog W (2014) Focal psychodynamic therapy, cognitive behaviour therapy, and optimized treatment as usual in outpatients with anorexia nervosa (ANTOP study): randomised controlled trial. *The Lancet* **383**, 127–137.