Correspondence

Author for correspondence:
James Lock, E-mail: jimlock@stanford.edu

CAMBRIDGE
UNIVERSITY PRESS

When meta-analyses get it wrong: response to 'treatment outcomes for anorexia nervosa: a systematic review and meta-analysis of randomized controlled trials'

James Lock[1], Helena Chmura Kraemer[1], Booil Jo[1] and Jennifer Couturier[2]

[1]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA 94305, USA and [2]Department of Psychiatry and Neurosciences, McMaster University, Hamilton, ON, Canada

Systematic reviews and meta-analyses are important tools in evaluating data supporting evidence-based medicine (Ioannidis, 2016). At the same time, results from meta-analyses can also be confusing, misleading, and harmful when the methods are misused or misinterpreted. It is unfortunate that when endeavoring to provide new insights about treatment using a meta-analysis, the recent paper entitled 'Treatment Outcomes for Anorexia Nervosa: A Systematic Review and Meta-Analysis of Randomized Controlled Trial' (Murray *et al.*, 2018) may have some of these unintended consequences. There were a variety of methodological decisions made when conducting this meta-analysis that limit the value of its conclusions, such as the inclusion of studies with non-comparable interventions and/or controls, use of widely varying assessment end points as end of treatment (EOT), the combining of highly differing prognostic age groups, and reliance on grossly underpowered studies.

It is mandatory in a meta-analysis that meta-analysts locate and evaluate all studies done on a particular issue. However, it is the responsibility of the meta-analysts to exclude from their analysis studies of questionable validity, and to separate studies addressing different research questions. To include all studies one might find in a literature search is highly problematic. Including all types of treatment for AN in this report creates yet another problem, unfortunately common in meta-analyses – the comparison of 'apples to oranges' rather than 'apples to apples'. Treatment types used in the studies included in the meta-analysis are highly variable: medications, placebo, family therapy, inpatient treatment, family workshops, various forms of cognitive behavioral therapy, and acupuncture. To assume that all these are equally effective is no more acceptable than assuming that all surgeries or all drugs are equally effective for any physical disorder. Including this range of intervention types reduces the ability of the meta-analysis to shed light on the treatment effects of any of them. In a valid and powerful meta-analysis mixing interventions/controls, a non-statistically significant or non-clinically significant result indicates that not all treatments are effective, *not* that all treatments are not effective. 'Absence of proof is not proof of absence'.

Are all the treatments categorized as 'specialized' and 'comparator' treatment groups really the same at the group level used for the analysis? It is hard to see how Quetiapine and Focal Dynamic Psychotherapy, for example, are similar enough to be grouped together (in this case categorized as 'specialized' treatments). Further, the *same* treatments are sometimes categorized as 'specialized' and at other times 'comparator' treatments [e.g. Family-based Treatment (FBT) is a specialized treatment in Lock *et al.*, 2010 and a comparator treatment at others]. The strategy to address this by conducting what the authors call a 'potential moderator analysis' cannot overcome the problems of too much heterogeneity of the treatment types included or the overlapping categorization of treatment type by group.

A further problem is the failure to account systematically for the effect of time in the meta-analysis. For EOT, the time point at which data were collected varies greatly – from as short as 7 weeks to as long as 12 months. It is difficult to imagine how outcomes collected over this wide time interval can be reasonably considered comparable. Relatedly, one of the findings the authors purport that is most important is that while specialized treatment is more adept in promoting weight gain by EOT (as noted, defined very broadly), there was no advantage to specialized treatment over time. The authors admit that only about half of the studies actually included had follow-up data and acknowledge that 'the scarcity of these data compromises the interpretation of results beyond EOT' (Murray *et al.*, 2018). We agree that this is a major limitation for both the research studies and the meta-analysis. Thus, the broad range of time points used of EOT and the lack of robust follow-up data raise major questions about the outcomes reported at either time point.

Moreover, the authors decided to include studies of treatment for both adolescents and adults with AN in the same meta-analysis. On the surface, this might seem reasonable; however, there are considerable data that suggest that available treatments for younger patients (under the age of 18 years) are likely more effective than those available for adults

(Treasure and Russell, 2011; Lock, 2015). Not unlike other illnesses, AN tends to become more intractable with time (Hay *et al.*, 2012; Watson and Bulik, 2013). By combining studies of younger patients with older ones, the effects of treatment for this younger age group are, not surprisingly, obscured. The result of embedding the adolescent AN treatment studies in this meta-analysis is to diminish the treatment effects of at least three treatments for this age group that are effective according to adequately powered randomized controlled trials: FBT (Lock *et al.*, 2010), Adolescent Focused Therapy (Lock *et al.*, 2010), and Systemic Family Therapy (Agras *et al.*, 2014).

In addition to age groups, the authors examined various aspects of heterogeneity across the included studies (e.g. age, type of weight outcome measure, illness duration, year of publication, risk of bias, follow-up length) as potential moderators. However, the current meta-analytic study is in fact underpowered to examine such moderating effects (only 35 studies included), and therefore examining potential moderators could not adequately address the issue of comparing highly heterogeneous studies (i.e. apples to oranges). Interpreting insignificant moderating effects as absence of heterogeneous treatment effects could be misleading. In addition, using mean age as a moderator of effect at the study level is not the same as individual age as a moderator of treatment effect at the patient level, the long-known 'ecological fallacy'.

