Rev Saúde Pública 2011;45(3)

Renan M V R Almeida

The role of plausibility in the evaluation of scientific research

ABSTRACT

The paper discusses the impact of plausibility (the *a priori* probability) on the results of scientific research, according to the approach proposed by Ioannidis, concerning the percentage of null hypotheses erroneously classified as "positive" (statistically significant). The question "what fraction of positive results are true-positives?", which is equivalent to the positive predictive value, is dependent on the combination of true and false hypotheses within a given area. For example, consider an area in which 90% of hypotheses are false and $\alpha = 0.05$ and power = 0.8: for every 1,000 hypotheses, 45 (900 x 0.05) are false-positives and 80 (100 x 0.8) are true-positives. Therefore, the probability of a positive result being a false-positive is 45/125. In addition, the reporting of negative results as if they were positive would contribute towards an increase in this fraction. Although this analysis is difficult to quantify, and these results are likely be overestimated, it has two implications: i) plausibility should be considered in the analysis of the ethical adequacy of a research proposal, and ii) mechanisms aimed at registering studies and protocols should be encouraged.

DESCRIPTORS: Hypothesis-Testing. Reproducibility of Results. Statistical Methods and Procedures.

INTRODUCTION

In the contemporary statistical methodology,[11] which is widely in use across all fields of science, the "null hypothesis" (H_0) represents the inexistence of a given effect (therefore its name). H_0 is either "rejected" or "not rejected" based on an appropriate test statistic (for example, Student's *t* for assessing the difference between two means). Thereafter, the analysis strategy consists of calculating a probability – known as the p-value – associated with this statistic. In cases in which this value is lower than a threshold defined *a priori* (α), the effect is considered to exist, or to be "statistically significant." Two types of error are intrinsic to this procedure, and are known as Type I (rejecting H_0 when it is true) and Type II (not rejecting H_0 when it is false). These errors occur with probabilities "α" and "β," respectively. In general, α is set arbitrarily to 5%, and experimental designs often aim at a level of up to 20% for β (that is, 80% probability of correctly rejecting H_0 when it is false, or the test's "power").

Among other things, a shortcoming of the traditional approach is that it does not consider the effect of plausibility when evaluating a hypothesis. Especially among statisticians of non-classical persuasion, there is the idea that a p-value may overestimate the evidence against a hypothesis, since the effect of plausibility is not evident in classical analyses (that is, $p = 0.001$ is considered as evidence for rejecting both a plausible and an implausible hypothesis).[3,4]

Thus, the present article discusses the impact of initial plausibility in the results of scientific research, based on the approach of Ioannidis.[6-9] This approach relates to the percentage of null hypothesis H_0 erroneously classified as "positive

Programa de Engenharia Biomédica. Coppe. Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ, Brasil

Correspondence:
Renan MVR Almeida
Programa de Engenharia Biomédica
Universidade Federal do Rio de Janeiro
Caixa Postal 68510
Cidade Universitária
21945-970 Rio de Janeiro, RJ, Brasil
E-mail: renan.m.v.r.almeida@gmail.com

Received: 7/14/2010
Accepted: 11/14/2010

Article available from: www.scielo.br/rsp

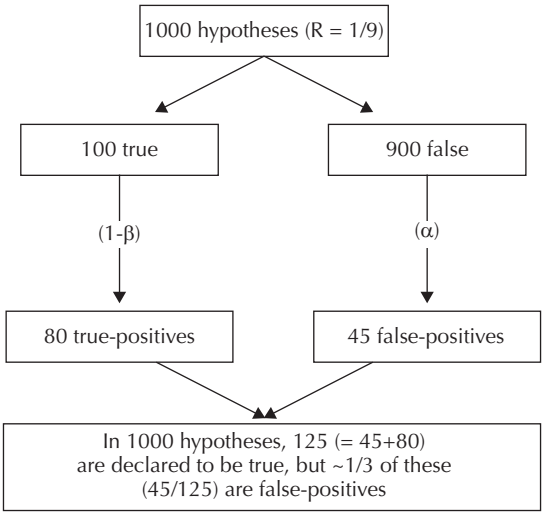
results" (statistically significant) in different fields of science. According to this, the question "what proportion of positive results is truly positive?" essentially depends on the proportion of true and false hypotheses tested within a given field of knowledge – or the *a priori* plausibility. This analysis is important for our understanding of the limitations inherent to scientific research, especially with respect to the *priors* (initial probabilities) of a given study.

THE IOANNIDIS APPROACH

In a recent series of articles, Ioannidis analyzed the role of replication and initial plausibility on the results of scientific research.[6-9] His central argument was presented in an article with the provocative title "Why most published research findings are false,"[6] in which the author states that "it can be proven that most claimed research findings in most areas of research are false." This article has been cited hundreds of times in the scientific literature.

