BJPsych
The British Journal of Psychiatry (2015)
207, 357–362. doi: 10.1192/bjp.bp.113.143701

Replication and contradiction of highly cited research papers in psychiatry: 10-year follow-up

Aran Tajika, Yusuke Ogawa, Nozomi Takeshima, Yu Hayasaka and Toshi A. Furukawa

Background

Contradictions and initial overestimates are not unusual among highly cited studies. However, this issue has not been researched in psychiatry.

Aims

To assess how highly cited studies in psychiatry are replicated by subsequent studies.

Method

We selected highly cited studies claiming effective psychiatric treatments in the years 2000 through 2002. For each of these studies we searched for subsequent studies with a better-controlled design, or with a similar design but a larger sample.

Results

Among 83 articles recommending effective interventions, 40 had not been subject to any attempt at replication, 16 were contradicted, 11 were found to have substantially smaller effects and only 16 were replicated. The standardised mean differences of the initial studies were overestimated by 132%. Studies with a total sample size of 100 or more tended to produce replicable results.

Conclusions

Caution is needed when a study with a small sample size reports a large effect.

Declaration of interest

None.

Copyright and usage

The number of publications in medicine and in psychiatry is increasing exponentially year after year. About 20 million articles have been published in more than 5000 MEDLINE-indexed journals.[1] How do we identify, read and evaluate new information of interest in this sea of research? The impact factor has largely replaced recommendations and reputations as an indicator of the value of scientific journals. According to Journal Citation Reports (http://thomsonreuters.com/journal-citation-reports), 3000 journals have been given impact factors in biomedicine. These do not directly reflect the worth of individual studies. However, when we consider the credibility of a published study, we often refer to the impact factor of the journal in which the study is published. We can also evaluate the importance of an individual medical publication by its own citation count. The common logic behind counting journal or study citations is the belief that highly cited papers must have had a major impact on science. However, frequent citation is no guarantee that the study results are true. Ioannidis identified studies that were cited more than 1000 times among journals with a high impact factor in general medicine and internal medicine; when these studies were compared with subsequent studies, which theoretically had a better-controlled design, only half of the randomised controlled trials (RCTs) and none of the observational studies were replicated.[2] When statistically significant and extremely favourable initial reports of intervention effects were examined, however, it was found that the majority of such large treatment effects had emerged from small studies, and when additional trials were performed the effect sizes typically became much smaller.[3] Psychiatric research may not be immune to these biases.[4–6] Indeed, psychiatry may be more vulnerable than general medicine to publication and citation bias, as psychiatry typically has to rely on 'soft' outcomes, which have been found to lead to results that are less robust than unequivocal and universally agreed 'hard' outcomes (e.g. scores on the Hamilton Rating Scale for Depression or the Positive and Negative Syndrome Scale *v.* death or recurrence of myocardial infarction).[3,7–9] We therefore aimed to examine what proportion of highly cited studies in psychiatry are or are not confirmed by subsequent studies examining the same clinical questions.

Method

We selected three general medicine journals and five psychiatry journals with the highest impact factors for the year 2000 according to Journal Citation Reports. These journals were the *New England Journal of Medicine* (29.51), *JAMA* (15.40), *The Lancet* (10.23), *Archives of General Psychiatry* (11.78), *Molecular Psychiatry* (8.93), *American Journal of Psychiatry* (6.58), *Schizophrenia Bulletin* (6.09) and *Journal of Clinical Psychopharmacology* (5.05). From the original clinical research studies published in these journals for the years 2000 and 2002 we selected studies that claimed the effectiveness of psychiatric treatments in their abstracts. We did not consider studies reporting the non-effectiveness of treatment. We also excluded meta-analyses and some studies in which two or more studies were combined either systematically or non-systematically, because it is impossible to calculate the effect size of a single study from such papers, and also because such studies mixed studies from different periods including those older than 2000. Two investigators examined the titles and abstracts of the relevant references to check whether the study claimed the effectiveness of a certain psychiatric treatment. Disagreement was resolved by a discussion between the two assessors and, where necessary, in consultation with a third author. We then counted the number of citations of each selected article for the 3 years after the publication year using the Web of Science. We finally restricted the articles to those cited more than 30 times in the 3 years after publication, i.e. approximately the top 10% in terms of citation counts.

