Medical Hypotheses 113 (2018) 27–29

Contents lists available at ScienceDirect

Medical Hypotheses

journal homepage: www.elsevier.com/locate/mehy

It has not been proven why or that most research findings are false

John C. Ashton

Department of Pharmacology & Toxicology, Otago School of Biomedical Sciences, University of Otago, Dunedin, New Zealand

ARTICLE INFO

Keywords:
Bias
Experiment
Null hypothesis testing
Research finding
Pre-study odds

ABSTRACT

The claim has been made that it can be proven that most published findings in medical, biological, and allied sciences are false and that the reason for this can be proven and explained with a mathematical model. It has not, however, been mathematically proven that most research findings are false, and this can be proven. The model used in the proof is incoherent and has been falsified. Furthermore, advice to researchers derived from the model is misleading and distracts from more important issues in experimental standards.

Introduction

In a very highly cited paper, Ioannidis [1] argued that intrinsic logical properties of the experimental process lead to very high false positive finding rates. To do this he first assumed a model of scientific investigation based on null hypothesis testing (NHT) and then combined the mathematics of NHT with the concept of pre-study odds (that a hypothesis is true) to produce a model with which to estimate false positive finding rates. This approach was derived from the work of Walcholder et al. [2] who used a very similar model to calculate the rate of false positive hits in microarray analysis. Walcholder et al. [2] applied their model to a finite and tightly enumerable domain defined by the possible outcomes constrained by the dimensions of microarrays. Ioannadis [1], by contrast, applied the model to experimental science as such, where no such enumeration of hypotheses is possible. This leads to a paradox, and as a result his analysis and conclusions are incoherent. Furthermore, they are falsified by data and are scientifically pernicious.

The model is incoherent

The critical parameter in the Ioannidis model is *R*, the pre-study odds. Ioannidis also investigates the role of bias and statistical power in his model, but these are minor factors compared with *R* (see Table 4, Ioannidis [1]). The central assumption of the Ioannidis model is that *R* is a real number and can be estimated. However, although this is true for finite systems as in the work of Walcholder et al. [2] on microarrays, this is not the case for hypotheses across an experimental research field.

To estimate *R* it is necessary to first estimate a denominator defined by the set of possible hypotheses in a field of enquiry (the numerator will consist of the true hypotheses within that set). But what is this exactly? Is it all conceivable hypotheses? This cannot be correct because such a set would be infinite. Perhaps then it is all hypotheses that are actually investigated? But how is this defined, by all hypotheses that are thought through and written down? Or is it restricted to all those hypotheses that are communicated to others? Or, to those that undergo some preliminary experimental investigation, such as the tinkering that goes on prior even to something as formal as a pilot experiment? Or, only hypotheses that are subject to a fully planned NHT experiment? One answer to this conundrum is that the set of possible hypotheses is restricted to those that have sufficient explanatory power and logical consistency with other findings in the field (plausibility) to warrant experimental testing. It is clear that *restricting the set of possible hypotheses to those that are plausible in the sense of fitting the constraints set by existing knowledge in a research field reduces the denominator for R and thus R increases.* If this were the end of the matter, then it would lead immediately to some interesting results: some experimental fields are much more tightly constrained in this sense than others. For example, theoretical physics is so highly developed that new hypotheses or theories are subject to very strong logical constraints prior to even being considered worthy of experimental testing. Einstein's theory of General Relativity was remarkable even before empirical testing because it passed a very high bar set by the logical constraints of existing findings in the field, so much so that it justified high cost experimental tests. By contrast, in a field with relatively weak theoretical and empirical constraints such as social psychology, simple hypotheses consisting of little more than proposed associations between phenomena may be admitted for experimental testing. Furthermore, these may be relatively easily tested by low cost experiments. Hence, *R* in advanced physics would appear to be far higher than in social psychology.

However, other considerations about the logic of hypotheses lead to an opposite conclusion and hence to a contradiction at worst, or at best to an unsolved paradox. A hypothesis that is consistent with other findings in a field and that explains or predicts these findings in

E-mail address: john.ashton@otago.ac.nz.

https://doi.org/10.1016/j.mehy.2018.02.004
Received 1 December 2017; Accepted 6 February 2018

addition to making some new (hence testable) predictions will be more precise or specific than a hypothesis that does not. Vague hypotheses with relatively indeterminate predictions (such as are the norm in horoscopes for example) are virtually untestable because they rule out few possible outcomes. The more informative a hypothesis, the less probable *a priori* it is. A tautology is not empirically informative at all, and has a probability of being true of 1. By contrast, a testable hypothesis will rule out a number of conceivable outcomes and so be less probably true. This means that *R* for the more rigorously constrained hypothesis will be *smaller* than for a hypothesis in a field with weaker constraints. To use the Einstein example again, the precision of the predictions made by his theory and the universality of its scope means that it rules out a great number of possible outcomes, which means that *R* becomes vanishingly small. In fact, several logicians have argued cogently that for any significant causal theory the prior probability of its being true approaches zero [3,4].

Therefore, it can be shown that hypotheses with greater explanatory power are both more and less probable than weaker hypotheses, and so the concept of pre-study odds is incoherent.