Ioannidis systematized observations initially made by others, such as Browner & Newman[1] and Sterne & Smith.[15] Thus, the concepts of Type I and Type II errors were presented in a conceptually equivalent manner, such that the probability of a Type I error was defined as a the percentage of all H_0 hypotheses in a given field of research that are erroneously classified as statistically significant; and Type II error was defined as the percentage of false H_0 erroneously classified as non-statistically significant. Given a positive finding (that is, a rejected H_0), the probability that H_0 is indeed false is conditional upon the initial fraction of truly true and truly false hypotheses tested. This statement – which is analogous to the concept of positive predictive value, widely used in diagnostic testing[1,2] – can be understood by considering the following examples: i) all hypotheses tested in a given area are in actuality false. In this case, 100% of positive results would be false. ii) 100% of hypotheses tested are true. Analogously, all positive results would be true. And iii) in a field in which 90% of tested hypotheses are false (and maintaining the conventional values of $\alpha = 0.05$ and power = 0.8), for every 1,000 hypotheses, 45 will be false-positives (900 x 0.05), and 80 will be true-positives (100 x 0.8). Thus, given a positive result, the probability of it being a false positive is of roughly one-third (45/125) (Figure).

Between the extremes represented by cases i and ii, the relationship R = *true* H_0/*false* H_0 alters the equivalent to the positive predictive value for a given field of knowledge (given a positive result, the larger R is, the greater the probability of a true-positive). In other words, the lower a study's plausibility is, the greater the probability of a positive result being false. This phenomenon, according to Ioannidis, would help explain why even high-impact scientific publications often publish contradictory and non-replicable results.[6]



R = proportion of true hypotheses/false hypotheses; 1-β: statistical power; α: probability of a Type I error. Notes: This proportion is dependent on the initial proportion of true and false hypotheses under scrutiny. Figure adapted from Tabbarok A. Why most published research findings are false. [cited July 2010] Available from: http://marginalrevolution.com/marginalrevolution/2005/09/why_most_publis.html

Figure. Proportion of false hypotheses in relation to the total number of statistically significant results.

Ioannidis also introduced the concept of "*u* bias," defined as the probability of a negative result being erroneously reported as positive by selective use of secondary outcomes, alteration of cutoff points, use of inappropriate statistical methods, or fraud. Based on these concepts, a simulation of R and u for different types of study led to the conclusion that "in the described framework, a positive predictive value exceeding 50% is quite difficult to get"[6] (p. 699), and that "even well powered epidemiological studies may have only a one in five chance of being true, if R =1:10"[6] (p. 699). This justifies Ioannidis' claims regarding the prevalence of non-replicable results in science.

ANALYSIS AND EVALUATION

Ioannidis' analysis is contingent on two fundamental aspects: i) that the number of false hypotheses in any field is much greater than that of true hypotheses; and ii) that the u rate is indeed high (Ioannidis assumed values ranging from 10% to 80%). The former may be justified by the innovative nature of the scientific endeavor, as well as by the constant pressure for results, even in sterile or slow-moving fields, but is difficult to extrapolate to the majority of scientific fields. On the other hand, in the absence of so-called "file drawer effect,"[13,14] the influence of the phenomenon discussed

by Ioannidis is smaller, given that, if the pattern of R were known, the assessment of the true importance of a positive result would in principle be possible. However, as discussed above, the only way to determine R would be if all negative and positive results in a field were known (for example, if 100 hypotheses concerning a given phenomenon were tested and 95 were determined to be negative, global results would be compatible with a model that assumed the inexistence of this phenomenon). As to the u parameter, Goodman & Greenland[5,a] pointed out that: i) the definition of u is misleading, since it equates the selective reporting of secondary outcomes with direct fraud; and ii) the values for u assumed by Ioannidis (10%-80%) are speculative and dominated the simulation, that is, his conclusions are dependent on a high prevalence of "fraud." It follows that Ioannidis' assumptions are difficult to quantify, and that it is impossible to claim that their effect applies to a substantial fraction of scientific results.

According to the authors, Ioannidis' analysis failed to distinguish between different levels of evidence (in terms of p-value) against H_0.[5,8] Ioannidis dichotomized of results as either "statistically significant" or "non-statistically significant" based on the classically used α cutoff of 0.05. However, in practice, such dichotomization of results is unusual, the indication of specific p-values being generally preferred.