Subsequent studies

For each of these highly cited studies we searched for subsequent studies conducted up until June 2013 that examined the same

https://doi.org/10.1192/bjp.bp.113.143701 Published online by Cambridge University Press

clinical question, i.e. focused on the same diagnoses and on the same interventions or exposures. The journals that were searched were limited to those indexed in MEDLINE. All the selected studies (i.e. the original studies as well as the subsequent studies) were categorised in terms of evidence level as an RCT, an observational study or a case study (or case series).[10] If two studies were at the same level of evidence hierarchy, the one with the larger sample was regarded as constituting stronger evidence.[2] We selected newer studies whose intervention and control conditions were as similar to those of the previous study as possible. When the newer study had more study arms than the previous one, we checked each arm and selected the most appropriate one, i.e. the one closest to that of the previous study. When the dosage of medication was the focus of the research in the previous study, we searched newer studies using the dosage closest to that of the previous study. When the condition of participants was restricted (e.g. children or adults, acute or chronic disorder), we also matched the condition as closely as possible.

The interrater reliability in the selection of the relevant subsequent study was of paramount importance in this study. We therefore first pilot-tested our reproducibility with regard to a dozen studies. Two authors independently selected one eligible subsequent study according to sample size and research design in MEDLINE. The selection agreed in 9 out of 12 studies; disagreement for the 3 remaining studies was due to simple oversight by one of the two authors, and there was no need for discussion once the study in question was shared. For the remaining studies, therefore, we followed the following procedure: for each study in our cohort the first author screened broadly for relevant ensuing studies using a few important keywords, selected several newer candidate studies with better-controlled designs and then chose the most appropriate one. Another investigator independently selected the most appropriate one among these candidate studies. Disagreements were resolved by a discussion between the two assessors and, where necessary, in consultation with a third author.

Data extraction

Where the study authors presented their primary outcome, we extracted this information. If the authors failed to designate their primary outcome, we regarded the outcome described first as the primary one. The results for the primary outcomes of the original studies were extracted as continuous or dichotomous data. We gave preference to continuous data, because in psychiatry most outcomes use continuous data and also because, in general, continuous data are statistically more powerful than dichotomous data. When we were able to extract neither continuous nor dichotomous data from the original studies (e.g. case series) we extracted the description regarding the benefit and applicability of the treatment. We then identified outcomes of the subsequent study that were the same as or similar to those of the previous study, and extracted relevant data.

Standardised mean differences

Calculation from continuous data

When the studies showed effectiveness for a continuous outcome, we extracted the means and standard deviations of the end-point or change scores and calculated the standardised mean difference (SMD) using the formula (mean 1 − mean 2)/s.d., where s.d. represents the pooled s.d. of the intervention and control groups.

Converting dichotomous outcomes

When the studies showed effectiveness using only dichotomous data, we first calculated the odds ratio (OR) and then converted it into the SMD using the formula $\text{SMD} = (\sqrt{3}/\pi)\ln\text{OR}$.[11]

Study comparisons

We compared the SMD of each previous study with that of a newer study that was at a higher evidence level or at the same level but with a larger sample, and assigned each comparison to one of four categories:

(a) unchallenged: when there was no subsequent study with a higher level of evidence;

(b) contradicted: when the point estimate of the subsequent stronger study was opposite to that of the former, or the benefit and applicability of a previous study were denied;

(c) initially stronger effects: when the original study and the subsequent stronger study both concluded that the intervention was effective and the point estimate of the previous study was not included in the 95% confidence interval of the effect size of the newer study or the effect size of the previous study was 0.2 s.d. units or greater than that of the subsequent study;[12]

