Why Most Discovered True Associations Are Inflated

John P. A. Ioannidis

Abstract: Newly discovered true (non-null) associations often have inflated effects compared with the true effect sizes. I discuss here the main reasons for this inflation. First, theoretical considerations prove that when true discovery is claimed based on crossing a threshold of statistical significance and the discovery study is underpowered, the observed effects are expected to be inflated. This has been demonstrated in various fields ranging from early stopped clinical trials to genome-wide associations. Second, flexible analyses coupled with selective reporting may inflate the published discovered effects. The vibration ratio (the ratio of the largest vs. smallest effect on the same association approached with different analytic choices) can be very large. Third, effects may be inflated at the stage of interpretation due to diverse conflicts of interest. Discovered effects are not always inflated, and under some circumstances may be deflated—for example, in the setting of late discovery of associations in sequentially accumulated overpowered evidence, in some types of misclassification from measurement error, and in conflicts causing reverse biases. Finally, I discuss potential approaches to this problem. These include being cautious about newly discovered effect sizes, considering some rational down-adjustment, using analytical methods that correct for the anticipated inflation, ignoring the magnitude of the effect (if not necessary), conducting large studies in the discovery phase, using strict protocols for analyses, pursuing complete and transparent reporting of all results, placing emphasis on replication, and being fair with interpretation of results.

(*Epidemiology* 2008;19: 640–648)

The discovery and replication of associations is a core activity of quantitative research. This article will not deal with the debate on whether research findings are credible.[1] I will focus instead on the interesting subset of research findings that are true. Research findings discussed here encompass all types of associations that emerge from quantitative measurements, and are expressed as effect metrics. This includes treatment effects from clinical trials, measures of risk for observational risk factors, prognostic effects for prognostic studies, and so forth. I start here with the assumption that a research finding is indeed true (non-null), ie, it reflects a genuine association that is not entirely due to chance or biases (confounding, misclassification, selection biases, selective reporting, or other). The question is: do the effect sizes for such associations, at the time they are first discovered and published in the scientific literature, accurately reflect the true effect sizes?

The article has the following sections: a brief literature review on inflated early-effect sizes based on theoretical and empirical considerations; a description of the major reasons why early discovered effects are inflated and the major countering forces that may occasionally lead to deflated effects (underestimates); and suggestions on how to deal with these problems.

Evidence About Inflated Early-Effect Sizes

Table 1 cites articles suggesting that early studies give (on average) inflated estimates of effect.[2–34] I list here only selected evaluations that cover either many different articles/effects or a whole research domain or method. This list is nowhere close to exhaustive. For some topics, such as the inflation of regression coefficients for variables selected through stepwise statistical-significance-based processes, the literature is vast. The theme of inflated early effects has been encountered in various disguises in many scientific disciplines in the biomedical sciences and beyond. For empirical studies, it may not be known whether the subsequent studies are more correct than the original discovery, but when a pattern is seen repeatedly in a field, the association is probably real, even if its exact extent can be debated. One should also acknowledge the difficulty in differentiating between an early inflated but true (non-null) effect and an entirely false (null) one. In addition to empirical studies, however, Table 1 also includes theoretical work that proves why inflation is anticipated; some of these arguments are discussed in the next section.

I mention here a few examples to demonstrate the seriousness of the problem. The prognostic significance of a 70-gene expression signature for lymph-node-negative breast cancer is accepted beyond doubt.[35] However, while the first study published in *Nature* showed almost perfect sensitivity and specificity, even in an independent replication exercise of 19 patients,[36] subsequent evaluation in a cohort of 307 women showed sensitivity of 90% and specificity of only 40% (AUC for survival 0.648).[37] Prognostic ability is

Submitted 17 March 2008; accepted 27 May 2008; posted 14 July.
From the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece; and Department of Medicine, Tufts University School of Medicine, Boston, Massachusetts.
Editors' note: *Related articles appear on pages 649, 652, 655, and 657.*
Correspondence: John P. A. Ioannidis, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, 45 110 Ioannina, Greece. E-mail: jioannid@cc.uoi.gr.

ISSN: 1044-3983/08/1905-0640
DOI: 10.1097/EDE.0b013e31818131e7

TABLE 1. Selected Evaluations Suggesting That Early Discovered Effects Are Inflated

Research Field	Theoretical Work or Empirical Evidence and References
Highly cited clinical research	A quarter of most-cited clinical trials and 5/6 most-cited epidemiological studies were either fully contradicted or found to have exaggerated results[2]
Early stopped clinical trials	Early stopping results in inflated effects in theory[3,4] and shown also in practice[5]
Clinical trials of mental health interventions	More likely for effect sizes of pharmacotherapies to diminish than to increase over time[6]
Clinical trials on heart failure interventions	"Regression to the truth" in phase III trials for interventions with early promising results[7]
Clinical trials on diverse interventions	Effectiveness shown to fade over time[8]
Multiple meta-analyses on effectiveness	Eleven independent meta-analyses on acetylcysteine show decreasing effects over time[9]
Epidemiologic associations	Expected to be inflated in multiple testing with significance threshold; empirical demonstration for occupational carcinogens[10]
Pharmacoepidemiology	"Phantom ship" associations that do not stand upon further evaluation[11]
Gene-disease associations	Several empirical evaluations showing dissipation of effect sizes over time[12–15]
Linkage studies in humans	Theory anticipates large upward bias ("winner's curse") in effects of discovered loci[16–18]
Genetic traits in experimental crosses	As above (actually literature on the "Beavis effect" precedes literature on humans)[19–22]
Genome-wide associations	Large winner's curse anticipated for discovered effects in underpowered conditions[23,24]
Ecology and evolution	Empirical demonstration that relationships fade over time[25,26]
Psychology	Replication studies in psychology failing to confirm true effects because the new studies were underpowered due to reliance on the estimate of effect from the original positive study[27]
Early repeated data peaking in general	Simulations to model inflation of effects with repeated data peaking[28]
Prognostic models	Overestimated prognostic performance with stepwise selection of variables based on significance thresholds[29–32]
Regression models in general	Exaggerated effects (coefficients) with stepwise selection based on significance thresholds and small datasets[32–34]; may correct substantially if a very lenient alpha = 0.20 is used for selection[34] [thus having enough power]