Moreover, including studies addressing the research question of interest that are grossly underpowered to detect clinically significant treatment effects slows progress to reaching a correct and definitive conclusion (Kraemer and Blasey, 2015). For 80% power using a 5% significance level and a one-tailed t test to detect a moderate treatment effect of $d = 0.5$ (NNT = 4) requires a minimum of 50 per group (Kraemer and Thienemann, 1987). Of the studies included, only seven out of 35 met this threshold. Some of the included studies have <10 subjects in each group. In fact, most of the effect sizes included in the meta-analysis are derived from inadequately powered studies. These underpowered studies introduce 'noise' to the meta-analysis, decreasing power of tests and precision of effect size estimates. Including these data in a meta-analysis not only obscures, rather than clarifies treatment effect (Kraemer *et al.*, 1998), but also encourages more such inadequately powered studies in the future.

The interpretation of treatment effect as significant or insignificant is also problematic (Kraemer and Kupfer, 2006). For example, the authors report that for weight outcomes at EOT $g = 0.16$ with a 95% CI (0.05–0.28) as a significant treatment effect ($P = 0.006$), but found at follow-up $g = 0.11$, with a 95% confidence interval (CI) (−0.04 to −0.27) interpreted as not significant based on the p value ($p = 0.15$), although the CIs are almost completely overlapping (Kraemer, 2017). It is very unlikely that these treatment effects differ from each other. Either both indicate an advantage of specialized treatment or neither does. Further, as the authors noted, attrition rates are considerably high (20–40%), which is particularly true in long-term follow-ups. Given limited resources and difficulties in recruiting a large number of subjects, unless the studies aim for the outcomes at these follow-up assessments as primary end points, the studies are most likely underpowered to detect treatment effects at follow-ups.

Why does all of this matter beyond considerations of improving meta-analytic methods? When a meta-analysis is conducted with the kinds of problems that this one has and thereby reaches conclusions that are questionable, unintended consequences may follow. Meta-analyses are necessary to determine whether consensus has been reached on a particular research question, thus either encouraging or discouraging further research on that question, and providing the evidence base for clinical decision-making. Consequently meta-analyses often have high impact. In this case, the concern is that patients and families will interpret these findings as suggesting that no treatment is effective for AN and either not seek treatment or not seek treatments that actually have an evidence base. At the same time, clinicians could interpret these results as suggesting that there is no reason to learn the evidence-based treatments available for AN and therefore patients will have even more limited access to effective treatment. Finally, researchers will be discouraged from further research on such questions, assuring that any erroneous conclusions here will never be corrected. These are major burdens that those who conduct meta-analyses need to keep in mind when they design, interpret, and publish meta-analyses related to treatment effectiveness.

References

Agras WS, Lock J, Brandt H, Bryson SW, Dodge E, Halmi KA, Jo B, Johnson C, Kaye W, Wilfley D and Woodside B (2014) Comparison of 2 family therapies for adolescent anorexia nervosa: a randomized parallel trial. *JAMA Psychiatry* **72**, 1279–1286.

Hay P, Touyz S and Sud R (2012) Treatment for severe and enduring anorexia nervosa: a review. *Australian and New Zealand Journal of Psychiatry* **46**, 1136–1144.

Ioannidis J (2016) The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Millbank Quarterly* **94**, 485–514.

Kraemer H (2017) 'Evidence'-based medicine in eating disorders research: the problem of 'confetti p values'. *International Journal of Eating Disorders* **50**, 307–311.

Kraemer H and Blasey CLA (2015) *How Many Subjects? Statistical Power Analysis in Research*, 2nd Edn. Los Angeles, London; CA, USA: Sage Publications.

Kraemer H and Kupfer D (2006) Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry* **58**, 990–996.

Kraemer H and Thienemann S (1987) *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA, USA: Sage Publications, Inc.

Kraemer H, Yesavage J, Gardner C and Brooks J (1998) Advantages of excluding underpowered studies in meta-analysis: inclusionist versus exclusionist viewpoints. *Psychological Methods* **3**, 23–31.

Lock J (2015) An update on evidence based psychosocial interventions for children and adolescents with eating disorders. *Journal of Clinical Child and Adolescent Psychology* **44**, 707–721.

Lock J, Le Grange D, Agras WS, Moye A, Bryson S and Jo B (2010) A randomized clinical trial comparing family based treatment to adolescent focused individual therapy for adolescents with anorexia nervosa. *Archives of General Psychiatry* **67**, 1025–1032.

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*. doi: 10.1017/s0033291718002088 [Epub ahead of print].

Treasure J and Russell G (2011) The case for early intervention in anorexia nervosa: theoretical exploration of maintaining factors. *British Journal of Psychiatry* **199**, 5–7.

Watson H and Bulik C (2013) Update on the treatment of anorexia nervosa: review of clinical trials, practice guidelines and emerging interventions. *Psychological Medicine* **43**, 2477–2500.