(d) replicated: when the original study and the subsequent stronger study both concluded that the intervention was effective and the point estimate of the previous study was included in the 95% confidence interval of the effect size of the subsequent study and the two effect sizes were within 0.2 s.d. units apart or the effect size of the subsequent study was larger than that of the previous study (0.2 s.d. units would signify a small effect difference according to Cohen's rule of thumb).[12]

Other comparisons

When we could not obtain SMDs we compared the benefits and applicability of the two studies and made qualitative judgements. Two investigators made these judgements independently. Disagreements were resolved by discussion between the two assessors, where necessary in consultation with a third author. Two independent raters assessed the quality of the previous and subsequent studies using the Cochrane Collaboration risk of bias tool,[13] which assesses a trial's quality in the following domains: random sequence generation, allocation concealment, masking of participants and personnel, masking of assessment, completeness of outcome data and selective outcome reporting.

Outcomes

Our primary outcome was the percentage of studies where the results were replicated for all the studies, defined as follows:

$$\text{Percentage of replicated studies} = \frac{\text{replicated}}{\text{total} - \text{unchallenged}} \times 100(\%)$$

In subgroup analyses we classified the original studies according to the journals in which they were published, their research design, the diagnoses of the participants and the therapies that were examined (pharmacotherapy, psychotherapy or others). We calculated the percentage of replicated studies for each subgroup as our secondary outcomes.

Statistical analysis

Statistical tests were performed using SPSS version 22.0. The level of significance was set at the conventional level of $P<0.05$ (two-tailed). Differences between SMDs of previous studies and newer studies were tested by Wilcoxon matched pairs signed rank test. The linear relation between the categories of comparison and

https://doi.org/10.1192/bjp.bp.113.143701 Published online by Cambridge University Press

the sample sizes of previous studies was analysed with the Jonckheere–Terpstra trend test. We estimated the threshold between replicated and non-replicated studies by using receiver operating characteristics (ROC) analysis.

Results

In the three general medicine journals 163 articles related to psychiatry and were cited over 30 times (agreement between independent raters 95.0%, $\kappa = 0.68$). In the five psychiatry journals 390 such articles were cited over 30 times. In total 553 articles concerned psychiatry. Among them were 159 articles about psychiatric treatments (agreement between independent raters 83.5%, $\kappa = 0.66$). However, about half of these suggested non-effectiveness or harmful effects. Finally, we found 83 articles that recommended certain psychiatric treatments (Fig. 1). The numbers of articles finally selected from each journal were 9 from the *New England Journal of Medicine*, 14 from *JAMA*, 5 from *The Lancet*, 17 from *Archives of General Psychiatry*, 1 from *Molecular Psychiatry*, 31 from *American Journal of Psychiatry*, 0 from *Schizophrenia Bulletin* and 6 from *Journal of Clinical Psychopharmacology*. These articles are summarised in Table 1 in terms of their study design, diagnosis and treatments examined. They included 74 RCTs, 7 cohort studies and 2 case series. Most interventions were assessed using a scale such as the Hamilton Rating Scale for Depression or the Positive and Negative Syndrome Scale. In a few studies other outcomes such as relapse or readmission to hospital were used. The details of each of these 83 articles in order of citation counts are tabulated in online Table DS1.