The model is falsified by data

Table 4 in [1] makes explicit predictions about false positive rates in various types of investigation. These take the form of PPV (positive predictive values; the false positive probability is the complement of this). The predictions for exploratory research are alarming, with false positive rates predicted to be well over 99% *even for relatively highly powered experiments with low bias*. Hence, irreproducible results are explained by low pre-study odds *irrespective of bias or experimental standards*.

Strangely, that this is not been confirmed by replication studies has received little or no comment. Successful replication rates have been reported that are far in excess of that predicted the Ioannidis model. In one highly cited study that sought to replicate 100 experimental or correlational studies [5] 36% of findings tested were reproduced, a far greater rate than predicted by the Ioannidis model, *and* in a field with relatively weak theoretical constraints (psychology). In a study that *specifically selected highly unexpected findings* to replicate [6], which according to Table 4 in Ioannidis [1] should be reproduced at a rate of 0.1–0.15%, the actual replication rate was 11%, *a hundred-fold difference*. Hence, Ioannidis's model and assumptions about pre-study odds have been falsified.

The model is pernicious

There are many reasons why experiments are not reproduced. For example, in recent decades many researchers were unaware of the lack of quality control for many antibody probes in commercial production. As a result, papers were published based upon antibodies that were subsequently challenged (e.g. [7]). This was (and is) a clear problem with clear solutions; data sharing on quality control for antibodies among scientists, and greater awareness of potential problems. It also highlights the risks that come with scientific financial bubbles [8], where rapid production of new technologies in highly funded fields may lead to inferior quality control. Publication bias, data manipulation and both low and high level scientific fraud are also definite problems with various solutions. The problem of lack of checking and rechecking of results (experimental rigour) and design problems (such as lack of blinding when outcome measures have an element of subjectivity) are also well known. However, all of these are quite distinct from the core idea in the Ioannidis model, which is the overwhelming influence of pre-study odds on experimental reproducibility. This is pernicious because it casts scepticism on experimental science as such, rather than simply on badly carried out science. This is damaging to science, and such claims for the invalidity of the experimental enterprise have now extended beyond science with potential consequences for the rationality of our society [9–11].

Making science more rigorous

The focus of the Ioannidis critique is null hypothesis testing. There are problems with NHT that have been widely discussed, such as widespread confusion over the difference between the likelihood of data given a (null) model with the probability that a hypothesis is true given a set of data. A deeper and less discussed problem with NHT lies in its algorithmic nature. It is possible to follow the forms of NHT quite properly, but only to test a trivial hypothesis. This is an example of what Richard Feynman called “Cargo cult science” [12]. Hence, in this way NHT can become an experimental box-ticking exercise and thus a smokescreen for poor standards in other ways, such as quality control tests of reagents or the theoretical consistency of the hypothesis [13]. Despite this, NHT has a very clear purpose in experimental research, and that is to test the alternative explanation that results are due to chance. NHT is not perfect for this, and other statistical approaches are hotly debated [14] but the problem that NHT addresses is very important and the need for NHT or something like it very real.

It is therefore extremely important to understand the limited role of NHT in science (testing for chance) and not to confuse it with the whole of experimental science. A highly testable theory is one that rules out a large number of possible experimental outcomes. As discussed above, any attempt to reduce this to pre-study odds runs into a paradox, but all this means is that the application of probabilities to causal theories is misguided [15]. Attempts at experimental testing are properly aimed at such theories, not only at simple hypotheses about the outcomes of particular experiments that such theories predict. Highly testable theories that over time pass tests and explain or solve research problems are thereby corroborated and considered to be provisionally true and scientifically significant. Ioannidis [1] acknowledges that negative results for “major concepts” are more informative than for narrow questions but does not recognise how this is related to testability as such. A greater acknowledgement of this, and of the limits on the role and scope of NHT in the scientific enterprise could do much to correct the identification of NHT with the whole of experimental science, the analysis of NHT in terms of the incorrect Ioannidis model, and the crisis of public trust in science [9–11].

Nevertheless, there is a relatively straightforward way of making even the results of individual NHTs more reproducible, and that is to set stringent standards on what are considered to be meaningful effect sizes (and not just significant P values). Proposals based on reducing the value of P that is considered significant [16] in the absence of any consideration of effect sizes run into a paradox discussed by Lindley [17] and Meehl [18], wherein increasing sample sizes can increase the chance of finding significant P values but for tiny and meaningless effect sizes. This proposal has been borne out by replication studies: The Open Science Collaboration [5] pointed out that in their results replication success was better predicted by the strength of original associations than by researcher status. Alternatively, for original studies meaningful effect sizes can be set in advance of experiment on theoretical and/or practical grounds [19].

Refutational strength as a guiding principle for experimental research

Powerful experiments are those that are not only well constructed on experimental design principles, but also test predictions that logically follow directly from the hypothesis being tested. When this is not the case loosely associated experiments can be performed without strongly testing the hypothesis. For example, a theory might predict that drugs targeting protein X can improve life expectancy in patients with disease Y. A well performed clinical trial with hard outcomes is a very direct test of this hypothesis. However, the predicted result for some outcome measure for a cellular or animal model is a much less