PLoS One | 3(8): e2732, July 30, 2008
Published online 2008 July 30. doi: 10.1371/journal.pone.0002732

PLOS One | www.plosone.org

PMCID: PMC2518391

Why Most Published Research Findings Are False

John P. A. Ioannidis

Joint IPG, Stanford-Hoover Chair of Disease Prevention and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Stanford University School of Medicine, Stanford, California, United States of America, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece

Corresponding author. To cite this article: jioannidis@cmh.edu

Copyright © 2005 John P. A. Ioannidis

This article is part of the collection entitled Some of the Classics in Public Health Genomics Literature, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

See "Welcome Readers and Contributors" (PLoS Medicine, vol. 2(1): e1).
"Patent, Statistical, and Homogeneous Science" (PLoS Medicine, vol. 2(2): e123).
"Two Critical Interpretations of Statistical" (PLoS Medicine, vol. 2(3): e223).
"Statistical Significance" (PLoS Medicine, vol. 2(4): e 224).
"Scale, Complexity, and Contextuality" (PLoS Medicine, vol. 2(5): e 225).
"Research Hypotheses and Research Questions" (PLoS Medicine, vol. 2(6): e 226).
"Why Most Published Research Findings Are False" (PLoS Medicine, vol. 2(8): e 124).
"Why Current Statistical Practices Make Research Findings Inconsistent" (PLoS Medicine, vol. 2(9): e 225).

This article has been cited by other articles in PMC.

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias. The number of other studies on the same question, and, importantly, the status of the science in question. We propose a simple framework for such a scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical randomized trials and epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–12] that the high rate of nonreplication (lack of confirmation of research discoveries) is a consequence of the interaction of the complex phenomena that scientists deal with (biological, clinical, epidemiological, and social phenomena) and the hypothesis-testing stereotype for statistical inference that focuses on avoiding 1 false positive at all costs (P<0.05) and disregards false negatives. Ioannidis [13] has shown that most claimed research findings are false under these circumstances and for a wide range of research designs. Here I extend this work and suggest that multiple research attributes should be interpreted based only on p-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, associations, or causal relationships. "Negative" research also is important, but it is usually hampered by lack of resources, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

It can be proven that most claimed research findings are false.

As has been shown previously, 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 [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard of the true relationship (Table 1). In a research field where the true relationships that exist are to be sought among many candidates, the ratio of the number of "true relationships" to "no relationships" among those tested is crucial. Let R be the ratio of the number of true relationships to other relationships in the field where the relationships to be tested are chosen. In genetic epidemiology, it may be that one out of a few thousand polymorphisms in a candidate gene may be hypothesized or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (1 minus the Type II error rate). The probability of a study finding a true relationship is $(1 - \beta)R/(R + 1)$. The probability of a study finding any relationship is $(R + \alpha)/(R + 1)$. Assuming that a relationship is either true or not, the probability that it is false if a significant association is noted is $\Pr(\text{false} \mid \text{significance}) = (R + \alpha - (1 - \beta)R)/(R + \alpha)$, and PPV decreases with increasing α, since $\partial \Pr(\text{false} \mid \text{significance})/\partial \alpha > 0$ for most situations. Thus, with increasing bias, the chances that a research finding is true diminish considerably. This is shown for different levels of power and different ratios of true to false relationships in Figure 1, for fixed $\alpha = 0.05$. In a "hot" scientific field (high R), even a well-conducted study may have a PPV that is lower than that of a poorly conducted study in a "cold" field (low R).

Corollary 1: Most research findings are false for most research designs and for most fields. Let us suppose that in a field, 100 relationships are tested by teams of investigators, of which 10 are in reality true (R = 0.1). Let us suppose that the study has power of 0.8 to detect a true relationship ($1 - \beta = 0.8$). Let us also suppose that the study has 5% power to find an association with an odds ratio of 1.3 ($\alpha = 0.05$), and 99% chances to find a false association (1 minus the Type I error rate). Then it can be estimated that a statistically significant association is found with the p-value barely crossing the 0.05 threshold, the post-study probability that this is true increases about 12-fold compared with the pre-study probability, but it is still only 23 ($1 - \beta$)^n.

