Epidemiology and Psychiatric Sciences (2016), **25**, 436–438. © Cambridge University Press 2016
doi:10.1017/S2045796015000888

COMMENTARY TO SPECIAL ARTICLE

Most psychotherapies do not really work, but those that might work should be assessed in biased studies

Received 27 July 2015; Accepted 1 September 2015; First published online 8 March 2016

Key words: Bias, randomised trials, psychotherapy.

Cuipers and Cristea (this issue) present a smart Decalogue on how to prove that psychotherapy is effective even when it is not. Given such an impressive list of 'techniques' for biasing results and empirical estimates for their prevalence, most current evidence on psychotherapies is probably heavily distorted. No one wishes this situation perpetuated. Therefore, using this Decalogue in inverse should decrease bias. However, here I argue that not all of these bias 'techniques' should be eliminated. Some 'techniques' are always detrimental, while some others may have a niche in the development of psychotherapies. Psychotherapies may need to be tested under biased conditions, but bias should be of the right type.

Improper biases

Let me start with the five 'techniques' that are always bad. I will call them 'improper biases'. There is nothing to say in defence of these research practices. Using the 'weak spots' of randomised trials, not concealing treatment allocation to assessors of outcome, analysing only the participants who completed the intervention and ignoring dropouts, using multiple outcome instruments and selectively reporting only the significant ones and not publishing results unless positive represent clear cheating. Any treatment that does not work may seem to work when these recipes are followed. Empirical evidence suggests that this is happening frequently, as summarised by Cuipers and Cristea. Given the prevalence of these 'improper biases' alone (Ioannidis *et al.* 2014), most psychotherapies probably have no or little benefit, even when published literature suggests they are very effective. For example, published psychotherapies for depression have an average $d = 0.69$ (Cuipers *et al.* 2008), but the true average effect may be $d = 0.2–0.3$ or less. For many other conditions the situation may be even worse.

These 'improper biases' vary in the extent to which they are avoidable. Preemptive action is preferable, but some biases are difficult to eliminate entirely. In particular, concealing allocation to assessors of outcomes is sometimes difficult. Even minimal contact with the assessor may reveal what psychotherapy the patient has been on. Moreover, treatment dropouts and losses to follow-up are frequent even in short-term studies and, indeed, they often reflect lack of effectiveness or poor tolerability. However, this is difficult to judge. Imputation methods are better than ignoring missing observations, but still leave substantial uncertainty. What this all means is that at least the other improper biases (those possible to deal thoroughly with) should be eliminated. There is absolutely no reason nowadays for a trial not to be performed with robust randomisation, allocation concealment and pre-specified outcomes and not to get published as pre-specified. I leave some room only about the need to occasionally modify the analysis plan if something ensues during the study conduct that could not be anticipated upfront. Then this still needs to be transparently acknowledged, the modified analysis plan justified and results interpreted with caution.

Potentially proper biases

The other five 'techniques' listed by Cuipers and Cristea are not necessarily always bad. In fact, under some circumstances, they are very appropriate. I will call them 'potentially proper biases'. Let me explain.

When researchers test their own pet intervention themselves they may do everything possible to increase expectations in the participants. Allegiance bias has been documented to be powerful, although its ability to change results varies across interventions (Munder *et al.* 2013). We should set apart the

Address for correspondence: J. P. A. Ioannidis, Departments of Medicine, Health Research and Policy, and Statistics, and Meta-Research Innovation Center at Stanford, Stanford University, 1265 Welch Rd, MSOB X306, Stanford, CA 94305, USA.
(E-mail: jioannid@stanford.edu)

components of allegiance bias that operate through the five ‘improper biases, as discussed above. If we safeguard that this cheating does not occur, it is not unreasonable to have the researcher who developed a method to do the initial testing. If indeed most psychotherapies do not work or have only weak effects, it makes sense that the best experts should perform the first studies on how to implement the therapy. Conversely, administration of therapies by inexperienced students is the best way to decrease the observed effect (Cuipers *et al.* 2008). If the best experts can get no clinically meaningful effect, they can safely quit further testing, since the intervention will do worse in the hands of others. If a psychotherapy cannot attain $d = 0.3$ under optimised circumstances, it is unlikely to attain $d = 0.1$ when used more widely. The incentives system has to be such that a researcher is rewarded for showing that his idea for a new psychotherapy did not work. Given that most new psychotherapies do not really work, perhaps journals should always accept for publication well-done studies with ‘negative’ results, and may request further, additional, replication evidence (Open Science Collaboration, 2015) by independent investigators before they publish significant, ‘positive’ results (Ioannidis, 2006).

This is different to the testing of drugs for diseases such as cancer or cardiovascular disease, where I have argued that the entire clinical testing agenda should be kept independent from the manufacturer (Naci & Ioannidis, 2015). For a 5 mg pill, we do not need to have the CEO of the company by the bedside to administer it. Conversely, a new psychotherapy may need its developer to implement it initially. This gives it the best possible shot. Moreover, most drugs go through a long and expensive screening preclinical process of testing (*in vitro, in vivo*, in animals). Even though biases at these stages are prominent (Ioannidis & Begley, 2015), the whole process, even if inefficient, ends up eliminating many candidate drugs that have weak or unfavourable preclinical evidence (Goodman and Gerson, 2013). Conversely, psychotherapies can hardly be assessed in test-tubes, cell cultures or animals. They emerge from theoretical speculation and currently hit patients with little pre-screening. Thus the odds of success are probably weaker for psychotherapies than for drugs. (Of course, even for drugs, pre-clinical testing may have varying screening ability depending on how well the disease is reproduced in the test-tube, cell or animal. A preclinical model of depression is more elusive than one of clot formation).

