Commentary on John P.A. Ioannidis's 'Why Most Published Research Findings are False'

RAY HYMAN

In his provocative essay "Why Most Published Research Findings are False," John P.A. Ioannidis begins with the assumption that most research conclusions are false. He then creates a model to explain this state of affairs.[1] He defines a *research finding* as "any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations." As the editors of *PLoS Medicine* note, Ioannidis's definition fails to distinguish findings about data, conclusions, and hypotheses.[2]

Despite this and other quibbles about Ioannidis's essay, his model provides a useful framework for examining research claims in various domains of inquiry. I found it instructive, for example, to apply his model to the field of parapsychological research. Ioannidis points out that previous writers have argued "that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05."

Parapsychology is a field that depends exclusively for its conclusions on the significance test. J.B. Rhine advocated using the 0.01 level of significance because of the unusual nature of the psi claim. More recently, however, parapsychologists have been routinely using the 0.05 level. This change in the standard for rejecting the null hypothesis increases the number of false research findings by a factor of five. Ioannidis's model suggests other factors that would inflate the rate of false findings in psi research.

The level of significance chosen for the null hypothesis test, by itself, does not tell us the probability that the research finding is false. As I have written elsewhere,[3] the null hypothesis test entails serious logical problems. What the investigator wants to know is the *probability that the null hypothesis is true given the observed outcome*. Unfortunately, the test does not provide this information. Instead, the statistical test states *the probability of the observed outcome given that the null hypothesis is true.* This seemingly subtle distinction has important consequences that researchers routinely ignore. The most important consequence is that we need to consider the *prior or pre-study probability that a finding will be true* to find the posterior or post-study probability that the finding is true.

This principle underlies the famous and controversial Bayes Theorem. This theorem provides a way to revise the original probability that a finding is true given the new data provided by an experiment. The controversy arises because the prior or original probability is often difficult to determine. The relationship between the pre-study probability and the post-study probability is central to Ioannidis's model. This is because for a given level of significance, the probability that a research finding is false increases for lower pre-study probabilities.

We need to consider the prior or pre-study probability that a finding will be true to find the posterior or post-study probability that the finding is true.

Consider parapsychological research findings. Most scientists would consider the prior odds that psi is true as quite low. In the Ioannidis model, the post-study probability that a research finding is false increases as the pre-study plausibility of

Ray Hyman is professor emeritus of psychology at the University of Oregon. E-mail: rayhyman@comcast.net.

that finding goes down, as the significance level goes up, and as the power of the test goes down. Parapsychological claims begin life with low plausibility. Parapsychologists typically complain about the low power that characterizes their studies.[4] The low pre-study probability that psi is real, increasing the significance level from 0.01 to 0.05, and the low power of psi research, guarantees that most of the research findings in parapsychology will be false. This is only part of the story.

Ioannidis includes much more in his model. The model considers bias and also the effects of testing the same association by several independent teams. Ioannidis defines *bias* as "the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced." Bias increases the probability that research findings will be false. Readers who are familiar with my critiques of parapsychological research[5] are aware that I have focused on various flaws that create bias. The biases are just the kinds that, according to the model, greatly enhance the probability that research findings will be false.

In this brief commentary, I will consider just one corollary that follows from Ioannidis's model. This corollary states that, "The smaller the studies conducted in a scientific field, the less likely the research findings are to be true." In other words, a field that conducts many studies with low power will generate many false findings. Ironically, this corollary directly conflicts with the positions of parapsychologists such as Utts and Radin. These commentators argue that the reason that parapsychological research findings may be unreplicable is that they have low power. By pooling studies, the meta-analysis becomes a powerful tool, in their opinion, to show that psi effects, however small, are real.

I have documented a variety of flaws that plague parapsychological research. For the most part, parapsychologists have not denied the existence of such flaws.

Ioannidis uses his model to make the opposite point. According to him, in fields where most of the studies have low power and where the alleged effects are small, "the field is likely to be plagued by almost ubiquitous false claims. . . . " His simulations show that "a meta-analytic finding from inconclusive studies where pooling is used to 'correct' the low power of single studies, is probably false if R [the pre-study odds of being true] is less than or equal to 1:3." This means that if the pre-study probability of a research outcome being true is less than .25, then a significant outcome of a meta-analysis is probably false. Although we do not know the actual pre-study probability that a significant outcome in favor of psi is true, it is reasonable to suppose that it would be much, much smaller than .25.

During my fifty years as a critic of paranormal claims, I have documented a variety of flaws that plague parapsychological research. For the most part, parapsychologists have not denied the existence of such flaws. The most typical defense is to argue that a given flaw, by itself, would be insufficient to account for all the findings.

The parapsychologists could be correct; any given flaw, by itself, may very well be insufficient to account for all the significant parapsychological findings. However, that is not my point. Let's assume that each flaw contributes just a very small amount of bias. The question is, what is the total bias produced by the combination of all these minor biases operating in concert? I tried to answer this question for a few of the flaws that I identified with respect to testing significance in the original ganzfeld psi data base.[6] I ran a simulation using a few of the flaws. The studies in this data base tested their outcomes using the 0.05 level of significance. The simulation showed that, in effect, the experimenters were operating with a significance level of .30 or higher. The false alarm rate was more than six times what was advertised! This simulation used only a few of the flaws and weaknesses I uncovered in that database.

Ioannidis has given us a valuable tool for quantifying the probabilities that the findings from a series of investigations in a given field will be false. Hopefully, this will lead to additional and better ways to gauge how many results from a given program of research are spurious.

Indeed, Ioannidis indicates that the significant results and effect sizes that go with them in certain fields may simply reflect nothing more than bias.

Notes

1. Ioannidis, J.P.A. 2005. Why most published research findings are false. *PLoS Medicine*, 2(8): e124.

2. PloS Medicine Editors. 2005. Minimizing mistakes and embracing uncertainty. *PLoS Medicine*, 2(8), e272.

3. Hyman, R. 2005. Statistics and the test of Natasha. Available at www.csicop.org/specialarticles/natasha2.html. A shorter version was published as Statistics of the Natasha Test: Response to Concerns and Questions, SKEPTICAL INQUIRER, 29(5): 58–60, 2005 (September/October).

4. For example: Utts, J. 1991. Replication and meta-analysis in parapsychology. Statistical Science, 6, 363–403 and Radin, D. 1997. *The Conscious Universe: The scientific Truth of Psychic Phenomena.* NY: HarperEdge.

5. For example: Hyman, R. 1989. *The Elusive Quarry: A Scientific Appraisal of Psychical Research.* Buffalo, NY: Prometheus Books; and Hyman, R. 2003. How *Not* to Test Mediums: Critiquing the Afterlife Experiments. SKEPTICAL INQUIRER, 27(1): 20–30 (January/February).

6. Hyman, R. 1985. The ganzfeld psi experiment: a critical appraisal. *Journal of Parapsychology*, 49, 3–50. □