The False-positive to False-negative Ratio in Epidemiologic Studies

John P. A. Ioannidis,[a] *Robert Tarone,*[b] *and Joseph K. McLaughlin*[b]

Abstract: The ratio of false-positive to false-negative findings (FP:FN ratio) is an informative metric that warrants further evaluation. The FP:FN ratio varies greatly across different epidemiologic areas. In genetic epidemiology, it has varied from very high values (possibly even >100:1) for associations reported in candidate-gene studies to very low values (1:100 or lower) for associations with genome-wide significance. The substantial reduction over time in the FP:FN ratio in human genome epidemiology has corresponded to the routine adoption of stringent inferential criteria and comprehensive, agnostic reporting of all analyses. Most traditional fields of epidemiologic research more closely follow the practices of past candidate gene epidemiology, and thus have high FP:FN ratios. Further, FP and FN results do not necessarily entail the same consequences, and their relative importance may vary in different settings. This ultimately has implications for what is the acceptable FP:FN ratio and for how the results of published epidemiologic studies should be presented and interpreted.

(*Epidemiology* 2011;22: 450–456)

There has been an ongoing concern in epidemiology regarding false-positive findings.[1,2] However, erroneous inferences from any study include not only false positives, but also false negatives.[3] In this study, the ratio of false positives to false negatives (FP:FN ratio) has been explored. On the basis of theoretical and empirical evidence, we argue that in most traditional areas of epidemiologic investigation this ratio is much higher than 1, whereas under certain circumstances the ratio can decrease sharply and become substantially smaller than 1. We provide evidence from human genome epidemiology, where the FP:FN ratio has decreased dramatically over time. We also probe whether similar advances are feasible for other areas of epidemiologic investigation. Finally, we discuss the interpretation and implications of FP:FN ratios.

Submitted 19 October 2010; accepted 4 January 2011; posted 13 April 2011.
From the [a]Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, CA; and [b]International Epidemiology Institute, Rockville, MD.
Editors' note: *Related articles appear on pages 457, 460, 464, and 467.*
Correspondence: John P. A. Ioannidis, Stanford Prevention Research Center, Stanford University School of Medicine, MSOB X306, 251 Campus Dr, Stanford, CA 94305. E-mail: jioannid@stanford.edu.

ISSN: 1044-3983/11/2204-0450
DOI: 10.1097/EDE.0b013e31821b506e

CONCEPTUALIZING THE FP:FN RATIO

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 (meta-analyses). 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 to identify a true underlying association. The Figure summarizes this framework, where N_1 denotes the number of hypotheses tested corresponding to actual underlying relationships and N_2 denotes the number of hypotheses tested corresponding to null relationships (ie, no actual relationship between exposure and disease risk exists). The ratio of N_1/N_2 is the prestudy odds, R.[2] R may vary substantially across different types of epidemiologic studies, depending on the maturity of the particular area of interest and on whether investigators select hypotheses to test based on extensive, limited, or no prior evidence. Using this framework, as previously proposed,[2] it can be shown that the FP:FN ratio is a function of R, the type I error rate α (ie, the nominal significance level), and the type II error rate β (ie, the complement of power):

$$\text{FP:FN} = \alpha/(\beta\text{R}) \qquad (1)$$

The aforementioned formula holds true in the absence of any biases. However, we want to also consider the proportion of probed analyses that would not have been nominally significant, but nevertheless ends up reported as such for any reason other than chance (ie, due to selective analysis/outcome reporting, confounding, or any combination of hundreds of possible biases)[2]—in other words, the effect of bias (denoted by u). The FP:FN ratio can then be expressed by the equation:

$$\text{FP:FN} = [\alpha\,(1 - \text{u}) + \text{u}]/(1 - \text{u})\beta\text{R} \qquad (2)$$

A similar amendment can be made for estimating the FP:FN ratio when we are interested in at least one of several