present, but the difference between an almost-perfect predictor and a modest-to-poor predictor is prominent.[35]

Many high-profile clinical trials are stopped early during their conduct. This is performed according to robust rules that suggest termination when a demanding threshold of statistical significance is crossed during an interim analysis.[3,4] These interventions are indeed effective (the null of "no effectiveness" is correctly rejected). However, as shown both in theory[3,4] and in practice,[5] the effect sizes derived from such early terminated trials are inflated. With very early termination, the effect sizes may be markedly inflated,[5] with implications for decision-making in the use of these interventions.

Theoretical considerations prove that linkage signals of genome-wide linkage studies are inflated.[12–15] These studies have aimed to reveal loci that harbor genetic variants that are related to various phenotypes. Several thousands of such studies conducted over 2 decades have yielded very few replicated hits. Although the replication record is better with genome-wide association studies, theoretical considerations again show the early discovered effects are inflated.[23,24] Furthermore, if the observed effects are used as estimates in designing replication studies, these subsequent studies will be underpowered, and genuine effects will be falsely nonreplicated.[38]

Inflated Effect Sizes Due to Selection Thresholds and Suboptimal Power

Effect sizes of newly discovered true (non-null) associations are inherently inflated on average. This is due to the key characteristic of the discovery process. Inflation is expected when, to claim success (discovery), an association has to pass a certain threshold of statistical significance, and the study that leads to the discovery has suboptimal power to make the discovery at the requested threshold of statistical significance. Both conditions are necessary to inflate effect sizes. If investigators were not fixated on claiming discoveries based on *P* value thresholds, this would not be an issue. Similarly if the discovery studies were fully powered, inflation would not be an issue. Selection usually entails *P* values, but a similar pattern may be seen if selection is based on effect size or some other threshold measure.

For illustrative purposes, I use here a simulation approach to demonstrate this phenomenon and the relationship between inflation and lack of power. Suppose that the true odds ratio (OR) for an association is 1.10 or 1.25 and that the proportion of exposed individuals in the control group is 30%. We can simulate a set of studies that have an equal number of participants (n) in each of the 2 compared groups. The number of exposed in the control group in each simulated study is drawn randomly from a binomial distribution with probability 0.30. The number of exposed in the case group in each simulated study is drawn randomly from a binomial distribution with probability 0.3203 or 0.3488, so as to correspond to OR = 1.10 and 1.25, respectively. The median OR of these simulated studies is expected to be 1.10 or 1.25, respectively. However, this is not so when we focus only on

TABLE 2. Simulations for Effect Sizes Passing the Threshold of Formal Statistical Significance ($P = 0.05$)

True OR	Control Group Rate (%)	Sample n Per Group	Observed OR in Significant Associations: Median (IQR)	Observed OR in Significant Associations: Median Fold Inflation
1.10	30	1000	1.23 (1.23–1.29)	1.11
1.10	30	250	1.51 (1.49–1.55)	1.37
1.25	30	1000	1.29 (1.26–1.39)	1.03
1.25	30	250	1.60 (1.50–1.67)	1.28
1.25	30	50	2.73 (2.60–3.16)	2.18

IQR indicates interquartile range.

the simulated studies that have a *P* value for the association crossing a specific level of statistical significance. Table 2 shows the median and IQR of the ORs that cross the "*P* value = 0.05" threshold of statistical significance for different values of n. As shown, even though the true OR is 1.10, the median observed OR when a study discovers this association ($P < 0.05$) is 1.51 when n = 250 (a study of 500 participants total). With similar sample sizes, when the true OR = 1.25, the discovered median OR is 1.60. When the studies have n = 50 (100 participants total), the median discovered OR is 2.73 instead of 1.25, representing huge inflation. One should note also the skewed nature of the distributions of discovered effects.

One may argue that we do not know the true effect sizes necessary to make these simulations for specific hypotheses. In the example above, if the true OR were 500, then studies with 250 participants per group would have excellent power to detect it at $\alpha = 0.05$ and the discovered effects would not be inflated compared with the true OR = 500. In some fields, there may be considerable uncertainty about the magnitude of the true effect sizes. However, in most fields, we can make reasonable guesses about the effect sizes, with only modest uncertainty. For example, in genetic associations of common variants with common diseases, we have repeatedly found that effect sizes of consistently and extensively replicated associations tend to be small or even very small (most ORs = 1.1–1.4; a few, 1.4–2).[39–41] Similarly, for most medical interventions with hard clinical outcomes (including mortality) relative risk decreases of 10%–30% are the best we can hope for. Some fields that have proposed much larger effect sizes may simply need a reality check. Perhaps some of these fields have been stuck in doing underpowered studies, and thus effects circulating in their literature appear large when they are actually much smaller.