Now let us suppose that the investigators manipulate their designs, analyses, and reporting so as to make more relationships cross the $p = 0.05$ threshold even though this would not have been noticeable if a smaller bias were at play. This manipulation could be done, for example, with outcomes lacking a gold standard, with unvalidated measurements, with a "significant" difference between two groups being sought when the two groups are compared, with subgroup analyses being performed ad hoc or selectively, with endpoints being chosen among many possible ones, or with subsequent changes in the definitions of disease or of other study parameters. Let us suppose that these practices increase the probability of finding a true relationship from 0.8 to 0.9 but increase the probability of finding a false relationship from 0.05 to 0.1. Then, the post-study probability that a significant finding is true is now only 15. Thus, the "significance" that is attained has 1.5-fold lower meaning than that attained in the previous case. The "significance" is also more easy to achieve by a factor of 2 (0.1 versus 0.05). This is shown for different levels of power and for different ratios of true to false relationships in Figure 2, for a value of bias of 0.1 (10%). Finally, if we also assume that a proportion u of the relationships that are to be tested would be tested in any case (for instance, because they are the only ones that are hypothesized or they are the only ones that are feasible to be tested), then the post-study probability that a research finding is true is only $(R + u\beta R - u\alpha R + u(1 - \beta)R)/(R + \alpha + \beta R)$. It can be shown that this is even lower than the above values for any $u > 0$. Thus, the more a field is susceptible to bias, the lower the probability that its research findings are true. Even though bias increases the probability that a research finding is true when it is true, it decreases the probability that a research finding is true when it is false. Thus, the net effect is complex. It depends on the exact values of the various parameters, but the probability that a research finding is true is only 1.5×10^{-4} hardly any higher than the prior probability we had before any of the studies conducted in this scientific field. In this example, the "significant" research findings are almost all (99.7%) false.

Corollary 2: The smaller the studies conducted in a scientific field, the less likely the research findings are to be true. Small sample sizes means smaller power and, for all the reasons explained above, a lower probability that research findings are true. This is particularly important when the pre-study probability of the research finding is low, as in the example above. The same small sample sizes result also in wider confidence intervals, and more importantly, in lower probability that the research finding is not just a random fluctuation. In fact, as a research field moves toward smaller and smaller sample sizes, the probability that the research finding is true approaches zero. Small-sample studies may be easier to conduct than large-scale, well-powered studies. However, the ease of conducting research is a poor compass for the direction of research. Moreover, the ease of conducting small studies creates a vicious cycle, because it allows many more groups to be involved in a scientific field, which further decreases the probability that any given research finding is true. This is shown in the lower portion of Figure 1.

Corollary 3: The greater the number of relationships that are tested in a scientific field, the less likely the research findings are to be true. This is a corollary of Corollary 1. The more relationships are tested, the lower is the ratio of true to false relationships, which decreases the probability that a significant finding is true. Even a well-designed study loses some of its ability to confirm a true relationship when it is one among many that are tested. This is also shown in the lower portion of Figure 1.

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. Flexibility increases the possibility that research findings will satisfy preconceived notions and biases. For instance, in epidemiology, it may be possible to manipulate the way exposure is defined in a case-control study, or the way a control group is selected, in such a way as to yield the desired odds ratio. In randomized clinical trials, there may be a wide spectrum of permissible ways to define the eligible patient population, the intervention, the comparator treatment, the outcomes, and the statistical analysis. Such flexibility allows for the production of statistically significant results even when the research question is not quite right, or even when the research question is quite wrong. For example, if the wrong comparator is chosen, or if the dosing of the active drug is too low or too high, many randomized trials may yield results that are statistically significant in favor of the experimental treatment, but these results will be misleading. Nevertheless, such results may be easier to publish than results from well-conducted studies that happen to ask the wrong questions or to use the wrong comparators.