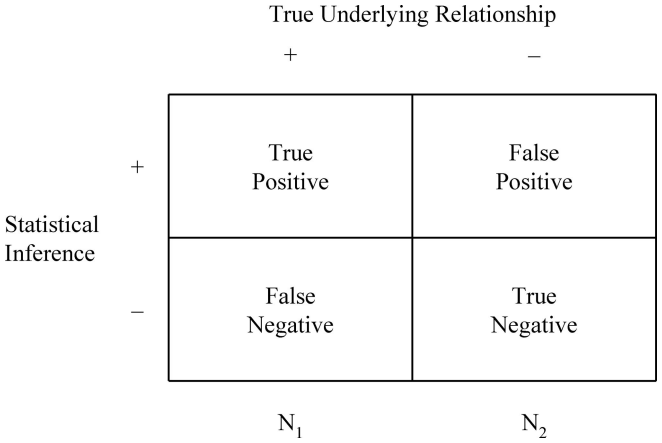

FIGURE. The 4 possible outcomes of dichotomous statistical inferences about hypothesized relationships between exposures and disease status investigated in epidemiologic studies; N_1 denotes the number of true underlying relationships investigated and N_2 denotes the number of null relationships investigated.

studies on the same topic showing false-positive results (see Ioannidis[2] for details). The relevant values of α, β, u, and R may vary considerably by research area. Even within the same research area, the ratio is expected to decrease when more stringent (lower) levels of type I error are required, and when bias is reduced or, ideally, eliminated. Conversely, larger studies with higher power produce an increased FP:FN ratio. Finally, the FP:FN ratio is inversely proportional to R, so that targeting a sample of hypotheses that are more enriched in true effects decreases the FP:FN ratio.

It would also be useful to contrast the FP:FN ratio with 2 other metrics, the ratio of false positives to true positives (FP:TP ratio), which is the inverse of the poststudy odds that a statistically significant association is true,[2] and the Bayes factor, which expresses the magnitude of the difference between the poststudy and prestudy odds.[4] The FP:TP ratio also depends on α, β, u, and R, but does not factor in the problem of false negatives. In the absence of bias, this ratio is given by $\alpha/[(1 - \beta)R]$. Therefore, the FP:TP ratio equals the FP:FN ratio when power is 50%, whereas in general the FP:FN ratio is equal to $[(1 - \beta)/\beta] \times$ (FP:TP). The Bayes factor does not depend on R and, in the absence of bias, it is given by $\alpha\beta/[(1 - \alpha)(1 - \beta)]$. Although the Bayes factor becomes more favorable (smaller) with reduced type I and type II error, the FP:FN ratio becomes smaller with reduced type I error and with increased type II error.

EVOLUTION OF THE FP:FN RATIO IN HUMAN GENOME EPIDEMIOLOGY

Human genome epidemiology offers a useful paradigm, because there has been a major transformation in this field with prominent changes in the key parameters that influence the FP:FN ratio. Until 5 years ago, investigators followed the path still followed in most traditional epidemiologic areas: a few candidate risk factors were selected based on diverse considerations (often under the umbrella of "biologic plausibility") that suggested their possible importance; studies were often of small sample size (underpowered, based on the current hindsight about the size of genetic effects); discovery hunting was performed using conventional levels of nominal statistical significance (typically 5% type I error rate, despite more prudent suggestions for alternatives, eg, false-report-rate probability[5]); confounding (population stratification) was often unaccounted for; and the field was subject to substantial selective reporting of "positive findings" that—not surprisingly—were often rapidly refuted by subsequent studies.[6] Many investigators suspected early on that the large majority of "discovered" candidate-gene associations were false,[2,5,7–9] despite the coexistence of more optimistic views, especially when data were enhanced by meta-analyses.[10]

Over time, human genome epidemiology has benefited from the routine adoption of replication practices.[11] Moreover, it has become easy to measure a large number of genetic risk factors concurrently, and genome-wide association (GWA) studies are currently the norm for discovery and replication of genetic associations.[12–14] As a collateral benefit from such enhanced measurement ability, we now have systematic replication efforts for inclusive, large sets of previously proposed nominally statistically significant candidate-gene associations. This allows one to estimate, with substantial confidence and without the threat of selective reporting of one-association-at-a-time, the ratio of false positives to true positives across proposed "discovered" associations of the candidate-gene era.