Inflated Effects Due to Flexible Analyses (Vibration of Effects) and Selective Reporting

Until now, we have assumed that the (simulated) studies arise out of the play of chance alone. We have assumed that there is no human intervention in the analysis process and there is only one analysis based on the observed results. This situation is rare in discovery research. The hallmark of discovery is the performance of exploratory analyses. Flexible analyses lead to vibration of effects. Vibration conveys the extent to which an effect may change in alternative analytical approaches.

Vibration is mostly due to the availability of alternative options in statistical model selection (eg, Cox model for time-to-death vs. logistic regression for death in 30 days); statistical inference machine (eg, different methods for computation of the odds ratio [eg, with or without Wolf correction and with different corrections of zero cells] and its variance[42]); data selection (eg, possibility to exclude or include some participants based on some partly prespecified, prespecified but ambivalent, or entirely post hoc criteria); dependent arbitration of equivocal data; and wide choice of adjustments for other covariates (especially when there are many such). Changes may affect not only the analytic core but also the question formulation itself, eg, changing eligibility criteria may modify the research question.

I define the vibration ratio for effect size as the ratio between the extremes of effect sizes that can be obtained in the same study under different analytical options. In Figure 1, I have analyzed the same dataset (250 participants) with different approaches. Unadjusted analysis yields OR = 2.10 (95% confidence interval [CI] = 1.18–3.72). I simulate 2 random variables and also perform analyses adjusting the association for each one of them. The vibration ratio is only 1.01. I simulate another random variable and perform analyses where the top 6% or the top 10% of the participants for this random variable are considered noneligible for the analysis. The vibration ratio is 1.18. Then, I also simulate 5 observations (only 2% of the data) for which exposure is considered equivocal, and is either changed to specifically agree with the direction of the association or is changed to specifically disagree with the direction of the association. The vibration ratio is 1.55. The possible combinations of random adjustment, random eligibility, and dependent arbitration, as above, yield a vibration ratio of 1.95: ORs as divergent as 1.48 (CI 0.81–2.70) and 2.88 (1.55–5.35) are obtained with these relatively subtle options. Without trying hard, I changed the OR 2-fold.

The vibration ratio will be larger in small datasets and in those with hazy definitions of variables, unclear eligibility criteria, large numbers of covariates, and no consensus in the field about what analysis should be the default. In most discovery research, this explosive mix is the rule. It is difficult to obtain funding to run very large studies for taking a first shot into the dark, and discovery is inherently related to situations where hazy definitions and iterative searching abound. The wealth of databases in covariates has also grown over time.

Even if enormous, vibration alone would not lead to inflated discovered effects if one eventually presents all the

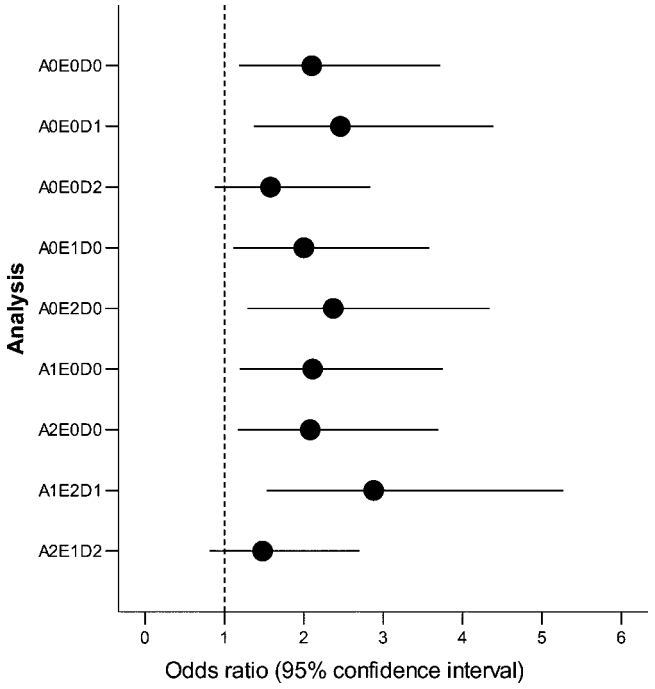

FIGURE 1. Vibration of an effect size: the odds ratio with 95% confidence interval is obtained for a simulated study, with or without adjustment (A0 indicates no adjustment; A1, adjustment for one randomly generated variable; A2, adjustment for another randomly generated variable), application of various eligibility criteria (E0 indicates all participants included; E1, excluding 6% of the participants according to high values on a random variable; E2, excluding 10% of the participants according to high values on a random variable) and arbitration of 2% of the data on the exposure based on knowledge of outcome (D0 indicates no arbitration; D1, 2% of the exposure data [5 observations] changed to be consistent with the direction of the association; D2, 2% of the exposure data [5 observations] changed to be against the direction of the postulated association).

applied analytical options without any preference. However, typically only one or a few analyses are presented. Moreover, vibration would not be a problem if the one or few analyses selected for presentation were a random choice of the possible ones, selected with an impartial view and no interest in making a discovery. However, this is counterintuitive to the discovery process. One makes exploratory analyses specifically to find something. The effects selected for presentation are likely to be among the largest observed, if not the largest possible. Secondary analyses similarly may be chosen to show that they are consistent with the main selected analysis.