Corollary 5: The greater the interest and prejudice in a scientific field, the less likely the research findings are to be true. Conflicts of interest and prejudice may increase the probability that research findings are false. If a lot of money is at stake, small biases and conflicts of interest may shift the balance of probabilities to the side of falsehood. For instance, financial interests may favor the production of spurious findings in favor of the sponsor's product [14–21]. Prejudices may not necessarily have financial roots. Scientists in a given field may be predisposed toward the acceptance of their belief in a scientific hypothesis or theory. Such a belief may be based on previous research findings, but it may also be based on personal or political preferences or on other factors. For instance, a theory or a hypothesis may be favored because it has been proposed by a Nobel Prize winner or by the head of a powerful laboratory. Such preferences may lead to the publication of research findings that are biased in favor of the theory or hypothesis. Prestigious investigators may suppress via the peer review process the appearance and dissemination of findings that refute their findings, thus condemning their field to perpetuate their dogma. Empirical evidence suggests that such suppression is a common scientific sin.

Corollary 6: The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true. This seemingly paradoxical corollary follows because, as a hot area emerges, the pre-test probability of at least one research finding being true increases. But the more teams are involved, the more likely it is that at least one team will find a false relationship, because of the combination of bias and of the large number of tests being performed. Competition may also lead to shortcuts in replication of studies by other teams, leading to the publication of more false findings than would be the case in a less competitive field. In such circumstances, the "winner" of the race may be the team that is most willing to cut methodological corners to be the first to publish. The "winner's curse" is that the first claim is more likely to be false than subsequent claims that validate the initial claim. Competition may also lead to publication of results that are not the most accurate but those that are the most spectacular. This may be the---

Table 1
Research Findings and True Relationships

	True Relationship	
Research Finding	Yes	No
Yes	$(1 - \beta)R/(R + 1)$	$\alpha/(R + 1)$
No	$\beta R/(R + 1)$	$(1 - \alpha)/(R + 1)$

What is less well appreciated is that bias and the extent of repeated independent testing by different teams of investigators around the globe may further distort this picture and may lead to even smaller probabilities of the research findings being indeed true. We will try to model these two factors in the context of similar 2 × 2 tables.

Bias

First, let us define bias as the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced. Let u be the proportion of probed analyses that would not have been "research findings," but nevertheless end up presented and reported as such, because of bias. Bias should be confined within classifiers that cross some findings to be false by chance even though the study design, data, analysis, and presentation are perfect. Bias can consist manipulation in the analysis or reporting of findings. Selective or distorted reporting is a typical form of such bias. We may imagine that, in the 2 × 2 table, a fraction u of formerly "research findings" (both true and false) are now reported as being "research findings" (Table 2). If we now repeat the calculations that led to the formulas of Table 1, we get: $\Pr(\text{true} \mid \text{research finding}) = (R + u\beta R)/(R - \alpha R + u\beta R + \alpha + u - u\alpha)$, and PPV decreases with increasing u, since $\partial \Pr(\text{true} \mid \text{research finding})/\partial u < 0$ for most situations. Thus, with increasing bias, the chances that a research finding is true diminish considerably. This is shown for different levels of power and different ratios of true to false relationships in Figure 3, for fixed $\alpha = 0.05$. In a "hot" scientific field (high R), even a well-run study may have a PPV that is lower than that of a poorly run study in a "cold" field (low R).

Corollary 7: The more biases of various kinds in a scientific field, the less likely research findings are to be true. Small sample size means smaller power and, for all the reasons explained above, a lower probability that research findings are true. This is particularly important when the pre-study probability of the research finding is low, as in the example above. The same small sample sizes result also in wider confidence intervals, and more importantly, in lower probability that the research finding is not just a random fluctuation. In fact, as a research field moves toward smaller and smaller sample sizes, the probability that the research finding is true approaches zero. Small-sample studies may be easier to conduct than large-scale, well-powered studies. However, the ease of conducting research is a poor compass for the direction of research. Moreover, the ease of conducting small studies creates a vicious cycle, because it allows many more groups to be involved in a scientific field, which further decreases the probability that any given research finding is true. This is shown in the lower portion of Figure 1.

