DERLEME REVIEW

DOI: 10.5336/biostatic.2016-53445

The Published Research Findings are Trustable?: Review

Yayınlanmış Araştırma Bulguları Güvenilir mi?

İsmet DOĞAN,[a]
Nurhan DOĞAN[a]

[a]Department of Biostatistics and Medical Informatics, Afyon Kocatepe University Faculty of Medicine, Afyonkarahisar

Geliş Tarihi/*Received:* 30.09.2016
Kabul Tarihi/*Accepted:* 19.12.2016

Yazışma Adresi/*Correspondence:*
Nurhan DOĞAN
Afyon Kocatepe University
Faculty of Medicine,
Department of Biostatistics and Medical Informatics, Afyonkarahisar, TURKEY/TÜRKİYE
idogan@aku.edu.tr

ABSTRACT The achievements of scientific research are amazing. However, there is increasing concern that most current published research findings are false. In a provocative article Ioannidis (2005b) argues that, in disciplines employing statistical tests of significance, professional journals report more wrong than true significant results. Currently, findings of many published research are false or exaggerated, and a large part of the resources allocated to research is wasted. Also, it has become apparent that an alarming number of published results cannot be reproduced by other people. However, this should not be surprising. It can proven that most claimed research findings are false. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that is true is the positive predictive value (PPV). The probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance. After calculating the PPV value of a scientific research, it is possible that an initial statistically significant finding will turn out to be a false positive, even for large, well designed, and well conducted studies. Broadly, in the literature there are many different methods other than PPV about this topic, like total error rate, negative predictive value, and false positive report probability. Consequently for statistical significance in the study, not only P values but also PPV values given would be appropriate in terms of the reliability of the results. The purpose of this article is to review and to evaluate the potential usefulness of PPV.

Keywords: Published research finding; positive predictive value; false positive report probability

ÖZET Bilimsel araştırma sonucu elde edilen başarılar hayret verici düzeyde olmasına rağmen yayınlanmış araştırma bulguları ile ilgili endişeler artmaktadır. Ioannidis (2005b) tarafından yazılan makalede, istatistik anlamlılık testlerini kullanan disiplinlere ait mesleki dergilerde yer alan yanlış anlamlı sonuçların doğru anlamlı sonuçlardan daha fazla olduğu ifade edilmektedir. Yayınlanan araştırma bulgularının birçoğu yanlış veya abartılı olup araştırmaya ayrılan kaynakların büyük bir kısmı israf edilmektedir. Ayrıca, yayınlanmış sonuçların önemli bir kısmının diğer araştırmacılar tarafından yapılan çalışmalarda elde edilemediği ortaya çıkmıştır. Ancak bu şaşırtıcı bir durum değildir. Araştırma bulgularının çoğunun yanlış olduğu kanıtlanabilir. İstatistik anlamlılık sonucu iddia edilen bir araştırma bulgusu elde edildikten sonra çalışmadan elde edilen sonucun doğru olma olasılığı pozitif öngörü değeridir. Bir araştırma bulgusunun gerçekten doğru olma olasılığı, söz konusu bulgunun önsel (çalışma yapılmadan önceki) doğru olma olasılığına, çalışmanın istatistiksel gücüne ve istatistik anlamlılık düzeyine bağlıdır. İyi tasarlanmış, yürütülmüş ve yeteri büyüklüğe sahip bir çalışmadan başlangıçta elde edilen istatistik anlamlı bir bulgunun pozitif öngörü değeri hesaplandıktan sonra, yanlış pozitif bir bulgu olduğunun belirlenmesi mümkündür. Bu konu hakkında genel olarak literatürde, pozitif öngörü değeri dışında toplam hata oranı, negatif öngörü değeri ve yanlış pozitif rapor olasılığı gibi birçok farklı yöntem vardır. Sonuç olarak, istatistik anlamlılık için çalışmalarda yalnız P değeri değil aynı zamanda PPV değerinin de verilmesi sonuçların güvenirliği bakımından uygun olacaktır. Bu çalışmanın amacı pozitif öngörü değerinin potansiyel yararlılığını gözden geçirmek ve değerlendirmektir.