Selective analyses and outcome reporting have been extensively demonstrated in clinical-trials research comparing protocols against reported results.[43–45] In theory, randomized trials have more inflexible protocols compared with observational epidemiology and fully exploratory research. For observational research, similar evaluations are more difficult to conduct because protocols are not readily available—often there is no protocol at all. Empirical evidence has demonstrated across a large sample of 379 epidemiologic studies that investigators selected the contrasts for continuous variables so as to show effects as being larger: more extreme contrasts were presented, when effects were inherently smaller.[46]

Post hoc demonstration of selective analysis and outcome reporting is difficult. Recently, a test was proposed to examine whether the number of reported study results that pass certain levels of statistical significance is reasonable or larger than what one would expect, even if the effect sizes for the proposed associations (eg, as suggested by meta-analyses of all relevant studies) were true.[47] Testing has suggested substantial selective reporting biases in both clinical trials and observational epidemiology.[12,47–49]

Inflated Interpretation for Effect Sizes

Inflated interpretation is the toughest of all sources of inflation to tackle. In a culture that rewards discovery, investigators may make an extra effort to present results in the most favorable way. This goes beyond selective reporting and enters the realm of qualitative interpretation of quantitative effects. Typical variants of inflated interpretation include unwarranted extrapolations and over-stated generalizability,[50] silencing or downplaying limitations and caveats,[51] mishandling external evidence,[52,53] and extension of promises to different inferential levels. In the last category, some typical leaps of faith in the epidemiologic literature include the interpretation of association as causation, the interpretation of association or even causation as anticipated treatment effects, and the interpretation of optimal efficacy as effectiveness in everyday life and clinical practice. In the molecular literature, a typical leap of faith is the interpretation that a modest association pointing to a new biologic pathway can be translated into a major benefit for treatment of diseases that may somehow be involved in this pathway. The sparse successful clinical translation of major promises made in the most high-profile basic science journals shows that this over-interpretation is common.[54]

Why Published True Associations May Sometimes Have Deflated Effects

Contrary to the above, some discovered associations may have deflated effect sizes compared with the true ones. For example, this may occur with overpowered studies, where interim looks at the data are performed at early stages and discovery happens late. If the association does not cross the desired threshold of significance at the interim looks, but only at the very end, the effect may be deflated, although the deflation is typically small.[3,4] The same situation would arise if the discovery process occurs as a regularly updated prospective meta-analysis, a true association gets discovered

(becomes formally significant for the first time) only after many studies have been performed and combined in the meta-analysis, and the power of these combined studies is high to detect such as association. Nevertheless, in most fields, overpowered studies at the discovery phase are still a small minority compared with underpowered studies[55–60]; moreover, the paradigm of prospective cumulative meta-analysis as a discovery tool has not been widely disseminated.

Another reason for deflated effect sizes is independent nondifferential misclassification due to measurement error in the associated variables. There is an extensive literature on misclassification and how to correct effect sizes for misclassification.[61] However, such corrections have never become main stream. Perhaps this is because usually nonindependent and differential misclassification has been difficult to exclude, and these can either deflate or inflate observed effects.[62,63] Measurement error has decreased over time for many fields of research in the current era. For example, genetic measurements have very minor measurement error if measurement platforms are used properly. Conversely, for some other variables, (eg, lifestyle), measurement error may remain substantial. Even in molecular/genetic epidemiology, misclassification remains important for evaluating gene-environment interactions.[64–67] Of note, when effects diminish because of misclassification, power to detect them also diminishes sharply[68]; this enhances the inflation upon discovery (inflation of a deflated effect), as above.

Furthermore, vibration of effects with selective reporting and interpretation of effects may sometimes reflect reverse biases. Various conflicts of interest may work in the direction of silencing or diminishing newly discovered associations that don't fit financial or other dogmatic perspectives. For therapeutic research, although financial conflicts may lead to inflation of treatment effects for new interventions,[69] they may similarly lead to deflation of the magnitude of adverse events.[70] For example, although most meta-analyses[71,72] of rosiglitazone found ORs for myocardial infarction in the range of 1.43, a meta-analysis originally conducted by Glaxo found a more conservative OR and the company did not consider it to be of concern.[73] However, the literature on adverse events of interventions is small compared with the literature on effectiveness.[74] Most harms probably remain unknown rather than silenced.[75]

Finally, conflicts may be of nonfinancial nature. Some investigators may fervently support their line of research and beliefs. For example, even the most strongly refuted associations continue to have supporters many years after the refutation.[76] Investigators may suppress new findings when they do not suit their beliefs.

What To Do

At the time of first postulated discovery, we usually can not tell whether an association exists at all,[1] let alone judge its effect size. As a starting principle, one should be cautious about effect sizes. Uncertainty is not conveyed simply by CIs (no matter if these are 95%, 99%, or 99.9% CIs) (Table 3).

For a new proposed association, credibility and accuracy of its proposed effect varies depending on the case. One may ask the following questions: does the research community in this field adopt widely statistical significance or similar selection thresholds for claiming research findings? Did the discovery arise from a small study? Is there room for large flexibility in the analyses? Are we unprotected from selective reporting (eg, was the protocol not fully available upfront)? Are there people or organizations interested in finding and promoting specific "positive" results? Finally, are the counteracting forces that would deflate effects minimal?