Corollary 8: The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely research findings are to be true. Flexibility increases the possibility that research findings will satisfy preconceived notions and biases. For instance, in epidemiology, it may be possible to manipulate the way exposure is defined in a case-control study, or the way a control group is selected, in such a way as to yield the desired odds ratio. In randomized clinical trials, there may be a wide spectrum of permissible ways to define the eligible patient population, the intervention, the comparator treatment, the outcomes, and the statistical analysis. Such flexibility allows for the production of statistically significant results even when the research question is not quite right, or even when the research question is quite wrong. For example, if the wrong comparator is chosen, or if the dosing of the active drug is too low or too high, many randomized trials may yield results that are statistically significant in favor of the experimental treatment, but these results will be misleading. Nevertheless, such results may be easier to publish than results from well-conducted studies that happen to ask the wrong questions or to use the wrong comparators.

Corollary 9: The greater the interest and prejudice in a scientific field, the less likely research findings are to be true. Conflicts of interest and prejudice may increase the probability that research findings are false. If a lot of money is at stake, small biases and conflicts of interest may shift the balance of probabilities to the side of falsehood. For instance, financial interests may favor the production of spurious findings in favor of the sponsor's product [14–21]. Prejudices may not necessarily have financial roots. Scientists in a given field may be predisposed toward the acceptance of their belief in a scientific hypothesis or theory. Such a belief may be based on previous research findings, but it may also be based on personal or political preferences or on other factors. For instance, a theory or a hypothesis may be favored because it has been proposed by a Nobel Prize winner or by the head of a powerful laboratory. Such preferences may lead to the publication of research findings that are biased in favor of the theory or hypothesis. Prestigious investigators may suppress via the peer review process the appearance and dissemination of findings that refute their findings, thus condemning their field to perpetuate their dogma. Empirical evidence suggests that such suppression is a common scientific sin.

Corollary 10: The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true. This seemingly paradoxical corollary follows because, as a hot area emerges, the pre-test probability of at least one research finding being true increases. But the more teams are involved, the more likely it is that at least one team will find a false relationship, because of the combination of bias and of the large number of tests being performed. Competition may also lead to shortcuts in replication of studies by other teams, leading to the publication of more false findings than would be the case in a less competitive field. In such circumstances, the "winner" of the race may be the team that is most willing to cut methodological corners to be the first to publish. The "winner's curse" is that the first claim is more likely to be false than subsequent claims that validate the initial claim. Competition may also lead to publication of results that are not the most accurate but those that are the most spectacular. This may be the---

Table 2
PPV of Research Findings for Various Combinations of Power (1 - β), Ratio of True to Not True Relationships (R), and Bias (u)

Claimed Research Findings May Often Be Simply Accurate Measures of the Prevailing Bias

As shown, the majority of modern biomedical research is operating in areas with very low pre- and post-study probability for true findings. Let us suppose that in a scientific field, no bias at all is operating. That means that what is claimed as a research finding is indeed the only one that is being tested in that field. This is unlikely to be true even approximately, or even in the vast majority of cases, but it is the ideal case. In this ideal case, the post-study probability that a finding is true is only equal to the prior probability. It is still quite low in the fields of interest. The extent that observed findings deviate from what is expected by chance also would be simply a mere measure of the prevailing bias.

