Editorial for JAMA Psychiatry

To accompany the article by Brown et al.

The Optimism of small Ns – a false economy for psychiatric clinical trials

Author Details:

Shitij Kapur, MBBS, PhD, FRCPC

Faculty of Medicine, Dentistry and Health Sciences

University of Melbourne, Melbourne 3010, Australia

Shitij.kapur@unimelb.edu.au

Tel: +61 03 8344 5894

Marcus Munafò, PhD

School of Psychological Science, University of Bristol, Bristol BS8 1TU, United Kingdom

MRC Integrative Epidemiology Unit at the University of Bristol, Bristol BS8 2BN, United Kingdom

Marcus.Munafo@bristol.ac.uk

In 2013 the *Journal* published an exciting new finding by Hallak et al.[1] : patients with schizophrenia treated with a single infusion of the antihypertensive sodium nitroprusside showed a dramatic, instantaneous and sustained improvement in psychotic and negative symptoms. The Hallak study was not only double-blind, randomized and placebo-controlled, but the authors had also presented a prospective power analysis, ensured good inter-rater reliability, and reported all the relevant aspects of the trial. Understandably this generated tremendous excitement and several attempts were made to replicate the finding. In this issue, Brown et al.[2] report an equally systematic study that fails to replicate the original finding. In fact, Brown's findings join two other prior studies[3,4] that also failed to replicate the original finding. A careful analysis of the competing studies fails to reveal an obvious explanation for this discrepancy – except that the original finding was based on a rather small sample, just ten patients per treatment arm.

Failure to reproduce such dramatic findings is neither new, nor limited to psychiatry. As provocatively claimed by Ioannidis in 2005, perhaps most published studies are indeed false.[5] Alert to this the Academy of Medical Sciences in the UK published a report on the reproducibility and reliability of biomedical research, and identified the many causes for the low replicability of published studies are plenty – from deficiencies in the conduct and reporting of studies, to incentive structures that promote publication over rigour.[6,7] Yet, in the matter of the sodium nitroprusside studies the methodology of the studies seems be strong and the reporting thorough. But rigour of conduct and reporting cannot overcome the challenge of starting with a small "N".

A review of 83 psychiatric intervention studies indicated that only half had been subject to any attempt at replication.[8] Of those that had been subject to a replication attempt, only 37% replicated

the original effect size, while the others either contradicted the original finding or observed a much smaller effect. The review clearly showed that the larger the original study, the more likely it was to be replicated.

It is a statistical inevitability that (all other things being equal) smaller studies will provide more imprecise estimates of any true effect. Therefore, if a number of small studies are done there will be wider variation of findings (sometimes described as “vibration of effects”). If many small studies are conducted, only those which generate a large effect size (and therefore reach p-value below the conventional 5% threshold) will be published and these findings from small N studies are likely to represent inflated effect size estimates, and at worst, false positives.

Based on this it would be easy to recommend larger sample sizes in general given evidence that studies in the biomedical sciences are usually too small to detect credible or likely effect sizes.[9] But any such general commandment would be too simplistic. Ultimately, studies should be large enough to give a sufficiently precise estimate, but not so large as to be wasteful. Funders and journals increasingly require a justification of sample size. However, sample size calculation is not a perfect science and require an estimation of the expected outcomes in the drug and placebo groups and an estimation of the expected variance. Getting these estimates right and realistic is the heart of the matter.

Hallak et al.[1] predicted an effect size of d = 1.5. How does this compare to effect sizes in this field? We provide three comparators. The study of clinical interventions in medicine and psychiatry across a range of domains indicates a median observed effect size of 0.37 and 0.41 respectively. [10] A review of over 7,000 patients treated with antipsychotics in randomised controlled trials showed

that the effect size of the newer atypical antipsychotics, as compared to placebo, was 0.48.[11] And finally, studies of schizophrenia clinicians show that they consider an improvement from baseline (which includes placebo response + drug response) equivalent to an effect size of d ~ 1.0 to be clinically significant.[12] Thus, seen from these three perspectives, an expectation of d = 1.5 versus placebo looks rather optimistic. The temptation is understandable. One would like to discover the big findings, and such an assumption comes with the payoff that it decreases the N required by sample size calculators. However, such over-optimism comes with two wasteful consequences – a higher than appropriate chance of a false-negative; and in the event of a positive finding, a higher chance that the positive finding might be a false-positive with an inflated effect.

So, what is the right effect size to aim for? There is not, and cannot be, a single answer. What seems prudent is that trials of any new treatment should assume the median observed in the field (which usually in the range of 0.3-0.5), and those who hope for a much larger effect size should be required to provide a strong justification for such optimism. Because, over-optimism which leads to too small a sample may just be a false-economy.

References

1. Hallak JE, Maia-de-Oliveira JP, Abrao J, et al. Rapid improvement of acute schizophrenia symptoms after intravenous sodium nitroprusside: a randomized, double-blind, placebo-controlled trial. *JAMA Psychiatry.* 2013;70(7):668-676.

2. Brown et al. Adjunctive intravenous sodium nitroprusside treatment for outpatients with schizophrenia: a randomized clinical trial. *JAMA Psychiatry.* 2019.

3. Stone JM, Morrison PD, Koychev I, et al. The effect of sodium nitroprusside on psychotic symptoms and spatial working memory in patients with schizophrenia: a randomized, double-blind, placebo-controlled trial. *Psychol Med.* 2016;46(16):3443-3450.

4. Wang X, Zhao J, Hu Y, et al. Sodium nitroprusside treatment for psychotic symptoms and cognitive deficits of schizophrenia: A randomized, double-blind, placebo-controlled trial. *Psychiatry Res.* 2018;269:271-277.

5. Ioannidis JP. Why most published research findings are false. *PLoS Med.* 2005;2(8):e124.
6. Committee TS. *Reproducibility and reliability of biomedical research: improving research practice.* London: Academy of Medical Sciences;2015.

7. Higginson AD, Munafo MR. Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. *PLoS Biol.* 2016;14(11):e2000995.

8. Tajika A, Ogawa Y, Takeshima N, Hayasaka Y, Furukawa TA. Replication and contradiction of highly cited research papers in psychiatry: 10-year follow-up. *Br J Psychiatry.* 2015;207(4):357-362.

9. Dumas-Mallet E, Button KS, Boraud T, Gonon F, Munafo MR. Low statistical power in biomedical science: a review of three human research domains. *R Soc Open Sci.* 2017;4(2):160254.

10. Leucht S, Hierl S, Kissling W, Dold M, Davis JM. Putting the efficacy of psychiatric and general medicine medication into perspective: review of meta-analyses. *Br J Psychiatry.* 2012;200(2):97-106.

11. Leucht S, Arbter D, Engel RR, Kissling W, Davis JM. How effective are second-generation antipsychotic drugs? A meta-analysis of placebo-controlled trials. *Mol Psychiatry.* 2009;14(4):429-447.

12. Hermes ED, Sokoloff D, Stroup TS, Rosenheck RA. Minimum clinically important difference in the Positive and Negative Syndrome Scale with data from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE). *J Clin Psychiatry.* 2012;73(4):526-532.