Modeling or correcting some of the sources of inflation is possible with (more) appropriate methods, such as for genetic linkage or association[17,23] or for regression coefficients in general.[33,77] These methods are probably more useful in estimating expected effect sizes, so as to perform more proper power calculations for future replication efforts, rather than for claiming that accurate "corrected" estimates of effect are known. In each case, one has to ask whether it is appropriate to ignore completely the effect size for a new proposed association. It may be best to wait for additional, larger studies and cumulative evidence to reach a more firm conclusion on whether an effect exists at all, and then worry about its size later. Most fields can wait for the conduct of replication studies.

The conduct of larger studies in the discovery phase will diminish inflation due to suboptimal power. However, this is not always feasible. Discovery may sometimes arise from small investigations or even unanticipated case observations.[70] However, even if many discoveries in the past arose out of haphazard encounters of scientists with phenomena, this does not mean that we cannot improve in the future by running larger discovery-oriented studies. Agnostic genome-wide associations provide such an example.[78]

Using a strict protocol for the design, conduct, and analysis of a study can diminish vibration, but would this stifle creativity? Flexible analyses will not cause a problem if

TABLE 3. Avoiding Being Misled on Effect Sizes of True Associations in Early Discovery

Be cautious about effect sizes (and even about the mere presence of any effect in new discoveries)
Consider rational down-adjustment of effect sizes
Consider analytical methods that correct for anticipated inflation
Ignore effect sizes arising from discovery research
Conduct large studies in discovery phase
Use strict protocols for analyses
Adopt complete and transparent reporting of all results
Use methodologically rigorous, unbiased replication (potentially ad infinitum)
Be fair with interpretation

they are accompanied by complete and transparent reporting of all results. Despite demonstrable progress and the availability of evidence-based guidance for reporting, such as CONSORT,[79,80] STROBE,[81] and STARD,[82] full reporting remains an unattained target even in fields such as randomized trials, which are further ahead in registration and reporting efforts.[83,84] Making databases publicly available is more easily said than done, and there are many challenges in making this a widespread practice.[85,86] Still, the antithesis of practices among various fields is striking. For example, genome-wide associations studies currently test hundreds of thousands of associations, ask for very demanding thresholds (eg, $P < 10^{-7}$), report all results in a single paper, and then often make the data publicly available.[87,88] Conversely, in traditional risk-factor epidemiology (eg, nutritional epidemiology), each (or a few) of the thousands of tested associations is reported as a single separate paper, "$P < 0.05$" rules are still widespread, and databases rarely become public. Imagine what would happen if the criteria of genome-wide association studies were applied to nutritional epidemiology associations. There are clearly other major differences among such fields,[89] but one wonders whether such widely discrepant practices are justified. Inclusive consortia of investigators may also help enhance transparency and completeness of reporting of results.[90]

Discovery can be unfettered, haphazard, exploratory, opportunistic, selective, and highly subjectively interpreted. Conversely, these same characteristics that are perfectly fine for discovery are not desirable of replication. Replication is essential for all discoveries and with few exceptions (eg, treatment effects in interventional studies) only resource constrains and prioritization issues would prohibit replication ad infinitum. Replication offers a wider evidence base on which to try to make inferences about the truth and biases that may affect it.

A crucial question is whether replication suffices to correct the inflated effects that arise in early studies.[91,92] For example, should a meta-analysis worry about including an early terminated study? In principle, the replication process, if unbiased, should correct the inflation[91] and if stopping is not very early, inflation is small regardless.[93] However, the replication process may not be unbiased, and may sometimes suffer from similar problems as (or more problems than) the discovery. Observational evidence has been attacked as unreliable, and even the best meta-analyses of observational data meet with skepticism for their spurious precision.[94] Problems may arise, however, even for the supposedly more rigorous design of randomized trials. To demonstrate this problem, an evaluation of the whole Cochrane Library shows 1011 systematic reviews that have at least one meta-analysis with at least 4 studies.[95] Selecting the largest meta-analysis in each of these reviews, 256 of the 1011 meta-analyses have formally statistically significant results ($P < 0.05$) by random effects calculations in the OR scale. The effect sizes of these

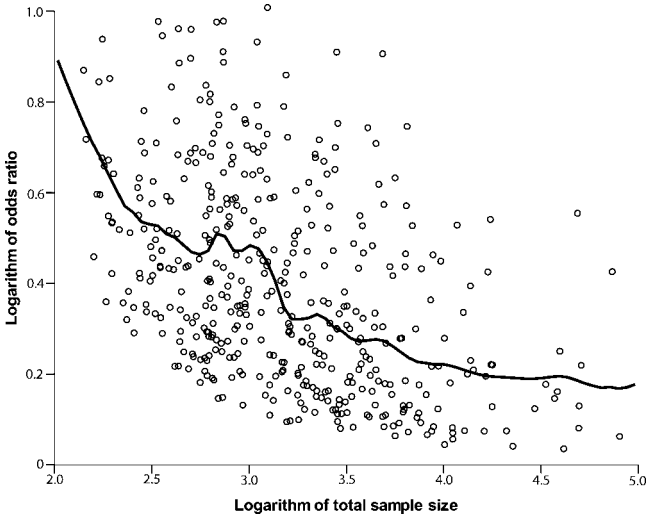

FIGURE 2. Relationship between total sample size and the effect size (odds ratio) for 256 Cochrane meta-analyses with formally statistically significant results ($P < 0.05$ according to random effects calculations) and at least 4 included studies. Both axes are in log10 scale. Also shown is a fit LOESS line. All odds ratios have been coined to be >1.00 for consistency. The median effect size for the 40 meta-analyses with at least 10,000 subjects is 1.53. Not shown are 5 outliers with extreme sample size or effect size.