Anahtar Kelimeler: Yayınlanmış araştırma bulgular; pozitif öngörü değeri; yanlış pozitif rapor olasılığı

Scientific progress depends on the slow, steady accumulation of data and facts about the way the world works. The scientific process is also hierarchical, with each new result predicated on the results that came before. When developing new experiments and theories, scientists rely on the accuracy of previous discoveries, as laid out in the published literature. The accuracy of published medical research is critical for scientists, physicians and patients who rely on these results. However, the fundamental belief in the medical literature was called into serious question by a paper suggesting that most published medical research is false.[1] This statement seems absurd on the first reading. Scientific research is carried out by highly trained and skilled scientists, vetted through peer review, and publicly scrutinized once it appears in journals. The entire scientific publishing infrastructure was originally conceived to prevent the publication of incorrect results and provide a forum for correcting false discoveries. It seems inconceivable that most of the findings that pass through this process are false.[2]

For any tested association, in a binary framework, the resulting inference could be categorized as a true negative, false positive, false negative, or true positive. The categorization can be applied to single studies as well as to collective results derived from many data sets. Although it may not be optimal to categorize results in a dichotomous fashion, such an approach is common in the field, and it allows for probabilistic estimations about how likely it is identify a true underlying association. There has been an ongoing concern in all disciplines regarding false-positive findings. However, erroneous inferences from any study include not only false positives, but also false negatives.[3]

The rate of findings that have later been found to be wrong or exaggerated has been found to be 30 percent for the top most widely cited randomized, controlled trials in the world's highest-quality medical journals. For non-randomized trials that number rises to an astonishing five out of six.[4] You make a fool of yourself if you declare that you have discovered something, when all you are observing is random chance. From this point of view, what matters is the probability that, when you find that a result is statistically significant, there is actually a real effect. If you find a significant result when there is nothing but chance at play, your result is a false positive, and the chance of getting a false positive is often alarmingly high.[5] There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims. However, this should not be surprising. It can be proven that most claimed research findings are false.[6]

The whole point of a diagnostic test is to use it to make a diagnosis, so we need to know the probability that the test will give the correct diagnosis. The sensitivity and specificity do not give us this information. Instead we must approach the data from the direction of the test results, using predictive values.[7]

Ioannidis (2005b), quantified the theoretical basis for lack of replication by deriving the positive predictive value (PPV) of the truth of a research finding on the basis of a combination of factors. He showed elegantly that most claimed research findings are false. One of his findings was that the more scientific teams involved in studying the subject, the less likely the research findings from individual studies are to be true. Ioannidis showed that the probability of a research finding being true when one or more studies find statistically significant results declines with increasing number of studies.[8]

The false positive report probability (FPRP) is the complement of the PPV which is the probability that, when you get a “significant” result there is actually a real effect. So, for example, if the FPRP is 70%, the PPV is 30%.[5] The FPRP is a more self-explanatory term so it will be preferred here. In classical theory,

the truth of H_0 and H_A is considered unknown, not random. Therefore, Wacholder et. al.,(2004) must go outside classical theory to consider H_0 and H_A probabilistically. They define the prior probability (P) as $P = Pr(H_A\ is\ true)$. Thus false positive report probability is,

$$FPRP = \frac{\alpha(1-P)}{[\alpha(1-P)+(1-\beta)P]}$$

The FPRP, the probability of no true association between two variables given a statistically significant finding, depends not only on the observed P value but also on both the prior probability that the association between two variable is real and the statistical power of the test. The FPRP approach offers guidelines for publication and interpretation of study results. It provides a way for editors and readers of articles to protect themselves from being misled by statistically significant findings that do not signify a true association. Table 1 presents the joint probabilities of statistical significance of a single test of association and truth of the alternative hypothesis (Table 1).[9]

After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that is true is the PPV. The PPV is derived easily. For a true alternative hypothesis, the probability that the out-come is significant (i.e. that the null hypothesis is correctly rejected) is $(1-\beta P$. For a true null hypothesis, the probability of a significant result is (i.e. that the null hypothesis is wrongly rejected) is $\alpha(1-P)$. Then, the PPV is:[6,10]

$$PPV = \frac{(1-\beta)P}{[(1-\beta)P+\alpha(1-P)]}$$

The prior probability of a hypothesis is a critical determinant of its probability after observing a study result, which the p-value does not reflect. This point has been made repeatedly by statisticians, epidemiologists and clinical researchers for at least 60 years, but is still underappreciated.[11]

The calculations for PPV values based on PPV formula are shown in Table 2.