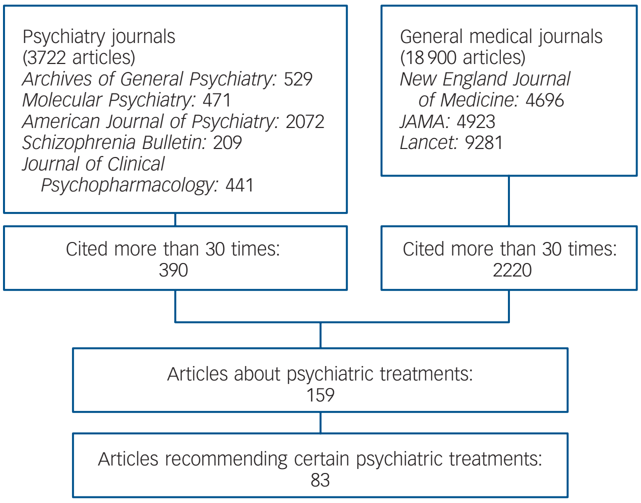
Psychiatry journals
(3722 articles)
Archives of General Psychiatry: 529
Molecular Psychiatry: 471
American Journal of Psychiatry: 2072
Schizophrenia Bulletin: 209
Journal of Clinical Psychopharmacology: 441
General medical journals
(18 900 articles)
New England Journal of Medicine: 4696
JAMA: 4923
Lancet: 9281
Cited more than 30 times: 390
Cited more than 30 times: 2220
Articles about psychiatric treatments: 159
Articles recommending certain psychiatric treatments: 83

Fig. 1 Flowchart of study identification process from studies published in 2000–2002.

Subsequent studies

Of the 83 articles we found that 43 had subsequent studies (52%) that dealt with the same clinical question. The remaining 40 articles (48%) were therefore highly cited but had not been subject to any attempt at replication in the 10 years following their publication (see online Table DS1).The design of the former study in the 43 pairs was RCT in 37 studies (including 4 crossover RCTs and one factorial design RCT), prospective cohort study in 4 studies and case series in 2 studies, whereas all the subsequent studies were RCTs.

Comparisons of study pairs

Sixteen of the 43 studies were categorised as replicated (37%). Two of the 16 studies that replicated the earlier results had SMDs that were 0.2 s.d. units larger than the earlier study (Table 1). The mean SMDs of the original studies and the subsequent studies were 0.72 (s.d. = 0.39) and 0.31 (s.d. = 0.32) respectively. According to Cohen's rule of thumb,[12] the mean SMD of the earlier studies represents a medium to large effect, whereas that of the later studies represents a medium to small effect. There was a highly significant difference between these effect sizes (median difference 0.35, interquartile range 0.03–0.66, $P < 0.001$, Wilcoxon matched pairs signed rank test). The assessment of the risk of bias of

Table 1 Replication and contradiction of highly cited research papers in psychiatry

	Total	Unchallenged	Contradicted	Initially stronger	Replicated	Percentage of replicated studies[a]
Total	83	40	16	11	16	37
Journal						
General medicine	28	16	3	3	6	50
Psychiatry	55	24	13	8	10	32
Design						
Randomised controlled trial	74	37	13	10	14	38
Cohort	7	3	2	1	1	25
Case series	2	0	1	0	1	50
Diagnosis						
Dementia or cognitive impairment	9	4	1	2	2	40
Depression	24	15	4	2	3	33
Mania	4	1	1	0	2	67
Schizophrenia	12	2	5	3	2	20
Dependence	7	4	1	0	2	67
Other	27	14	4	4	5	38
Treatment						
Pharmacotherapy	61	28	9	10	14	42
Psychotherapy	11	7	2	1	1	25
Combined therapy	4	3	1	0	0	0
Other	7	3	3	0	1	25

a. Percentage of the 43 subsequent trials.

https://doi.org/10.1192/bjp.bp.113.143701 Published online by Cambridge University Press

the previous and subsequent studies (agreement between two independent raters 82.5%, κ = 0.70) revealed that their quality was comparable in terms of random sequence generation, allocation concealment, masking, completeness of outcome data and selective outcome reporting (see online Table DS2).

Examples

We cite an example for each category.