"positive" meta-analyses are inversely related to the amount of evidence accumulated (Fig. 2). Perhaps large anticipated effects lead to the conduct of small trials and small anticipated effects promote several large trials. However, the observed pattern is consistent with what one would expect based on the inflation biases described above. Most meta-analyses remain largely underpowered for small-to-modest effects.[96] Superimposed selective reporting can also be operating. Thus, even in the theoretically most rigorous study design (randomized trials), not only discoveries but also pragmatically limited replication efforts may not eliminate inflation of effects, and may not even ensure that any effect at all is present.

What constitutes fair interpretation of new discoveries is unavoidably subjective. However, critical discussion of limitations, caveats, and a reserved stance against one's findings is useful. Thresholds of significance that dictate a discovery may have to be abolished. Instead, all results would be reported, grading their credibility and the uncertainty thereof in a Bayesian framework. Suggestions to adopt Bayesian views of research results have long been made.[1,11,97–103] However, inflation of effects may still be an issue, even if effects are selected based on Bayes factor thresholds rather than P value thresholds. This depends on how Bayes factors are calculated. For example, direct translation[99] of P values (or z-scores) to minimum Bayes factors, $\exp(-z^2/2)$, would face the same problem, whereas if priors assume that small effects are

TABLE 4. Two Stances in Hunting Associations

	Aggressive Discoverer	Reflective Replicator
What matters is …	Discovery	Replication
Databases are …	Private goldmines not to be shared	Public commodity
A good epidemiologist …	Can think of more exploratory analyses	Is robust about design and analysis plan
One should report …	What is interesting	Everything
Publication mode	Publish each association as a separate paper	Publish everything as single paper
After reporting …	Push your findings forward	Be critical/cautious

plausible but large effects are implausible, Bayes factors become most promising for small effects.[103] Bayesian views are useful when coupled with unselective presentation of all results. In this way, one can see which results are more interesting based on different prior assumptions, and whether there is consistency in highlighting specific results. New results modify future priors. If new results are biased because of selection, priors get biased and we may keep pursuing, believing, and expecting nonexistent large effects.

Finally, Table 4 summarizes 2 stances in hunting associations—the aggressive discoverer versus the reflective replicator. These stances may underlie the root of the problems that I discussed here, and their possible solutions. In trying to reward or punish scientists for their stance and in shaping the new generation of scientists, we need to think hard about which of the 2 modes we want to promote, and whether some good elements can be picked from each list.

ACKNOWLEDGMENTS

I am grateful to Duncan Thomas for helpful comments.

REFERENCES

1. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2:e124.
2. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294:218–228.
3. Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. *Control Clin Trials*. 1989;10(4 suppl): 209S–221S.
4. Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. *Stat Med*. 1988;7:1231–1242.
5. Montori VM, Devereaux PJ, Adhikari NK, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA*. 2005;294:2203–2209.
6. Trikalinos TA, Churchill R, Ferri M, et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *J Clin Epidemiol*. 2004;57:1124–1130.
7. Krum H, Tonkin A. Why do phase III trials of promising heart failure drugs often fail? The contribution of "regression to the truth." *J Card Fail*. 2003;9:364–367.
8. Gehr BT, Weiss C, Porzsolt F. The fading of reported effectiveness. A meta-analysis of randomised controlled trials. *BMC Med Res Methodol*. 2006;6:25.
9. Bagshaw SM, McAlister FA, Manns BJ, et al. Acetylcysteine in the prevention of contrast-induced nephropathy: a case study of the pitfalls in the evolution of evidence. *Arch Intern Med*. 2006;166:161–166.
10. Thomas DC, Siemiatycki J, Dewar R, et al. The problem of multiple inference in studies designed to generate hypotheses. *Am J Epidemiol*. 1985;122:1080–1095.
11. Hauben M, Reich L, Van Puijenbroek EP, et al. Data mining in pharmacovigilance: lessons from phantom ships. *Eur J Clin Pharmacol*. 2006;62:967–970.
12. Ntzani EE, Rizos EC, Ioannidis JP. Genetic effects versus bias for candidate polymorphisms in myocardial infarction: case study and overview of large-scale evidence. *Am J Epidemiol*. 2007;165:973–984.
13. Ioannidis JP, Trikalinos TA, Ntzani EE, et al. Genetic associations in large versus small studies: an empirical assessment. *Lancet*. 2003;361: 567–571.
14. Ioannidis JP, Ntzani EE, Trikalinos TA, et al. Replication validity of genetic association studies. *Nat Genet*. 2001;29:306–309.
15. Ioannidis JP. Common genetic variants for breast cancer: 32 largely refuted candidates and larger prospects. *J Natl Cancer Inst*. 2006;98: 1350–1353.
16. Göring HH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet*. 2001;69:1357–1369.
17. Allison DB, Fernandez JR, Heo M, et al. Bias in estimates of quantitative-trait-locus effect in genome scans: demonstration of the phenomenon and a method-of-moments procedure for reducing bias. *Am J Hum Genet*. 2002;70:575–585.
18. Siegmund D. Upward bias in estimation of genetic effects. *Am J Hum Genet*. 2002;71:1183–1188.
19. Beavis WD. QTL analysis: power, precision, and accuracy. In: Paterson AH, ed. *Molecular Dissection Of Complex Traits*. Boca Raton, FL: CRC Press; 1998:145–173.
20. Kearsey MJ, Farquhar AG. QTL analysis in plants; where are we now? *Heredity*. 1998;80:137–142.
21. Melchinger AE, Utz HF, Schon CC. Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics*. 1998;149:383–403.
22. Utz HF, Melchinger AE, Schon CC. Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics*. 2000; 154:1839–1849.
23. Zollner S, Pritchard JK. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am J Hum Genet*. 2007;80:605–615.
24. Garner C. Upward bias in odds ratio estimates from genome-wide association studies. *Genet Epidemiol*. 2007;31:288–295.
25. Jennions MD, Moeller AP. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. *Proc R Soc Lond B Biol Sci*. 2002;269:43–48.
26. Leimu R, Koricheva J. Cumulative meta-analysis: a new tool for detection of temporal trends and publication bias in ecology. *Proc R Soc London B Biol Sci*. 2004;271:1961–1966.
27. Tversky A, Kahneman D. Belief in the law of small numbers. *Psychol Bull*. 1971;2:105–110.
28. Strube MJ. SNOOP: a program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behav Res Methods*. 2006;38:24–27.
29. Steyerberg EW, Eijkemans MJ, Harrell FE Jr, et al. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med*. 2000;19:1059–1079.
30. Steyerberg EW, Harrell FE Jr, Borsboom GJ, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*. 2001;54:774–781.
31. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer*. 1994;69:979–985.
32. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol*. 1999;52:935–942.
33. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol*. 1993;138:923–936.
34. Chatfield C. Model uncertainty, data mining and statistical inference. *J R Statist Soc Ser A*. 1995;158:419–466.

