# Finding large effect sizes – good news or bad news?

**John P.A. Ioannidis** on why we should be cautious, and how a Bayesian perspective and examination of all the evidence may help us approach the truth

In his classic essay on causation, Sir Austin Bradford Hill, put the strength of an association on the top of his considerations:

> What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation?
>
> (1) Strength. First upon my list I would put the strength of the association.
>
> (Hill, 1965)

Strength reflects what we would now call the effect size (*ES*) of an association. If this is so, finding a large *ES* is good news. We are more likely to have found something true, essential, causal. No?

Let us consider two studies A and B. Suppose study B finds a much larger *ES* than study A, say 1.0 vs. 0.2 on some standardised scale, and both studies get the same *p* for testing the null *ES* = 0 (say *p* = .01). This means that the two studies have different weights, i.e. study B is smaller than study A. Is it more likely that A or B has found a true (non-null) effect?

The answer depends on what we think about the nature of true effects. If we believe that effects in this field of research are quite modest, then finding a small *ES* is commensurate with our expectations; a large *ES* would be an oddity. It also depends on how strong the foundations of our beliefs are. If, in experiment after experiment and study after study, we have witnessed consistently (e.g. in well-conducted meta-analyses) that all effects that are robustly replicated in this specific field are small, then an atypically large *ES* either signals something truly major (perhaps a paradigm shift), or is just plain suspicious, precisely because of its large magnitude.

Given that paradigm shifts are probably not common, the sceptical explanation is more likely. For example, most studies on the effectiveness of drugs or psychotherapies for depression suggest modest efficacy at best and some investigators may even question whether they are effective at all in some patients. Now, if a single small study on a new antidepressant or psychotherapy finds a huge *ES*, much larger than what we are used to, the first reaction will probably be that we need to replicate it before we can accept this special extra large size. Most likely there is nothing special; possibly there is nothing at all.

One can see this with formal Bayesian calculations. Let us use a prior that has a spike and smear configuration (Ioannidis, 2008). The spike is the null (*ES* = 0) and the smear is spread as a normal distribution such that on average it is also 0, but it allows equally for both positive and negative *ES* values. Let us assume that the average positive *ES* is likely to be 0.2 (the average negative *ES* is likely to be – 0.2). Then if we observe an *ES* = 0.2 with *p* = .01, the Bayes factor (*BF*) is .16, while if we observe an *ES* = 1.0 with *p* = .01, the *BF* is .45. *BF* is the ratio of the pre-study odds versus the post-study odds that there is an effect. The inverse of BF is telling us how many times the data increase the odds that there is an effect versus that there is none. Therefore in this example, an observed *ES* = 0.2 increases the odds that there is any effect > 6-fold, while an observed *ES* = 1.0 increases the odds that there is any effect barely 2.2-fold. You can run the calculations for yourself in an Excel spreadsheet at www.dhe.med.uoi.gr/software.htm.

What if the typical anticipated effects are indeed large? Then the situation changes completely. If we anticipate that the average positive *ES* should be 1.0, the *BF* conferred by observed *ES* of 0.2 and 1.0 (again with *p* = .01) is .59 and .16, respectively. Different Bayesian variants may lead to somewhat different results, but the key message is the same: it depends on what we think about the ballpark where an *ES* should be, a reflection of our prior knowledge.

Here is a real example. In 1989, a trial reported on 86 women with metastatic breast cancer randomised to receive or not receive supportive-expressive group therapy (Spiegel et al., 1989). The therapy encouraged participants to express feelings and concerns about their disease in the supportive environment of a therapist-led group. The intervention

## Bite size

Finding large effect sizes does not mean that an observed effect is true. One should be cautious with too large effect sizes. These effects may be false positives or may reflect inflated estimates that lead to unrealistic conclusions and wrong choices and actions. A Bayesian approach can be used (for example, with spike and smear priors; see www.dhe.med.uoi.gr/software.htm). This approach can model what our expectations are about possible effect sizes and estimate the Bayes factors corresponding to the observed effect sizes.

references

Goodwin, P.J., Leszcz, M., Ennis, M. et al. (2001). The effect of group psychosocial support on survival in metastatic breast cancer. *New England Journal of Medicine, 345*, 1719–1726.

Hill, A.B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58*, 295–300.

Ioannidis, J.P.A. (2005). Why most published research findings are false. *Public Library of Science Medicine, 2*, e124.

Ioannidis, J.P.A. (2008). Calibration of credibility of agnostic genome-wide associations. *American Journal of Medical Genetics B: Neuropsychiatric Genetics*, published online March 24.

Ioannidis, J.P.A. (in press). Why most discovered true associations are inflated. *Epidemiology*.

Kissane, D. & Li, Y. (2008). Effects of supportive-expressive group therapy on survival of patients with metastatic breast cancer. *Cancer, 112*, 443–444.

Klein, J.B., Jacobs, R.H. & Reinecke, M.A. (2007). Cognitive-behavioral therapy for adolescent depression. *Journal of the American Academy of Child and Adolescent Psychiatry, 46*, 1403–1413.

Spiegel, D., Bloom, J.R., Kraemer, H.C. & Gottheil, E. (1989). Effect of psychosocial treatment on survival of patients with metastatic breast cancer. *The Lancet, 2*, 888–891.