Contradicted findings

A prospective cohort study published in the *New England Journal of Medicine* in 2001 was cited 164 times;[14] it suggested that the long-term use of non-steroidal anti-inflammatory drugs (NSAIDs) might protect against Alzheimer's disease. The authors concluded that the relative risk was 0.2. The corresponding SMD that we calculated by using the control risk given in this paper was 0.93. However, a subsequent study with an RCT design published in 2011 negated the effect of NSAIDs over placebo.[15]

Initially stronger effects

One RCT published in the *Archives of General Psychiatry* in 2000, cited 124 times, suggested that low-dose olanzapine (5 mg or 10 mg per day) was superior to placebo in patients with Alzheimer's disease with psychotic and behavioural symptoms, assessed using the sum of the agitation/aggression, hallucinations and delusions items of the Neuropsychiatric Inventory (NPI).[16] We combined the effect size of the two dosages recommended by the author, and the combined SMD was 0.41. The newer RCT on the same topic, which had a larger sample, was published in 2004;[17] in this trial four doses of olanzapine (1 mg, 2.5 mg, 5 mg and 7.5 mg per day) were compared with placebo, and a dosage of 7.5 mg per day was deemed effective when assessed using the NPI total score. First, we calculated the score of the three items of the NPI that the authors of the previous study used and pooled the effect sizes of the 5 mg and 7.5 mg olanzapine groups, which corresponded with the recommended dosages of the earlier study. The combined SMD was 0.04, and the authors concluded that olanzapine was efficacious. We categorised this finding as an initially stronger effect because the original authors emphasised the effectiveness of this treatment.

Replicated finding

Olanzapine demonstrated a greater efficacy than placebo in the treatment of acute bipolar mania in an article published in the *Archives of General Psychiatry* in 2000;[18] this article was cited 129 times. The SMD calculated using the Young Mania Rating Scale was 0.99. In 2009 a newer RCT examining the same clinical question with a larger sample also suggested the effectiveness of olanzapine;[19] the SMD was 1.19.

Subgroup analyses

Table 1 also presents the results of the subgroup analyses. Although there were nominal differences in the percentage of replicated studies for the subgroups that we examined (the percentage of replicated studies was higher for the general medicine journals than for the psychiatry journals, higher for the RCTs than for the observational studies, highest for studies on mania and dependence, and higher for pharmacological treatments than for psychological treatments), none of these differences was statistically significant. Of the 37 RCTs with a subsequent study, the median of the total sample size was 36 in the contradicted studies, 112 in the initially stronger effects studies and 161 in the replicated studies. There was a significant ordered difference among the three groups: the greater the sample size of the initial study, the more often the study was replicated (Jonckheere–Terpstra trend test, $P = 0.004$) (Fig. 2). The ROC analysis revealed that the best pair of sensitivity and specificity was obtained between $n = 92$ and $n = 120$ to distinguish between replicated and non-replicated studies. Approximately 75% of the replicated studies had a total sample size of more than 100. On the other hand, approximately 75% of the contradicted studies had a total sample of fewer than 100.

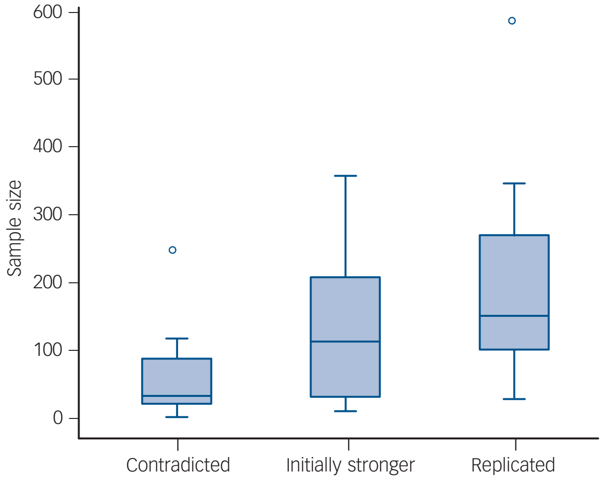

Fig. 2 Distribution of total sample size of the three study categories. Two extreme values are excluded: one was a contradicted study ($n = 1759$) and the other was a replicated study ($n = 3282$).