35. Ioannidis JP. Is molecular profiling ready for use in clinical decision making? *Oncologist*. 2007;12:301–311.
36. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415: 530–536.
37. Buyse M, Loi S, van't Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst.* 2006;98:1183–1192.
38. Ioannidis JP. Non-replication and inconsistency in the genome-wide association setting. *Hum Hered*. 2007;64:203–213.
39. Ioannidis JP, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am J Epidemiol*. 2006;164:609–614.
40. Khoury MJ, Little J, Gwinn M, et al. On the synthesis and interpretation of consistent but weak gene-disease associations in the era of genome-wide association studies. *Int J Epidemiol*. 2007;36:439–445.
41. 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.
42. Emerson JD. Combining estimates of the odds ratio: the state of the art. *Stat Methods Med Res*. 1994;3:157–178.
43. Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ*. 2005;330:753.
44. Chan AW, Krleza-Jerić K, Schmid I, et al. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ*. 2004;171:735–740.
45. Chan AW, Hróbjartsson A, Haahr MT, et al. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*. 2004;291:2457–2465.
46. Kavvoura FK, Liberopoulos G, Ioannidis JP. Selection in reported epidemiological risks: an empirical assessment. *PLoS Med*. 2007;4: e79.
47. Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. *Clin Trials*. 2007;4:245–253.
48. Pan Z, Trikalinos TA, Kavvoura FK, et al. Local literature bias in genetic epidemiology: an empirical evaluation of the Chinese literature. *PLoS Med*. 2005;2:e334.
49. Kavvoura FK, McQueen M, Khoury MJ, et al. Evaluation of the potential excess of statistically significant findings in reported genetic association studies: application to Alzheimer's disease. *Am J Epidemiol.* In press.
50. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?." *Lancet*. 2005;365:82–93.
51. Ioannidis JP. Limitations are not properly acknowledged in the scientific literature. *J Clin Epidemiol*. 2007;60:324–329.
52. Clarke M, Alderson P, Chalmers I. Discussion sections in reports of controlled trials published in general medical journals. *JAMA*. 2002; 287:2799–2801.
53. Ioannidis JP, Polyzos NP, Trikalinos TA. Selective discussion and transparency in microarray research findings for cancer outcomes. *Eur J Cancer*. 2007;43:1999–2010.
54. Contopoulos-Ioannidis DG, Ntzani E, Ioannidis JP. Translation of highly promising basic science research into clinical applications. *Am J Med*. 2003;114:477–484.
55. Chan AW, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet*. 2005;365:1159–1162.
56. Maddock JE, Rossi JS. Statistical power of articles published in three health psychology-related journals. *Health Psychol*. 2001;20:76–78.
57. Williams JL, Hathaway CA, Kloster KL, et al. Low power, type II errors, and other statistical problems in recent cardiovascular research. *Am J Physiol*. 1997;273:H487–H493.
58. Weaver CS, Leonardi-Bee J, Bath-Hextall FJ, et al. Sample size calculations in acute stroke trials: a systematic review of their reporting, characteristics, and relationship with outcome. *Stroke*. 2004;35: 1216–1224.
59. Keen HI, Pile K, Hill CL. The prevalence of underpowered randomized clinical trials in rheumatology. *J Rheumatol*. 2005;32:2083–2088.
60. Maxwell SE. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychol Methods*. 2004; 9:147–163.
61. Armstrong BG. The effects of measurement errors on relative risk regressions. *Am J Epidemiol*. 1990;132:1176–1184.
62. Kristensen P. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology*. 1992;3:210–215.
63. Flanders WD, Drews CD, Kosinski AS. Methodology to correct for differential misclassification. *Epidemiology*. 1995;6:152–156.
64. García-Closas M, Thompson WD, Robins JM. Differential misclassification and the assessment of gene-environment interactions in case-control studies. *Am J Epidemiol*. 1998;147:426–433.
65. Garcia-Closas M, Rothman N, Lubin J. Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. *Cancer Epidemiol Biomarkers Prev*. 1999;8:1043–1050.
66. Wong MY, Day NE, Luan JA, et al. Estimation of magnitude in gene-environment interactions in the presence of measurement error. *Stat Med*. 2004;23:987–998.
67. Zhang L, Mukherjee B, Ghosh M, et al. Accounting for error due to misclassification of exposures in case-control studies of gene-environment interaction. *Stat Med*. 2008;27:2756–2783.
68. Tung L, Gordon D, Finch SJ. The impact of genotype misclassification errors on the power to detect a gene-environment interaction using cox proportional hazards modeling. *Hum Hered*. 2007;63:101–110.
69. Lexchin J, Bero LA, Djulbegovic B, et al. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ*. 2003;326:1167–1170.
70. Vandenbroucke J. Observational research, randomised trials and two views of medical science. *PLoS Med*. 2008;5:e67.
71. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes [erratum in: *N Engl J Med* 2007;357:100]. *N Engl J Med*. 2007;356:2457–2471.
72. Singh S, Loke YK, Furberg CD. Long-term risk of cardiovascular events with rosiglitazone: a meta-analysis. *JAMA*. 2007;298:1189–1195.
73. Hernandez AV, Walker E, Ioannidis JPA, et al. Challenges in meta-analysis of randomized clinical trials for rare harmful cardiovascular events: the case of rosiglitazone. *Am Heart J*. 2008;156:22–30.
74. Papanikolaou PN, Christidi GD, Ioannidis JP. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *CMAJ*. 2006;174:635–641.
75. Ioannidis JP, Mulrow CD, Goodman SN. Adverse events: the more you search, the more you find. *Ann Intern Med*. 2006;144:298–300.
76. Tatsioni A, Bonitsis NG, Ioannidis JP. Persistence of contradicted claims in the literature. *JAMA*. 2007;298:2517–2526.
77. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361–387.
78. Todd JA. Statistical false positive or true disease pathway? *Nat Genet*. 2006;38:731–733.
79. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med*. 2001;134:663–694.
80. Ioannidis JP, Evans SJ, Gøtzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med*. 2004;141:781–788.
81. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Epidemiology*. 2007;18:805–835.
82. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *AJR Am J Roentgenol*. 2003;181:51–55.
83. Laine C, Horton R, DeAngelis CD, et al. Clinical trial registration: looking back and moving ahead. *Lancet*. 2007;369:1909–1911.
84. De Angelis C, Drazen JM, Frizelle FA, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *Lancet*. 2004;364:911–912.
85. McGuire AL, Hamilton JA, Lunstroth R, et al. DNA data sharing: research participants' perspectives. *Genet Med*. 2008;10:46–53.
86. Chokshi DA, Parker M, Kwiatkowski DP. Data sharing and intellectual property in a genomic epidemiology network: policies for large-scale research collaboration. *Bull World Health Organ*. 2006;84:382–387.