Table 1 catalogues several such empirical evaluations in studies published in 2007–2010 and examining at least 50 candidate genes each, in data sets with sample sizes of >1000 participants.[15–22] With one exception,[17] where the investigators genotyped only the gene variants that had been previously associated with a phenotype of interest, these investigations have used existing data from GWA studies to replicate previously proposed candidate associations. As shown (Table 1), robust evidence of replication of these previously proposed associations is infrequent. After adjusting for the multiplicity of comparisons in each study, only 13 gene loci-phenotype associations survive replication among the 1151 tested cumulatively in these studies. The crude replication rate is thus approximately 1.2%. Conceivably, the number of genuine associations is a bit larger than this disappointingly low percentage. For modest genetic effects, most of these replication exercises would have had modest power to detect significant associations at a type I error adjusted for multiplicity of comparisons. Consistent with this possibility, the largest study in Table 1[19] repli-

TABLE 1. Large-scale Efforts to Massively Replicate Reported Candidate-gene Associations[a]

First Author	Disease/Phenotype	Gene Loci Tested	Sample Size (Design)	Replicated Gene Loci[b]
Bosker et al[15]	Major depressive disorder	57	3540 (case-control)	1
Caporaso et al[16]	Smoking (7 phenotypes)	359	4611 (cohort[c])	1
Morgan et al[17]	Acute coronary syndrome	70	1461 (case-control)	0
Richards et al[18]	Osteoporosis (2 phenotypes)	150	19,195 (cohort[d])	3[e]:9[f]
Samani et al[19]	Coronary artery disease	55	4864; 2519 (case-control)	1[g]
Scuteri et al[20]	Obesity (3 phenotypes)	74	6148 (cohort)	0
Sõber et al[21]	Blood pressure	149	1644; 8023 (cohort[h])	0
Wu et al[22]	Childhood asthma	237	1476 (triads[i])	1

[a]The listed studies have been identified through a PubMed search using the strategy (replication [ti] or collaborative [ti] or genome-wide [ti]) and gene* [ti] and (candidate or "previously reported" or "previously proposed") for studies published between 2007 and 2010 (last search 29 July 2010) and complemented with eligible studies from those analyzed in the paper by Siontis et al.[24]
[b]With proper, stringent control for multiple comparisons.
[c]Two cohorts with combined analysis.
[d]Five cohorts with combined analysis.
[e]Femoral bone density.
[f]Spinal bone density; the 3 gene loci associated with femoral bone density were also associated with femoral bone density.
[g]2 case-control studies analyzed separately; the gene locus significantly associated in the smaller study also had the strongest association in the larger study.
[h]Initial discovery cohort of size 1644, with subsequent replication in 2 cohorts of size 1830 and 1823 and one case-control study with 2401 cases and 1969 controls.
[i]Case-parent triad design, with 492 triads consisting of an asthmatic child and both parents.

cated 3 of the 150 proposed candidate-gene loci for association with femoral bone-mineral density and 9 of the 150 for association with spine bone-mineral density, ie, a 2%–6% replication rate.

The data in Table 1 are consistent with an estimate of at least 20 false-positive findings for every one true-positive result, a ratio previously suggested for candidate-gene studies.[7] This is also consistent with most synopses of candidate-gene studies where grading of the evidence reveals very few variants with strong credibility.[23] Perhaps for some phenotypes other than those listed in Table 1, the replication percentage is higher. For example, in pharmacogenetics, the mechanism of action of a drug is usually well known, so one can target specifically the genes that code for the proteins in drug-related pathways.[24] However, for most disease phenotypes, a FP:TP ratio of 10 would likely be an optimistic estimate. Even under this optimistic scenario, approximately 1000 early gene loci-phenotype associations for the conditions listed in Table 1 were false positives from the candidate-gene approach.