Discussion

This is the first study to examine the fate of effect size estimates of psychiatric treatments recommended in highly cited clinical studies. We selected highly cited articles published in high-impact journals between 2000 and 2002 and compared their results with studies having a better-controlled design or a similar design but with a larger sample published in the subsequent decade. Of the 83 studies identified, 40 had not been subject to any attempt at replication; of the remaining 43 with replication studies, only 16 (37%) had replicated results. On average the SMD of the initial studies was overestimated by 132% in comparison with the subsequent studies (0.72 *v.* 0.31). The sample size of the initial study was the only statistically significant predictor of confirmation by later studies.

Comparison with general medicine

The percentage of unchallenged studies may be higher in psychiatry than in general medicine. Ioannidis reported that of 45 articles cited more than 1000 times among journals with a high impact factor in various fields of medicine, only 11 remained unchallenged.[2] In our study almost half of the highly cited articles were never re-examined in the following 10 years. We can speculate about the reasons for this. Ioannidis examined articles that had more than 1000 citations,[2] whereas we examined articles that had 30 or more; this may partly explain why the former papers had more replication studies. Another possible reason is the difference in the general levels of research activities in medicine and psychiatry.

https://doi.org/10.1192/bjp.bp.113.143701 Published online by Cambridge University Press

The percentage of contradicted or initially stronger studies also appears to be higher in psychiatry than in general medicine. Ioannidis found that of the 34 articles that had subsequent studies 20 (59%) were replicated,[2] whereas the percentage of replicated studies for our sample was much lower (37%). First, the subtle difference in the definitions of replication between the two studies may explain this difference. Second, however, it should be remembered that 'soft' outcome measures have more potential for bias than 'hard' ones,[20–22] but that 'hard' outcome measures are rare in psychiatry. This could be a reason why the effect size estimates of psychiatric treatments might be more unstable. So long as we have to rely on 'soft' outcome measurements we will need not only to use valid and reliable scales but also to assure validated procedures to administer them in future psychiatric research.

Effect sizes

It is important to note that if a study is significant at exactly $P = 0.05$ then the probability of finding statistical significance in a replication study (assuming that the replication followed exactly the same protocol of the original study) is only 50%, and not 95% as would be naively assumed.[23] Cumming recommends that we use effect sizes and their 95% confidence intervals which would give much better information about replication.[23] Our classification and definitions of replication were based on effect sizes, and the percentage of replicated studies thus defined was lower than expected.

Sample size turned out to be a factor in non-replication. In our study the sample sizes of the replicated studies were the largest and those of the contradicted studies were the smallest among the three categories. There was a significant linear relation. Trikalinos *et al* investigated effect sizes in cumulative meta-analyses of RCTs of psychiatric treatments and found that the magnitude of the effect size in mental health could change considerably.[6] If only 100 patients were randomised there could be a 3-fold to 5-fold relative change in the odds ratio when additional studies were combined; these changes would be relatively small when the cumulative sample size exceeded 1000. Such tendencies are not unique to psychiatry, however. Ioannidis noted that the observed effects of underpowered studies were inflated.[24] He also ran some simulation studies and suggested that, even if the RCT is well performed, the percentage of replicated studies of an underpowered RCT could be as low as 23%.[9]

The effect sizes found in our sample of subsequent studies, rather than those found in the initial studies, in fact appear to be in line with those in psychiatry and general medicine.[25] Leucht *et al* compared SMDs found in comprehensive meta-analyses in psychiatric and general medicine pharmacotherapy broadly, and found the median SMDs of psychiatric drugs and of general medicine ones to be similar (0.41 *v.* 0.37).[25] Caution is thus needed when reading an article reporting large or very large effects of novel treatments.