Spiegel, D., Butler, L.D., Giese-Davis, J. et al. (2007). Effects of supportive-expressive group therapy on survival of patients with metastatic breast cancer: A randomized prospective trial. *Cancer, 110*, 1130–1138.

apparently doubled survival from a mean of 18.9 months in the control group to 36.6 months in the intervention group ($p$ = .005). Taking a step back: How much of an effect on survival *should* such interventions have? Most people would probably argue that effects on survival should be small, if at all present; this is a terminal disease, talking about it may make the patient feel better but the widely spread tumour won't go away. Suppose that we anticipate that psychotherapies in terminal diseases, even if they do manage to postpone death, are expected to prolong survival by 10 per cent on average. Then, running the calculations through the Excel spreadsheet, we find that the observed trial result (doubling survival, $p$ = .005) corresponds to *BF* of only .53, the inverse of which is less than 2. Therefore, the trial result actually did not increase even by 2-fold the odds of having any prolongation of survival by supportive-expressive group therapy. If we thought before the study that there is a 3 per cent chance of any prolongation of survival by this therapy, after the study this 3 per cent becomes a little less than 6 per cent.

Several years later, a larger trial of 235 women found no benefit in survival from this same intervention (Goodwin et al., 2001). If anything the point estimate was in the opposite direction and the 95 per cent confidence interval even excluded a 12 per cent relative risk reduction in the risk of death (hazard ratio for death 1.23, 95 per cent confidence intervals 0.88–1.72). More recently, another trial by the same team that had published the very favourable results in 1989 also failed to document survival benefits (Spiegel et al., 2007). Unwilling to let it rest, the authors still claimed that in a small subgroup of 25 women with negative estrogen receptor status, supportive-expressive therapy *tripled* survival (median 29.8 months vs. 9.3 months). How likely is such a huge effect to be true? A letter was soon received (Kissane & Li, 2008) where the negative estrogen receptor subgroup ($N$ = 70) of a larger similar trial had witnessed no survival benefit from the therapy (15.5 vs. 17.2 months).

How about if two studies of equal size observe different effects? Then the larger effect reinforces more our belief that there is *some* (non-null) effect. However, is the larger observed *ES* likely to be more accurate (closer to the true *ES*) than the smaller observed *ES*? We need to dissect an effect to answer this difficult question. Observed effects are made of three components: the true *ES* (if any), random noise and bias. We should ask how big the anticipated true *ES* values are in the field, but also how much is the random noise and how big bias may be in this field.

Random noise can be taken care of with appropriate statistical methods. Textbooks state that random error by definition will have an equal tendency to increase or to decrease observed effects compared with the true ones. This is, however, not the case when the effects of interest are selected based on some statistical threshold, such as statistical significance (e.g. $p < .05$) (Ioannidis, in press). If we select only effects that are formally statistically significant, these are expected to be inflated compared with the true values. We cannot select and estimate at the same time. Selection of extremes is penalised by inflation of the estimate. The research discovery process itself is tightly related to the selection of such 'significant' extremes. The lower the $\alpha$ threshold for selection, the greater the average exaggeration of the observed effect sizes against their true values.

**How much of an effect on survival *should* interventions have?**

The impact of non-random bias is more difficult to tackle than random error. We need to ask: what proportion of evaluated effects may end up being seemingly non-null because of bias, while they are null? If this proportion is at least as large as the proportion of the truly non-null effects among those evaluated in a study, set of studies or whole field of research, then even if we unearth all the true (non-null) effects, the false ones will be at least as many. One can envision situations where the false effects are the large majority among the observed.

## Where does all this lead?

I have discussed elsewhere in more detail the explosive mix that can lead to this dreadful scenario, where a scientific field is plagued by false effects: factors include multiplicity of testing with low pre-study odds, small studies, flexible analysis, conflicts of interest, many furtively competing teams and selective reporting (Ioannidis, 2005). If we have a situation where the majority of seemingly non-null effects are false, then the average *ES* is simply an accurate measure of the net impact of all biases that have shaped this field. Within such a scientific field, a larger *ES* simply reflects a greater impact of bias than a smaller *ES*. Since studies with large *ES* are considered often more successful and most important (based on considerations similar to those expounded by Bradford Hill above), the most applauded studies are simply the ones that bear the greatest impact of their flaws. Fields with larger effects are those that suffer most from bias.

In a less extreme (and possibly common) scenario, bias may be responsible for some but not for all the observed effect. For example, cognitive-behavioural therapy is probably effective in adolescent depression. A meta-analysis of 26 studies found a summary post-treatment *ES* of 0.53, but while the first trials found post-treatment *ES* estimates as large as 1.3–1.6, more recent studies found post-treatment *ES* < 0.40 (Klein et al., 2007). Cumulative meta-analysis showed a decreasing summary *ES* over time. As the meta-analysts observed, the more recent studies were apparently more protected from bias than early studies: they used intent-to-treat analysis, they were conducted in clinical settings, and had overall greater methodological rigor on several design and reporting fronts (Klein et al., 2007).

To summarise: large effects are nice, if they are true. However, one should ask whether large effects are commensurate with the entirety of the prior evidence on the same or similar questions. Thus it is extremely important to be able to have unbiased views about this evidence in its totality, without selective reporting or selective availability of only the most impressive results. With millions of scientists working on a global level, for each question of interest there may be a lot of pertinent evidence to integrate and make sense of. A study cannot be seen in isolation. Second, due to both random errors and biases, too large effects require extra caution. One has to ask how much random error and bias may be operating in the specific study and in the wider field of interest. We have to learn from others working in the same field. A wider view may improve our ability to discern whether a large effect is a large success or a large disaster.

**John P.A. Ioannidis** *is at the University of Ioannina School of Medicine, Ioannina, Greece, and Tufts University School of Medicine, Boston, USA*
*jioannidl@cc.uoi.gr*