How about false negatives in candidate-gene studies? On the basis of inclusive catalogue of GWA studies,[25,26] investigators have currently discovered >1000 associations with genome-wide significance ($P < 5 \times 10^{-8}$) for diverse phenotypes and whose genuineness is beyond debate. To our knowledge, none of the GWA-documented associations had been tested in the candidate-gene era with consistently "negative" results. There are no documented false-negative results arising from candidate-gene studies. Therefore, for the phenotypes listed in Table 1, the numerator of the FP:FN ratio is over 1000, while the denominator is apparently 0.

In the absence of substantial empirical data on false negatives in candidate-gene studies, on the basis of formulae presented earlier, we can develop estimates of the FP:FN ratio in such studies for various values of study power and for an illustrative 20:1 FP:TP ratio. For a nominal significance level of 5% and power of 90% (as discussed earlier in the text), the 20:1 FP:TP ratio is consistent with an FP:FN ratio of 180:1. For power of 50%, the FP:FN ratio is equal to the FP:TP ratio, ie, 20:1. These estimates may underestimate the actual FP:FN ratio, perhaps considerably. The percentage of published findings that are false exceeds the nominal type I error level because of bias. Although confounding, selection and information biases may also contribute to false positives (and occasionally also false negatives), the main problems are selective reporting (publication bias, selective outcome, and analysis reporting bias) and other related biases.[1,2,6,27–29] Because of the same sources of bias, the false-negative rate determined from published results is usually less than the false-negative rate predicted by power considerations alone; when the main analysis is not nominally significant, some investigators may produce and report secondary or subgroup analyses with nominally significant results, regardless of whether the tested association is null or a true effect.

The transformation of the methodology of human genome epidemiology and the emergence of current GWA investigations have changed the values of all the variables of equation 2. First, far more stringent criteria are required for discovery, and typically investigators use $\alpha = 5 \times 10^{-8}$ instead of 0.05.[30,31] Second, increasingly large studies are performed, with collaborative consortia of multiple teams sharing data and performing GWA genotyping and replication efforts integrated into prospective meta-analyses.[32] Third, the use of comprehensive agnostic platforms allows concurrent testing of all genotypes, drastic curtailing of confounding (population stratification) with appropriate methods, and appraisal of all results without selective report-

TABLE 2. Illustrative Estimates of the FP:FN Ratio for Areas of Current Epidemiologic Research

	α	β	R	u	FP:FN
GWAS[a]; common variants with modest effects (OR >1.15); no bias	5×10^{-8}	0.2	10^{-5}	0.0	1:40
GWAS[a]; uncommon variants with modest effects (OR >1.15); no bias	5×10^{-8}	0.9	10^{-5}	0.0	1:180
Traditional case-control study; well-powered; substantial bias	0.05	0.2	0.05	0.3	48:1
Large traditional cohort study; some bias	0.05	0.1	0.05	0.1	32:1
As above; late confirmatory research	0.05	0.1	1	0.1	3.2:1
Large traditional cohort study; unbiased	0.05	0.1	0.05	0	10:1
As above; late confirmatory research	0.05	0.1	1	0	1:2
Large traditional cohort study; unbiased; more stringent threshold	0.001	0.1	0.05	0	1:50
As above; late confirmatory research	0.001	0.1	1	0	1:1000

[a]Genome-wide association studies of genetic variants that have modest effects; nominal level of significance for claiming discovery has been properly adjusted for the fact that the ratio of true positive to true null variants is very small.

OR indicates odds ratio; FP, false positive; FN, false negative.

ing.[11] Although R can vary for different phenotypes, for common variants with minor allele frequency >10% and modest effects, R = 0.00001 is probably a typical value, perhaps with roughly a log-scale window. Assuming no bias (u = 0), one can calculate (Table 2) an FP:FN ratio of 1:40. This is a complete reversal of the ratio compared with the candidate-gene era. False positives have practically disappeared, and there are still a number of common variants with modest effects that are currently false negatives and that can be discovered if larger studies are conducted. This is commensurate with the experience accumulating from many fields, where meta-analyses of GWA studies find a growing number of common variants, as sample size increases. For example, in type 2 diabetes, each single early GWA study identified only a couple of common variants; their meta-analysis led to the discovery of 8 risk variants; the list increased to 18 risk variants with larger meta-analysis, and then to 39 variants with even larger samples.[33–35] Moreover, there is increasing interest in identifying less common gene variants that were not properly targeted by initial GWA efforts. Unless these variants have large effects, power to detect them is low[36]; with power of 10%, the FP:FN ratio is 1:180. This means that in the current epistemic direction of GWA studies, false negatives are a far more common problem than false positives.[37–39]