Irrespectively of the criticisms made by Goodman & Greenland[5,a] (who in fact agree with the central points of Ioannidis' analysis), the effect discussed is highly dependent on the specific characteristics of each field of research. Ioannidis suggests two fields as being critical: genomic research and the search for associations between nutrients and epidemiological outcomes, in which hypotheses are often tested using a heuristic approach and effects are small and difficult to measure. Another important example[b] is that of the field known as Complementary and Alternative Medicine (CAM), given that it is not difficult to conclude, despite a number of attempts,[10] that the only common, core principle among the countless trends that fall under the CAM denomination is the clear implausibility of their claims. This argument can be added to other points made in the literature (such as the lack of impact of negative results on this field, the inadequate legitimacy conferred to implausible ideas, and the inappropriate allocation of limited resources) to conclude that is conducting CAM studies in human subjects is unjustifiable.[12]

On the other hand, absence of an effect is impossible to prove, for there is always the possibility of the effect being below the threshold of detection. Furthermore, the amount of resources for research is infinitely smaller than what would be necessary to analyze all phenomena that can theoretically be proposed. Therefore, at least in the realm of human research, phenomena should only be investigated when they are both relevant and plausible.

FINAL CONSIDERATIONS

Problems inherent to the methods of contemporary science facilitate the improper publication of results that are apparently positive. These problems are related to the plausibility of studies in a given field, but are also linked to the so-called "file drawer effect." As discussed above, quantifying these effects is difficult, since they are also dependent on the particular conditions of a given research field.

However, two implications are worth highlighting: the first relates to the importance of the operational principle recognized by the Helsinki declaration, which in its 11th article states that "medical research involving human subjects must conform to generally accepted scientific principles, be based on a thorough knowledge of the scientific literature, other relevant sources of information, and adequate laboratory and, as appropriate, animal experimentation."[16] Lack of plausibility should thus be regarded as an important violation of research ethics. The second implication refers to the need to develop mechanisms for registering study protocols,[13] so as to minimize and facilitate the detection of both the file drawer effect and of protocol alterations. Such registration mechanisms would also help to identify duplicate studies and to expedite meta-analysis, thus contributing to greater transparency and efficiency in scientific research.

[a] Goodman S, Greenland S. Assessing the unreliability of the medical literature: a response to "Why most published research findings are false". Baltimore: Johns Hopkins University; 2007[citado 2010 jul]. (Working paper, 135). Disponível em: http://www.bepress.com/jhubiostat/paper135

[b] Novella S. Are most medical studies wrong? *Neurologica Blog*. 2007[citado 2010 jul]. Disponível em: http://theness.com/neurologicablog/?p=8

REFERENCES

1. Browner W, Newman TB. Are all significant p values created equal? The analogy between diagnostic tests and clinical research. *JAMA*. 1987;257(18):2459-63.
2. Dawson B, Trapp RG. Basic & Clinical Biostatistics. New York: McGraw-Hill; 2004.
3. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med*. 1999;130(12):995-1004.
4. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med*. 1999;130(12):1005-13.
5. Goodman S, Greenland S. Why most published research findings are false: problems in the analysis. *PLoS Med*. 2007;4(4):e168. DOI:10.1371/journal.pmed.0040168
6. Ioannidis JPA. Why most published research findings are false *PLoS Med*. 2005;2(8):e124. DOI:10.1371/journal.pmed.0020124
7. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research *JAMA*. 2005;294(2):218-28.
8. Ioannidis JPA. Why most published research findings are false: author's reply to Goodman and Greenland. *PLoS Med*. 2007;4(6):e215. DOI:10.1371/journal.pmed.0040215
9. Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology*. 2008;16(4) 640-8. DOI:10.1097/EDE.0b013e31818131e7
10. Manzini T, Martinez EZ, Carvalho ACD. Conhecimento, crença e uso de medicina alternativa e complementar por fonoaudiólogas. *Rev Bras Epidemiol*. 2008;11(2):304-14. DOI:10.1590/S1415-790X2008000200012
11. Moore DS. Estatística Básica e sua Prática. Rio de Janeiro: LTC Editora; 2005.
12. Renkens CNM. Some complementary and alternative therapies are too implausible to be investigated. *Focus Alternat Complement Ther.* 2003;8(3):307-8. Disponível em:
13. Yamey G. Scientists who do not publish trial results are "unethical". *BMJ*. 1999; 319(7215):939.
14. Young NS, Ioannidis JPA, Al-Ubaydli O. Why current publication practices may distort science. *PLoS Med*. 2008;5(10):e201. DOI:10.1371/journal.pmed.0050201
15. Sterne JAC, Smith GD. Sifting the evidence: what is wrong with significance tests? *BMJ*. 2001;322:226-31.
16. World Medical Association. Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects - 2008 version. Ferney-Voltaire; 2008[citado 2010 jul]. Disponível em: http://www.wma.net/en/30publications/10policies/b3/index.html

---

Study presented at the 8th Congresso Brasileiro de Bioética, held in Buzios, Brazil, en 2009.
The author declares no conflict of interests.