87. GAIN Collaborative Research Group, Manolio TA, Rodriguez LL, Brooks L, et al. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet*. 2007;39:1045–1051.
88. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007;39:1181–1186.
89. Smith GD, Lawlor DA, Harbord R, et al. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Med*. 2007;4:e352.
90. Seminara D, Khoury MJ, O'Brien TR, et al. The emergence of networks in human genome epidemiology: challenges and opportunities. *Epidemiology*. 2007;18:1–8.
91. Goodman SN. Systematic reviews are not biased by results from trials stopped early for benefit. *J Clin Epidemiol*. 2008;61:95–96; author reply 96–98.
92. Bassler D, Ferreira-Gonzalez I, Briel M, et al. Systematic reviewers neglect bias that results from trials stopped early for benefit. *J Clin Epidemiol*. 2007;60:869–873.
93. Goodman SN. Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Ann Intern Med*. 2007;146:882–887.
94. Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. *BMJ*. 1998;316:140–144.
95. Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*. 2007;335:914–916.
96. Wetterslev J, Thorlund K, Brok J, et al. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *J Clin Epidemiol*. 2008;61:64–75.
97. Jeffreys H. *Theory of Probability*. 3rd ed. Oxford: Oxford University Press; 1961.
98. Goodman SN. Toward evidence-based medical statistics. Part 1: The P value fallacy. *Ann Intern Med*. 1999;130:995–1004.
99. Goodman SN. Toward evidence-based medical statistics. Part 2: The Bayes factor. *Ann Intern Med*. 1999;130:1005–1013.
100. Cornfield J. The Bayesian outlook and its application. *Biometrics*. 1969;25:617–657.
101. Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. *BMJ*. 1996;313:603–607.
102. Hughes MD. Reporting Bayesian analyses of clinical trials. *Stat Med*. 1993;12:1651–1263.
103. Ioannidis JP. Effect of statistical significance on the credibility of observational associations. *Am J Epidemiol*. 2008 Jul 8. [Epub ahead of print].