Limitations

Our study is not without its limitations. First, no newer study posed exactly the same clinical question as the earlier one. Consequently, there were bound to be differences in inclusion criteria, interventions, control conditions and outcome measures (e.g. age, dose range and measurement scale). The judgement 'almost the same' was thus susceptible to some subjective decisions. In order to avoid systematic and random errors caused by these and other inevitably subjective elements in our judgements as much as possible, we used two or more independent raters throughout the study procedures where possible, and were able to demonstrate a satisfactory degree of interrater agreement. However, we could not completely avoid arbitrariness in some decisions and judgements in our study. Second, generally speaking, the quality of masked RCTs is known to be higher than that of non-masked ones,[26] but we did not consider the masking of RCTs in our subgroup analyses. Such masking depends on the condition and theme of the study; for example, therapists and patients cannot be unaware of treatment in psychotherapy trials, and even in pharmacotherapy cluster randomised mega-trials are often conducted without masking (blinding). In other words, we reasoned that masking could be confounded with the types of interventions and the study designs.

Study implications

The clinical and research implications of our findings are clear. Clinicians should be more judicious when they read research studies, even if the studies are published in high-impact journals and are frequently cited. Even more caution is needed when the study had a small sample size and reported a large effect. Researchers should strive towards studies with larger samples and should employ reliable and valid measurements, and journal editors should place greater value on studies with a larger sample and possibly a smaller effect than eye-catching new studies with a small sample and a large effect.

Aran Tajika, MD, **Yusuke Ogawa**, MD, MPH, PhD, **Nozomi Takeshima**, MD, **Yu Hayasaka**, MD, **Toshi A. Furukawa**, MD, PhD, Department of Health Promotion and Human Behaviour, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto, Japan

Correspondence: Dr Aran Tajika, Department of Health Promotion and Human Behaviour, Kyoto University Graduate School of Medicine/School of Public Health, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan. Email: aran.tajika28@gmail.com

First received 23 Dec 2013, final revision 14 Nov 2014, accepted 17 Nov 2014

References

1 US National Library of Medicine. *Fact Sheet. MEDLINE, PubMed, and PMC (PubMed Central): How Are They Different?* US National Library of Medicine, 2015 (http://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html).

2 Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005; **294**: 218–28.

3 Pereira TV, Horwitz RI, Ioannidis JP. Empirical evaluation of very large treatment effects of medical interventions. *JAMA* 2012; **308**: 1676–84.

4 Nieminen P, Rucker G, Miettunen J, Carpenter J, Schumacher M. Statistically significant papers in psychiatry were cited more often than others. *J Clin Epidemiol* 2007; **60**: 939–46.

5 Hunt GE, Cleary M, Walter G. Psychiatry and the Hirsch h-index: the relationship between journal impact factors and accrued citations. *Harv Rev Psychiatry* 2010; **18**: 207–19.

6 Trikalinos TA, Churchill R, Ferri M, Leucht S, Tuunainen A, Wahlbeck K, et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *J Clin Epidemiol* 2004; **57**: 1124–30.

7 Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry* 1960; **23**: 56–62.

8 Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull* 1987; **13**: 261–76.

9 Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005; **2**: e124.

10 Ho PM, Peterson PN, Masoudi FA. Evaluating the evidence: is there a rigid hierarchy? *Circulation* 2008; **118**: 1675–84.

11 Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Stat Med* 2000; **19**: 3127–31.

12 Cohen J. *Statistical Power Analysis in the Behavioral Sciences*. Erlbaum, 1988.

https://doi.org/10.1192/bjp.bp.113.143701 Published online by Cambridge University Press

13 Higgins J, Green S. *Cochrane Handbook for Systematic Reviews of Interventions* (version 5.1.0). Cochrane Collaboration, 2011.

14 In t' Veld BA, Ruitenberg A, Hofman A, Launer LJ, van Duijn CM, Stijnen T, et al. Nonsteroidal antiinflammatory drugs and the risk of Alzheimer's disease. *N Engl J Med* 2001; **345**: 1515–21.

15 Breitner JC, Baker LD, Montine TJ, Meinert CL, Lyketsos CG, Ashe KH, et al. Extended results of the Alzheimer's disease anti-inflammatory prevention trial. *Alzheimers Dement* 2011; **7**: 402–11.