Ioannidis (2005b) presents a Bayesian analysis of the problem which most people will find utterly confusing. The idea of Ioannidis is shown in Figure 1.[1] Suppose that the null hypothesis is true for 99% of the hypotheses being considered by scientific investigators. If scientists test 1000 hypotheses with a statistical power of 80% then $1000 * 1\% * 80\% = 8$ true alternative hypotheses should be correctly detected. Even though the Type I error rate is much lower, the prevalence of null hypotheses is much higher. With a Type I error rate of 5% then we expect $1000 * 99\% * 5\% = 49.5$ null hypotheses will be incorrectly detected. In this situation $1 - \frac{8}{8+49.5} = 86\%$ of rejected hypotheses will actually be null (Figure 1).[2]

One may deduce several interesting corollaries about the probability that a research finding is indeed true.[6]

TABLE 1: Joint probability of significance of test and truth of hypothesis.

	Significance of Test		
Truth of Alternative Hypothesis	**Significant**	**Not Significant**	**Total**
True associaton	$(1-\beta)P$ (True positive)	βP (False negative)	P
No association	$\alpha(1-P)$ (False positive)	$(1-\alpha)(1-P)$ (True negative)	$1-P$
Total	$(1-\beta)P+\alpha(1-P)$	$\beta P+(1-\alpha)(1-P)$	1

α: Type I error rate β: Type II error rate
P: Denote the a priori probability of a hypothesis being true.

TABLE 2: Positive predictive values.

	$\alpha = 0.10$			$\alpha = 0.05$			$\alpha = 0.01$			$\alpha = 0.001$		
P	$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.20$	$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.20$	$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.20$	$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.20$
0.001	0.0094	0.0089	0.0079	0.0187	0.0177	0.0158	0.0868	0.0826	0.0741	0.4874	0.4739	0.4447
0.01	0.0876	0.0833	0.0748	0.1610	0.1538	0.1391	0.4897	0.4762	0.4469	0.9056	0.9009	0.8899
0.02	0.1624	0.1552	0.1404	0.2794	0.2687	0.2462	0.6597	0.6475	0.6202	0.9510	0.9484	0.9423
0.03	0.2271	0.2177	0.1983	0.3701	0.3576	0.3310	0.7461	0.7357	0.7122	0.9671	0.9653	0.9612
0.04	0.2836	0.2727	0.2500	0.4419	0.4286	0.4000	0.7983	0.7895	0.7692	0.9754	0.9740	0.9709
0.05	0.3333	0.3214	0.2963	0.5000	0.4865	0.4571	0.8333	0.8257	0.8081	0.9804	0.9793	0.9768
0.06	0.3775	0.3649	0.3380	0.5481	0.5347	0.5053	0.8584	0.8517	0.8362	0.9838	0.9829	0.9808
0.07	0.4169	0.4038	0.3758	0.5885	0.5753	0.5463	0.8773	0.8714	0.8576	0.9862	0.9855	0.9837
0.08	0.4524	0.4390	0.4103	0.6230	0.6102	0.5818	0.8920	0.8867	0.8743	0.9880	0.9874	0.9858
0.09	0.4844	0.4709	0.4417	0.6527	0.6403	0.6128	0.9038	0.8990	0.8878	0.9895	0.9889	0.9875
0.1	0.5135	0.5000	0.4706	0.6786	0.6667	0.6400	0.9135	0.9091	0.8989	0.9906	0.9901	0.9889
0.2	0.7037	0.6923	0.6667	0.8261	0.8182	0.8000	0.9596	0.9574	0.9524	0.9958	0.9956	0.9950
0.3	0.8028	0.7941	0.7742	0.8906	0.8852	0.8727	0.9760	0.9747	0.9717	0.9975	0.9974	0.9971
0.4	0.8636	0.8571	0.8421	0.9268	0.9231	0.9143	0.9845	0.9836	0.9816	0.9984	0.9983	0.9981
0.5	0.9048	0.9000	0.8889	0.9500	0.9474	0.9412	0.9896	0.9890	0.9877	0.9989	0.9989	0.9988
0.6	0.9344	0.9310	0.9231	0.9661	0.9643	0.9600	0.9930	0.9926	0.9917	0.9993	0.9993	0.9992
0.7	0.9568	0.9545	0.9492	0.9779	0.9767	0.9739	0.9955	0.9953	0.9947	0.9995	0.9995	0.9995
0.8	0.9744	0.9730	0.9697	0.9870	0.9863	0.9846	0.9974	0.9972	0.9969	0.9997	0.9997	0.9997
0.9	0.9884	0.9878	0.9863	0.9942	0.9939	0.9931	0.9988	0.9988	0.9986	0.9999	0.9999	0.9999
1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Corollary 1. The smaller the studies conducted in a scientific field, the less likely the research findings are to be true.