THE FP:FN RATIO IN TRADITIONAL EPIDEMIOLOGY

Most of the traditional areas of epidemiologic research more closely reflect the performance settings and practices of human genome epidemiology in the pre-GWA era and thus likely have high FP:FN ratios.[1,2,40–42] Under traditional epidemiology, we include disciplines such as nutritional, lifestyle, and occupational epidemiology; much of molecular epidemiology also still follows similar inferential and reporting practices. The operational features of these disciplines resemble the candidate-gene era of genomics—in fact the candidate gene-association studies were conceived, designed, performed, and reported using the templates that had been operational for at least a generation in chronic disease epidemiology. Associations are tested and reported one- or a-few-at-a-time, built around a single-theme hypothesis; statistical significance testing typically uses nominal significance levels of 0.05, despite repeated suggestions to shift to lower thresholds[43] or to Bayesian approaches[44]; confounding is difficult to exclude; raw data are usually not shared in public or across teams; and the pursuit of statistical significance leads to strong publication bias and selective analyses and outcome reporting.[27,45,46] There is considerable flexibility and creativity involved in defining and quantifying exposure in most traditional areas of epidemiology,[27] causing what has been termed "vibration" of effects,[29] the ability to get a wide range of different results, depending on how one analyzes the data. Thus, there is an even greater opportunity for selective analyses, reporting, and interpretation compared with candidate-gene epidemiology, where the exposure (the genotype) is fixed and not open to manipulation other than performing analyses with different models of inheritance. Finally, many studies in these traditional areas of epidemiology either continue to have small total sample sizes, or focus on extreme exposures[28] or subgroup analyses for which the effective sample size is small.

A commonly observed phenomenon in these fields is that case-control studies produce larger effect sizes than subsequent prospective cohort studies,[47] and when putative risk or protective factors are tested in randomized intervention trials, effects are often null.[38,39,48,49] There is a long-standing debate about the epistemic meaning of such nonreplication. Opinions range from the view that a large number[1,2] or even a vast majority[41,42,50–53] of traditional epidemiologic findings are false, to the view that randomized trials may not be appropriate to study all research questions and may provide wrong answers[54] or may be studying different questions,[55] and that the epidemiologic associations are likely correct and should be placed higher in the hierarchy of evidence.[56]

Even if we choose to ignore the failure of traditional epidemiologic results when tested in large-scale randomized trials, one can nonetheless envision a large-scale replication

approach for traditional epidemiologic associations, much as has been possible for human genomics. Very large cohort studies and cohort consortia are available and several hundreds or even a few thousand exposures can be measured and analyzed cumulatively in a standardized, consistent manner.[57,58] However, to date this is not happening in mainstream epidemiology.

Table 2 provides some estimates of the FP:FN ratio under diverse possible settings of traditional epidemiologic investigation. If we assume R = 0.05 (1 genuine effect tested for every 20 null effects tested, ie, a figure consistent with exploratory investigations), bias in the range of u = 0.1–0.3 results in very high FP:FN ratios (32:1 or 48:1 in the corresponding examples). These values of u may be underestimates in the current research environment. Empirical evaluations have shown that almost all published observational studies report some nominally statistically significant results.[28,59] An evaluation of 1915 publications on cancer prognostic markers shows that only 1.4% of them admit to fully "negative" results in their abstracts.[59] False negatives are by definition a subset of all reported negative associations. In literature where there are relatively few reported negatives, there cannot be many false negatives. Even if false positives in observational studies are not as prevalent as comparisons with randomized studies would suggest, one would still have to admit that much of current epidemiology operates at very high FP:FN values, given the relative paucity of negative (hence false negative) reports in the literature.