For example, let us suppose that no mutations or dietary patterns are actually important determinants for the risk of developing a specific tumor. Let us also suppose that the scientific literature has examined 100 mutations and dietary patterns and it has shown that 10 of them are associated with cancer with relative risks in the range of 1.2 to 1.4 for the comparison of the tested to the control or baseline category. The claimed effect sizes are simply measuring nothing but the net bias that has been generated in this scientific field. The more studies have been conducted in this scientific field, the stronger the bias, on average, that has been generated. Ioannidis [22] has shown that the most extreme, most "significant" results are those that tend to be the most biased. In this example, the most "significant" relationships will be those with the highest observed ratios. These are the ratios that have undergone the most amplification by the most biased studies, and they are likely to be the most inflated ones. The ratios that are the most "significant" in this example are those that are the most in need of being tested for their true magnitude of association. Ioannidis [22] has also shown that the most "significant" relationships are those that are the most susceptible to being the worst ones.

For fields with very low PPV, the few true relationships would not distort this overall picture much. Even if a few relationships are true, the shape of the distribution of the observed effects would still yield a clear measure of the bias involved in the field. This concept vaguely reverses the way we view scientific results. Traditionally, investigators have tried to find a statistical test that would "prove" a relationship. Nowadays, we could use the same tools to see how badly our fields are biased. Instead of searching for statistical significance, we should be concerned with quantifying the bias. Instead of being content with the few relationships that are considered relatively established, to see how often they are indeed confirmed. I suspect several established "classics" will fall by the wayside.

Nevertheless, most new discoveries will continue to stem from hypothesis-generating research with low or very low pre-study odds. We should then acknowledge that most of them will eventually prove to be wrong when tested and quantified. Hypothesis-generating research is fertile ground for subsequent testing. Its yields should be viewed as the richest field in large. Empirical, large statistical datasets are a treasure trove for hypothesis-generating research. It is impossible to decipher how much data dredging by the reporting authors or other research teams has preceded a reported research finding. Even if determining this were feasible, this would not inform us about the pre-study odds. There is no consistent or even roughly consistent mathematical connection between the extent of data dredging and the pre-study odds. It may be easier for a hypothesis proved across the relevant research fields and research designs. The wider field may yield some guidance for estimating this probability for the tested research. Meta-analyses and large consortia may sometimes be able to do that, if they are not biased themselves. Bias across the field can cause even such attempts to be misleading.