16 Street JS, Clark WS, Gannon KS, Cummings JL, Bymaster FP, Tamura RN, et al. Olanzapine treatment of psychotic and behavioral symptoms in patients with Alzheimer disease in nursing care facilities: a double-blind, randomized, placebo-controlled trial. The HGEU Study Group. *Arch Gen Psychiatry* 2000; **57**: 968–76.

17 De Deyn PP, Carrasco MM, Deberdt W, Jeandel C, Hay DP, Feldman PD, et al. Olanzapine versus placebo in the treatment of psychosis with or without associated behavioral disturbances in patients with Alzheimer's disease. *Int J Geriatr Psychiatry* 2004; **19**: 115–26.

18 Tohen M, Jacobs TG, Grundy SL, McElroy SL, Banov MC, Janicak PG, et al. Efficacy of olanzapine in acute bipolar mania: a double-blind, placebo-controlled study. The Olanzipine HGGW Study Group. *Arch Gen Psychiatry* 2000; **57**: 841–9.

19 McIntyre RS, Cohen M, Zhao J, Alphs L, Macek TA, Panagides J. A 3-week, randomized, placebo-controlled trial of asenapine in the treatment of acute mania in bipolar mania and mixed states. *Bipolar Disord* 2009; **11**: 673–86.

20 Marshall M, Lockwood A, Bradley C, Adams C, Joy C, Fenton M. Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. *Br J Psychiatry* 2000; **176**: 249–52.

21 Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 2008; **336**: 601–5.

22 Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med* 2012; **157**: 429–38.

23 Cumming G. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspect Psychol Sci* 2008; **3**: 286–300.

24 Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology* 2008; **19**: 640–8.

25 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**: 97–106.

26 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; **273**: 408–12.

psychiatry in history

Could Marcus Aurelius be the missing link in the insanity defence?

John H. M. Crichton

It is accepted that Modestinus (*c.* 320 CE) is the earliest source of the insanity defence, but Walker posed the question, '[w]here did Modestinus get his doctrine and his reasoning?' (*Annals of the American Academy of Political and Social Science* 1985). The surviving writing of Modestinus reveals a clue as to its origin:

> 'Truly, if anyone kills a parent in a fit of madness, he shall not be punished, as the deified brothers wrote in a rescript in the case of a man who had killed his mother in a fit of madness; for it was enough for him to be punished by the madness itself, and he must be guarded the more carefully, or even confined with chains' (*The Digest of Justinian*, University of Pennsylvania Press 1985).

It is possible this refers to a case discussed in a letter from joint emperors Marcus Aurelius and Commodus 177–180 CE:

> 'If you have ascertained that Aelius Priscus is so insane that he is permanently mad and thus he was incapable of reasoning when he killed his mother, and did not kill her with the pretense of being mad, you need not concern yourself with the question how he should be punished, as insanity itself is punishment enough. At the same time he should be kept in close custody, and . . . even kept in chains. This need not be done by way of punishment so much for his own and his neighbours' security . . . But since we learn . . . that he is in the hands of friends . . . , your proper course is to summon those in charge of him at the time and enquire how they were so remiss, and then to pronounce on each case separately, according to whether there is any excuse or aggravation for their negligence. The object of keepers for the insane is not merely to stop them from harming themselves, but from destroying others, and if this happens, there is some justification for casting the blame for it on those who were somewhat negligent in their duties (A. Birley, *Marcus Aurelius*, Eyre and Spottiswoode 1966).

The advice in the Aelius Priscus case is similar to Plato's writings both in terms of the reduced responsibility of the mentally unwell homicide perpetrator and also the vicarious responsibility of the 'keepers for the insane' (*The Laws*). Plato was also a major influence on Marcus Aurelius' philosophical writings. This surviving letter provides a possible link between Plato and English common law.

The British Journal of Psychiatry (2015) 207, 362. doi: 10.1192/bjp.bp.114.162065

https://doi.org/10.1192/bjp.bp.113.143701 Published online by Cambridge University Press