Even if one were to work in fields with R = 1 (1 genuine effect tested for every null effect tested)—ie, conducting research on topics that are already very well studied and on risk factors that already have very strong prior evidence—the FP:FN ratio would still be 3.2:1 unless bias is accounted for. Elimination of bias, of course, is an idealized target that is unlikely to be reached, particularly if we continue with the same practices of conducting and reporting research from case-control and cohort studies in a fragmented fashion, no matter how well-designed and conducted these studies are. However, if we assume hypothetically that we could eliminate bias, the FP:FN ratio would still be 10:1 for exploratory research, but would drop below 1 (1:2) for late-stage confirmatory research. If we adopted more stringent criteria for significance, eg, $\alpha = 0.001$ rather than $\alpha = 0.05$, and also eliminated bias, then the FP:FN ratio would become 1:50 for exploratory research and 1:1000 for late confirmatory research. We would then have matched current GWA studies in terms of the FP:FN ratio (Table 2).

ARE FALSE POSITIVES AS BAD AS, WORSE THAN, OR BETTER THAN FALSE NEGATIVES?

The same FP:FN ratio may be interpreted differently depending on the relative importance one attaches to a false positive compared with a false-negative result. Modeling these interpretations with decision analytic approaches[60–62] is difficult, because many nonscientific issues need to be considered. If getting a false positive is tolerable and without major consequences, but a false negative has detrimental consequences, then one would not be happy unless the ratio of FP:FN were high. Conversely, if getting a false negative has no major consequences, but a false positive has detrimental consequences, then one would only be satisfied with a low FP:FN ratio.

Much of the debate about false positives and false negatives in epidemiology stems from a poor understanding and communication of the consequences of each type of error, and the conflation of the aims of science[63] with those of public policy.[64] A systematic appraisal is needed on a case-by-case basis, and this often extends beyond the scope of epidemiology as a science, encompassing rather public health policy and political considerations. The eventual appraisal and decision depend on the perspective through which consequences are seen. Consequences differ if scientific validity and accuracy are the paramount concerns. If public health policy and political considerations are paramount, then an entirely different form of deliberation is necessary. In each situation, policy- and decision-makers may rationalize, discuss, and juxtapose the specific costs associated with false positives and the specific costs associated with false negatives, to interpret the estimated FP:FN ratio. In the best of situations, science may inform such a deliberation but cannot prescribe which action or approach to take. There is an inherent tension—if not conflict—between epidemiology as a science and epidemiology as public policy.[64]

One might tolerate false positives better than false negatives in scientific fields where findings can be confirmed or refuted quickly and with reasonable cost; however, this is rarely the situation in most areas of epidemiology. Even the most conclusively refuted associations such as beta-carotene for cancer prevention continue to get die-hard supporting citations from their followers 2 decades after their refutation.[65] False positives may also be better tolerated when interest is primarily in generating hypotheses for further scientific exploration, with no immediate clinical or public policy implications (ie, there are no medical or public actions to be taken). However, by its own nature of dealing with common diseases and common exposures, epidemiology captures the public's attention more than most other areas of scientific research; it has been demonstrated that newspapers prefer reporting bad news from observational studies compared with good news and compared with randomized trials.[66] This can have adverse consequences, including a loss of confidence (by the public as well as the wider medical research community) in the credibility of epidemiology as a science. Thus, we believe that epidemiologists should exercise caution in making definitive scientific claims, as should research journals and universities in their press releases, with the hope that such caution may lessen the impact of false-

positive findings. Conversely, public health policy decision-making sometimes operates at higher FP:FN ratios, but in such situations one should expect evidence that the public health intervention or action does more good than harm. The threshold of acceptability remains open to discussion and needs to be revisited in each area of epidemiologic research, with application on a case-by-case basis in the ever-shifting sands of the interface between science and public health policy.