Corollary 2. The smaller the effect sizes in a scientific field, the less likely the research findings are to be true.

Corollary 3. The greater the number and lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true.

Corollary 4. The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.

Corollary 5. The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true.

Corollary 6. The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.

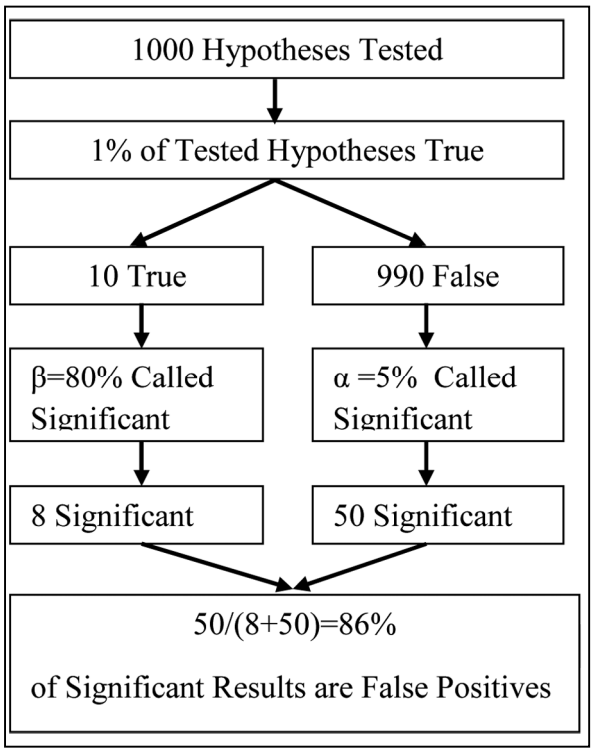

FIGURE 1: Tree diagram to illustrate the false discovery rate.

RESULTS

The seriousness of false-positives cannot be overemphasized such incorrect findings not only hinder any valid understanding of human nature but also can waste vast amounts of resources for those who believe

in false-positive findings. However, what is missing in the current debate is the explicit recognition of factors in conventional research practice in all disciplines that may go against the inflation of false-positive rates. Indeed, such factors sometimes make a substantial contribution to inhibiting the inflation of Type I error rates, making false-positive findings less likely.[12]

The false positive errors leading to the unwarranted publication of nonreplicable findings. Virtually all the critical arguments, and suggestions for improvement, that have been extracted from recent articles on "voodoo correlations", inappropriate statistical tests, questionable research practices, and replication are concerned with the problem of false positives.[13]

Ioannidis (2005b) suggested that most published medical research is actually false, calling into serious question the fundamental belief in the medical literature. The claim is based on the assumption that most hypotheses considered by researchers have a low prestudy probability of being successful. The suggested reasons for this low pre-study probability are small sample sizes, bias in hypothesis choice due to financial considerations, or bias due to over testing of hypotheses in “hot” fields. On the basis of this assumption, many more false hypotheses would be tested than true hypotheses. What can be done about these problems?[14]

1) In evaluating any study try to take into account the amount of background noise. That is, remember that the more hypotheses which are tested and the less selection which goes into choosing hypotheses the more likely it is that you are looking at noise.