Raising the expectations of the participants by boosting the placebo effect is also not a bad idea. Anyhow, most psychotherapies that do work, probably work to a good extent through the placebo effect. Some interventions may be more amenable by boosting the placebo effect. If this can be done efficiently, reproducibly and without cost, that is great! A treatment that exploits and magnifies the placebo effect from $d = 0.2$ to $d = 0.5$ is not worse than the one that achieves the same benefit through a different, more sophisticated mechanism. Conversely, a treatment is undesirable if it affects some favourable mechanism but concomitantly diminishes the placebo effect by a greater amount.

Small sample size is also not a vice on its own. The problem is that small studies are more susceptible to the five ‘improper biases’ than larger ones. A trial of $n = 20$ participants is easier to hide quietly in a file drawer than a trial of $n = 2000$ participants. However, if improper biases are eliminated (as they should), performing a few well-done studies of modest size is a reasonable choice (Inthout *et al.* 2012). If most psychotherapies do not work, it makes sense to test them initially with optimised studies of modest size, barely sufficient to exclude an effect of clinical interest. Otherwise, we waste resources to run large trials on low-yield experimental therapies. Large studies at an early stage make sense only if there is a reasonable chance to see a clinically meaningful effect *and* that clinically meaningful effect is small (thus needing a large study for sufficient power). This is probably more common for new drugs emerging from extensive preclinical screening than for new psychotherapies, as discussed above.

Similarly, using a waiting list control group and not comparing the tested intervention against an existing effective intervention are not necessarily vices on their own. Both of these approaches create a more discernible effect size. This effect size is not spurious (as in the case of the five ‘improper biases’). It simply reflects the comparative effect against nothing, rather than the incremental effect against something else that is already effective. If the few psychotherapies that are effective have small (but still clinically meaningful) effects, testing them against absolutely nothing can help discern their benefit without having to resort to huge sample sizes. If this initial screening is successful, then the relative benefits and overall merits of the new treatment should still be compared against other existing ones with large trials. However most psychotherapies that do not work even against nothing will be quickly screened out with small trials, failing even this favourably biased test. Again, incentives should reward publishing such ‘negative’ results and save the field from wasting effort chasing spurious claims.

Moving from pre-screening trials to large pragmatic trials

To summarise, some biases are always bad and should be eliminated, whenever possible. Other biases may be desirable. Carefully properly biased studies may show

https://doi.org/10.1017/S2045796015000888 Published online by Cambridge University Press

which psychotherapies might work. The few best survivors can then move to large-scale testing in pragmatic randomised trials of many thousands participants and many practitioners. The even fewer survivors at this second stage can be considered for wide, everyday use. From this viewpoint, most current trials of psychotherapies represent merely early stage, pre-screening research that has been done poorly to-date and can be done much better.

J. P. A. Ioannidis

Financial Support

The work of John Ioannidis is supported by an unrestricted gift from Sue and Bob O'Donnell.

References

Begley G, Ioannidis JP (2015). Reproducibility in science: improving the standard for basic and preclinical research. *Circulation Research* **116** (1), 116–126.

Cuipers P, Vanstraten A, Warmerdam L, Smits N (2008). Characteristics of effective psychological treatments of depression: a metaregression analysis. *Psychotherapy Research* **18**, (2), 225–236.

Cuipers P, Cristea I (this issue). How to prove that a therapy is effective even when it is not. *Epidemiology and Psychiatric Services*.

Goodman SN, Gerson J (2013). *Mechanistic Evidence in Evidence-Based Medicine: a Conceptual Framework* [*Internet*]. Agency for Healthcare Research and Quality, Rockville.

Inthout J, Ioannidis JP, Borm GF (2012). Obtaining evidence by a single well-powered trial or several modestly powered trials. *Statistical Methods in Medical Research* Oct 14. [Epub ahead of print].

Ioannidis JP (2006). Journals should publish all "null" results and should sparingly publish "positive" results. *Cancer Epidemiology Biomarkers & Prevention* **15**, 186.

Ioannidis JP, Munafò MR, Fusar-Poli P, Nosek BA, David SP (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Science* **18**, (5), 235–241.

Munder T, Brütsch O, Leonhart R, Gerger H, Barth J (2013). Researcher allegiance in psychotherapy outcome research: an overview of reviews. *Clinical Psychology Review* **33**, 501–511.

Naci H, Ioannidis JP (2015). How good is "evidence" from clinical studies of drug effects and why might such evidence fail in the prediction of the clinical utility of drugs? *Annual Review on Pharmacology and Toxicology* **55**, 169–189.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* **349**, (6251), aac4716.

https://doi.org/10.1017/S2045796015000888 Published online by Cambridge University Press