REFERENCES

1. Boffetta P, McLaughlin JK, La Vecchia C, Tarone RE, Lipworth L, Blot WJ. False-positive results in cancer epidemiology: a plea for epistemological modesty. *J Natl Cancer Inst*. 2008;100:988–995.
2. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2:e124.
3. Blair A, Saracci R, Vineis P, et al. Epidemiology, public health, and the rhetoric of false positives. *Environ Health Perspect*. 2009;117:1809–1813.
4. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* Chichester: Wiley; 2004.
5. 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. *J Natl Cancer Inst*. 2004;96:434–442.
6. 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.
7. Davey Smith G, Ebrahim S. Data dredging, bias, or confounding. *BMJ*. 2002;325:1437–1438.
8. Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. *Lancet.* 2003;361:865–872.
9. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med*. 2002;4:45–61.
10. Lohmueller 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.
11. Kraft P, Zeggini E, Ioannidis JP. Replication in genome-wide association studies. *Stat Sci*. 2009;24:561–573.
12. Hirschhorn JN. Genomewide association studies—illuminating biologic pathways. *N Engl J Med*. 2009;360:1699–1701.
13. Chanock S. High marks for GWAS. *Nat Genet*. 2009;41:765–766.
14. Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet*. 2009;10:318–329.
15. Bosker FJ, Hartman CA, Nolte IM, et al. Poor replication of candidate genes for major depressive disorder using genome-wide association data. *Mol Psychiatry*. In press.
16. Caporaso N, Gu F, Chatterjee N, et al. Genome-wide and candidate gene association study of cigarette smoking behaviors. *PLoS One*. 2009;4:e4653.
17. Morgan TM, Krumholz HM, Lifton RP, Spertus JA. Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study. *JAMA*. 2007;297:1551–1561.
18. Richards JB, Kavvoura FK, Rivadeneira F, et al. Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture. *Ann Intern Med*. 2009;151:528–537.
19. Samani NJ, Erdmann J, Hall AS, et al. Genomewide association analysis of coronary artery disease. *N Engl J Med*. 2007;357:443–453.
20. Scuteri A, Sanna S, Chen WM, et al. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet*. 2007;3:e115.
21. Sõber S, Org E, Kepp K, et al. Targeting 160 candidate genes for blood pressure regulation with a genome-wide genotyping array. *PLoS One*. 2009;4:e6034.
22. Wu H, Romieu I, Shi M, et al. Evaluation of candidate genes in a genome-wide association study of childhood asthma in Mexicans. *J Allergy Clin Immunol*. 2010;125:321.e13–327.e13.
23. Khoury MJ, Bertram L, Boffetta P, et al. Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. *Am J Epidemiol*. 2009;170:269–279.
24. Siontis KC, Patsopoulos NA, Ioannidis JP. Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *Eur J Hum Genet*. 2010;18:832–837.
25. Hindorff LA, Junkins HA, Manolio TA. A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies. Accessed August 1, 2010.
26. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009;106:9362–9367.
27. Phillips CV. Publication bias in situ. *BMC Med Res Methodol*. 2004;4:20.
28. Kavvoura FK, Liberopoulos G, Ioannidis JP. Selection in reported epidemiologic risks: an empirical assessment. *PLoS Med*. 2007;4:e79.
29. Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology*. 2008;19:640–648.
30. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol*. 2008;32:381–385.
31. Hoggart CJ, Clark TG, De Iorio M, Whittaker JC, Balding DJ. Genome-wide significance for dense SNP and resequencing data. *Genet Epidemiol*. 2008;32:179–185.
32. Zeggini E, Ioannidis JP. Meta-analysis in genome-wide association studies. *Pharmacogenomics*. 2009;10:191–201.
33. Zeggini E, Weedon MN, Lindgren CM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*. 2007;316:1336–1341.
34. Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*. 2008;40:638–645.
35. Voight BF, Scott LJ, Steinthorsdottir V, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet*. 2010;42:579–589.
36. Panagiotou O, Evangelou O, Ioannidis JP. Genome-wide significant associations for variants with minor allele frequency <5%: an overview. *Am J Epidemiol*. 2010;172:869–889.
37. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010;42:565–569.
38. Park JH, Wacholder S, Gail MH, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet*. 2010;42:570–575.
39. Gibson G. Hints of hidden heritability in GWAS. *Nat Genet*. 2010;42:558–560.
40. Kabat GC. *Hyping Health Risks: Environmental Hazards in Every Day Life and the Science of Epidemiology.* New York: Columbia University Press; 2008.
41. Maziak W. The triumph of the null hypothesis: epidemiology in an age of change. *Int J Epidemiol*. 2009;38:393–402.
42. Young S. Acknowledge and fix the multiple testing problem. *Int J Epidemiol*. 2010;39:934; author reply 934–935. doi: 10.1093/ije/dyp188.
43. Sterne JA, Davey Smith G. Sifting the evidence—what's wrong with significance tests? *BMJ*. 2001;322:226–231.
44. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. *Ann Intern Med*. 1999;130:1005–1013.
45. Dwan K, Altman DG, Arnaiz JA, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One*. 2008;3:e3081.
46. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev*. 2009:MR000006.
47. Riboli E, Norat T. Epidemiologic evidence of the protective effect of fruit and vegetables on cancer risk. *Am J Clin Nutr*. 2003;78(suppl 3):559S–569S.
48. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294:218–228.