Abbreviation

PPV positive predictive value

Footnotes

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124

References

1. Ioannidis JP, Haidich AB, Lau J. Any casualties in the clash of randomised and observational evidence? BMJ. 2001;322:879-880 [PMC free article] [PubMed]
2. Lawlor DA, Davey Smith S, Kundu D, Bruckdorfer KR, Ebrahim S. Those confounded vitamins: What can we learn from the differences between observational and randomized trials? Lancet. 2004;363:1728-1731 [PubMed]
3. Vandenbroucke JP. When are observational studies as credible as randomized trials? Lancet. 2004;363:1728-1731. [PubMed]
4. Michiels S, Koscielny S, Hill C. Prediction of cancer outcomes with microarrays: A multiple random validation strategy. Lancet. 2005;365:488-492. [PubMed]
5. Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genet. 2001;29:306-309. [PubMed]
6. Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. Lancet. 2003;361:865-872. [PubMed]
7. Ioannidis JP. Genetic associations: False or true? Trends Mol Med. 2003;9:135-138. [PubMed]
8. Hirschhorn JN, Lohmueler K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med. 2002;4:45-61. [PubMed]
9. Ioannidis JL, Dixon Moxley E, Herbison P, Appleton N, Burnett W, Smith GD, Conway DI. Genetic prognostic and association studies in relation to disease progression and bias. Lancet. 2004;363:1773-1777. [PubMed]
10. Wacholder S, Chanock S, Garcia-Closas M, Elghossein L, Rothman N. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. J Natl Cancer Inst. 2004;96:434-442. [PubMed]
11. Risch NJ. Searching for genetic determinants in the new millennium. Nature. 2000;405:847-856. [PubMed]
12. Lohmueler KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet. 2003;33:177-182. [PubMed]
13. Tsilidis KS, Ioannidis JPA. Meta-analysis for molecular epidemiology research: Assessing the validity of the literature. Proceedings of the 3rd International Conference of Molecular Tumor Markers Epidemiology; 2005 Feb 2-4; Athens, Greece. Athens: Hellenic Society for Biochemistry and Molecular Biology; 2005. p 17.
14. Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG. Genetic associations in large genome-wide association studies. N Engl J Med. 2004;351:233-241. [PMC free article] [PubMed]
15. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genet. 2001;29:306-309. [PubMed]
16. Lohmueler KE, Mindrinos M, Richards JB. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet. 2003;33:183-187. [PubMed]
17. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Tijssen JH, Rutgers EJ. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-1066. [PubMed]
18. Chalmers I. Underreporting research is scientific misconduct. JAMA. 1990;263:1405-1408. [PubMed]
19. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. JAMA. 2004;291:2457-2465. [PubMed]
20. Tang JL, Liu JLY, Neaton JD, Wang W, Blackburn G, Thomas D, Grimshaw J. Interrater reliability of functional outcome assessment instruments. J Clin Epidemiol. 1996;49:1375-1379. [PubMed]
21. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999;282:1054-1060. [PubMed]
22. Ioannidis JP, Dixon DO, Ware JH, Lau J. Improving the quality of reports of meta-analyses of randomized controlled trials: The QUOROM statements. Quality of Reporting of Meta-analyses. J Natl Cancer Inst. 2004;96:1448-1453. [PubMed]
23. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: A proposal for reporting. Meta-analysis of Observational Studies in Epidemiology (MOOSE) group. JAMA. 2000;283:2008-2012. [PubMed]
24. Marshall M, Lockwood A, Bradley C, Adams C, Joy C, et al. Unpublished rating scales: A major source of bias in randomised controlled trials of treatments for schizophrenia. Br J Psychiatry. 2000;176:249-252. [PubMed]
25. Altman DG, Goodman SN. Transfer of technology from statistical journals to the biomedical literature. Past trends and future predictions. JAMA. 1994;272:129-132. [PubMed]
26. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. JAMA. 2004;291:2457-2465. [PubMed]
27. Ioannidis JPA, Lau J. Evolution of treatment effects over time: Empirical insight from recent meta-analyses. Proc Natl Acad Sci USA. 2001;98:771-775. [PMC free article] [PubMed]
28. Papanikolaou GN, Baltogianni MS, Contopoulos-Ioannidis DG, Haidich AB, Giannakakis IA, et al. Reporting of conflicts of interest in guidelines of preventive and therapeutic interventions. BMC Med Res Methodol. 2001;1:3. [PMC free article] [PubMed]
29. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA. 1992;268:240-248. [PubMed]
30. Ioannidis JP, Trikalinos TA. Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. J Clin Epidemiol. 2005;58:543-549. [PubMed]
31. Lubin JH, Gail MH. On power and sample size of studies challenging an established theory. Am Stat Assoc Stat Med. 1984;3:35-43. [PubMed]
32. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer. 2004;4:309-314. [PubMed]
33. Lindley DV. A statistical paradox. Biometrika. 1957;44:187-192.
34. Bartlett MS. A comment on D.V. Lindley's statistical paradox. Biometrika. 1957;44:533-534.
35. Berry DA. Decision analysis and Bayesianism in clinical trials. JAMA. 1993;269:1772-1776.
36. Ioannidis JP, Ontiveros M, Iwasiw C, Jiang J, Avey L, et al. Clinical trial acceptance. A measure for the harmonization of Committee of Medical Journal Editors. N Engl J Med. 2004;351:1250-1251. [PubMed]
37. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. JAMA. 2005;294:218-228. [PubMed]
38. Hauth JRM, Chien H, Kodali RK. Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. J Biopharm Stat. 2003;13:675-689. [PubMed]

Articles from PLoS Medicine are provided here courtesy of Public Library of Science