2) Bigger samples are better. (But note that even big samples won't help to solve the problems of observational studies which is a whole other problem).

3) Small effects are to be distrusted.

4) Multiple sources and types of evidence are desirable.

5) Evaluate literatures not individual papers.

6) Trust empirical papers which test other people's theories more than empirical papers which test the author's theory.

7) As an editor or referee, don't reject papers that fail to reject the null.

Ioannidis (2005b) estimated that most published research findings are false, but he did not indicate when, if at all, potentially false research results may be considered as acceptable to society. The calculation of PPV tells us nothing about whether a particular research result is acceptable to researchers or not. Nevertheless, it can be shown that there is some probability (the “threshold probability”) which the results of a study will be sufficient for researchers to accept them as “true”.[15] False-positive results are an inherent feature of scientific research. They are a source of inconsistent and misleading evidence and have potential impact on approaches to prevent and cure diseases and on the allocation of research resources. The number of reports of clinical trials grows by hundreds every week. However, this does not mean that people making decisions about healthcare are finding it easier to obtain reliable knowledge for these decisions. Some of the information is unreliable. It’s not that you can’t believe anything that you read in the papers, just that you can’t believe everything.[16]

Conflict of Interest

Authors declared no conflict of interest or financial support.

Authorship Contributions

Opinion/Concept: Developing the hypothesis or idea of the research and/or the article: *İsmet Doğan;* ***Design: Designing a method to achieve the results:*** *İsmet Doğan;* ***Inspection/Consultancy: Organizing the conduct of the research, taking care of its progress and taking responsibility throughout the study:*** *İsmet Doğan, Nurhan Doğan;* ***Data Collection and/or Processing: Taking responsibility for the follow-up of patients, collection of relevant biological materials, regulation and reporting of data:*** *İsmet Doğan, Nurhan Doğan;* ***Analysis and/or Comment: Taking responsibility for evaluating the findings in a sensible way:*** *İsmet Doğan, Nurhan Doğan;* ***Literature Survey: Taking responsibility for doing a literature survey concerning the study:*** *İsmet Doğan, Nurhan Doğan;* ***Writing the Article: Taking responsibility for writing all or the most important parts of the work:*** *İsmet Doğan, Nurhan Doğan*

REFERENCES

1. Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics 2014;15(1):1-12.
2. Leek JT, Jager LR. Is most published research really false? bioRxiv 2016. bioRxiv 050575. Doi: 10.1101/050575.
3. Ioannidis JP, Tarone R, McLaughlin JK. The false-positive to false-negative ratio in epidemiologic studies. Epidemiology 2011;22(4):450-6.
4. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005;294(2):218-28.
5. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci 2014;1(3):140216.
6. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2(8):e124.
7. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ 1994;309(6947):102.
8. Moonesinghe R, Khoury MJ, Janssens AC. Most published research findings are false-but a little replication goes a long way. PLoS Med 2007;4(2):e28.
9. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 2004;96(6):434-42.
10. Diekmann A. Are most published research findings false? Journal of Economics and Statistics 2011;231(5/6):628-35.
11. Goodman S, Greenland S. Assessing the unreliability of the medical literature: a response to "why most published research findings are false". John Hopkins University, Department of Biostatistics, Working Paper 2007;135:1-25.
12. Murayama K, Pekrun R, Fiedler K. Research practices that can prevent an inflation of false-positive rates. Pers Soc Psychol Rev 2014;18(2):107-18.
13. Fiedler K, Kutzner F, Krueger JI. The Long Way From α-Error Control to Validity Proper: Problems with a Short-Sighted False-Positive Debate. Perspect Psychol Sci 2012;7(6):661-9.
14. Tabarrok A. Why most published research findings are false. Marginal Revolution. PLoS Med 2005;2(8):e124.
15. Djulbegovic B, Hozo I. When should potentially false research findings be considered acceptable? PLoS Med 2007;4(2):e26.
16. Clarke M. Can you believe what you read in the papers? Trials 2009;10:55.