49. Prentice RL, Pettinger M, Anderson GL. Statistical issues arising in the Women's Health Initiative. *Biometrics*. 2005;61:899–911; discussion 911–941.
50. Skrabanek P. Has risk-factor epidemiology outlived its usefulness? *Am J Epidemiol*. 1993;138:1016–1017.
51. Shapiro S. Looking to the 21st century: have we learned from our mistakes, or are we doomed to compound them? *Pharmacoepidemiol Drug Saf*. 2004;13:257–265.
52. Feinstein AR. Scientific standards in epidemiologic studies of the menace of daily life. *Science*. 1988;242:1257–1263.
53. Austin PC, Mamdani MM, Juurlink DN, Hux JE. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol*. 2006;59:964–969.
54. Stampfer M. Observational epidemiology is the preferred means of evaluating effects of behavioral and lifestyle modification. *Control Clin Trials*. 1997;18:494–499; discussion 514–516.
55. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. 2008;19:766–779.
56. Vandenbroucke JP. Observational research, randomized trials, and two views of medical science. *PLoS Med*. 2008;5:e67.
57. Ioannidis JP, Loy EY, Poulton R, Chia KS. Researching genetic versus non-genetic determinants of disease: a comparison and proposed unification. *Sci Transl Med*. 2009;1:7ps8.
58. Patel CJ, Bhattacharya J, Butte AJ. An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS One*. 2010;5:e10746.
59. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. *Eur J Cancer*. 2007;43:2559–2579.
60. Hozo I, Schnell MJ, Djulbegovic B. Decision-making when data and inferences are not conclusive: risk-benefit and acceptable regret approach. *Semin Hematol*. 2008;45:150–159.
61. Hozo I, Djulbegovic B. When is diagnostic testing inappropriate or irrational? Acceptable regret approach. *Med Decis Making*. 2008;28:540–553.
62. Djulbegovic B, Hozo I. When should potentially false research findings be considered acceptable? *PLoS Med*. 2007;4:e26.
63. Merton RK. The normative structure of science. In: Storer NW, ed. *The Sociology of Science. Theoretical and Empirical Investigations.* Chicago and London: Chicago University Press; 1973.
64. Pielke RA Jr. *The Honest Broker: Making Sense of Science in Policy and Politics*. Cambridge; Cambridge University Press; 2007.
65. Tatsioni A, Bonitsis NG, Ioannidis JP. Persistence of contradicted claims in the literature. *JAMA*. 2007;298:2517–2526.
66. Bartlett C, Sterne J, Egger M. What is newsworthy? Longitudinal study of the reporting of medical research in two British newspapers. *BMJ*. 2002;325:81–84.

