American Journal of Epidemiology
Vol. 168, No. 4
DOI: 10.1093/aje/kwn156
Advance Access publication July 8, 2008

Practice of Epidemiology

Effect of Formal Statistical Significance on the Credibility of Observational Associations

John P. A. Ioannidis[1,2,3]

[1] Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece.
[2] Biomedical Research Institute, Foundation for Research and Technology-Hellas, Ioannina, Greece.
[3] Department of Medicine, Tufts University School of Medicine, Boston, MA.

Received for publication May 22, 2007; accepted for publication January 4, 2008.

The author evaluated the implications of nominal statistical significance for changing the credibility of null versus alternative hypotheses across a large number of observational associations for which formal statistical significance ($p < 0.05$) was claimed. Calculation of the Bayes factor (B) under different assumptions was performed on 272 observational associations published in 2004–2005 and a data set of 50 meta-analyses on gene-disease associations (752 studies) for which statistically significant associations had been claimed ($p < 0.05$). Depending on the formulation of the prior, statistically significant results offered less than strong support to the credibility ($B > 0.10$) for 54–77% of the 272 epidemiologic associations for diverse risk factors and 44–70% of the 50 associations from genetic meta-analyses. Sometimes nominally statistically significant results even decreased the credibility of the probed association in comparison with what was thought before the study was conducted. Five of six meta-analyses with less than substantial support ($B > 0.032$) lost their nominal statistical significance in a subsequent (more recent) meta-analysis, while this did not occur in any of seven meta-analyses with decisive support ($B < 0.01$). In these large data sets of observational associations, formal statistical significance alone failed to increase much the credibility of many postulated associations. Bayes factors may be used routinely to interpret “significant” associations.

Bayes theorem; empirical research; epidemiologic methods; meta-analysis; observation; statistics

Abbreviation: RR, relative risk.

Editor’s note: *An invited commentary on this article appears on page 384, and the author’s response appears on page 389.*

Several sets of high-profile research findings from observational studies have been contradicted in the last few years by randomized trials (1, 2). Many associations identified in observational epidemiologic investigations may reflect false-positive findings (3). This applies to both traditional and molecular epidemiology (3–5). Multiplicity of comparisons, massive testing of hypotheses that have a low likelihood of being true, and various biases are invoked to explain false-positive findings.

In most observational studies, investigators use frequentist approaches to reject null hypotheses of no association. This may sometimes lead to misleading inferences. Several Bayesian methods have been proposed to measure the credibility of alternative versus null hypotheses (6–9). These methods have a well-established theoretical background. However, they have been applied in relatively few studies with real data (10–12).

Correspondence to Dr. John P. A. Ioannidis, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece (e-mail: jioannid@cc.uoi.gr).

With hypothesis-testing and nominal statistical significance having been entrenched in the interpretation of associations for decades, the paradigm may not change unless the shortcomings are shown with large-scale evidence. It is useful to examine a Bayesian interpretation of the results of a large sample of observational studies. One may ask: How much does a nominally statistically significant result increase the credibility of a postulated association? Here, this question was addressed in analyses of two large data sets of observational associations.

MATERIALS AND METHODS

Theoretical framework

The poststudy odds that a probed association is true can be estimated as the ratio of the prestudy odds over the Bayes factor B conferred by the study data. Prestudy odds depend on the specific research field and potentially other external evidence; therefore, deciding on a specific value carries some unavoidable subjectivity. All analyses presented here focus exclusively on B. $B < 1$ means that the study increases the odds that some probed association exists compared with what we thought before the study. $B > 1$ means that the study decreases the odds that some association exists compared with what we thought before the study. $B = 1$ means that the study does not change the odds that the probed association exists.

To allow closed form analysis in performing calculations, let us consider that in each study, the observed effect size (here, the relative risk) can be represented by a normal likelihood. This assumption is typically reasonable for studies that are not small, as in the studies included in the empirical evaluations presented here. The observed effect size is considered to be an estimate of the true effect θ with a certain variance

$$y_m \sim N[\theta, \mathrm{var}(\theta)]$$

or, equivalently,

$$y_m \sim N[\theta, \sigma^2/m], \tag{1}$$

with m being the effective number of events in the study.

The prior can be specified for convenience as a ‘‘spike and smear,’’ where a spike of $p(0)$ is placed at the null H_0 and the remaining $1 - p(0)$ is distributed under the alternative H_1 as

$$\theta|H_1 \sim N[0, \sigma^2/n_0]. \tag{2}$$

From equations 1 and 2, it follows that

$$y_m|H_1 \sim N[0, \sigma^2((1/m) + (1/n_0))]. \tag{3}$$

B is the ratio of $p(y_m|H_0)/p(y_m|H_1)$. Based on the above considerations of normality, computationally the Bayes factor is given by equation 4:

$$\begin{aligned} B &= [[\sqrt{m}/(\sigma\sqrt{2\pi})]\exp[-my_m^2/2\sigma^2]]/[[\sqrt{(mn_0/(m+n_0))}/\\ &\quad \sigma\sqrt{2\pi})]\exp[-mn_0y_m^2/(m+n_0)/2\sigma^2]]\\ &= \sqrt{(m+n_0)/n_0}\exp[(-y_m^2 m/\sigma^2)/(2(m+n_0)/m)]\\ &= \sqrt{(1+(m/n_0)}\exp[(-z_m^2)/(2(1+(n_0/m)], \end{aligned} \tag{4}$$

where $z_m = y_m\sqrt{m}/\sigma$ is the standardized test statistic for the null hypothesis. Let us set, following the method of Cornfield (9),

$$n_0 = 2\sigma^2/(\pi\theta_A^2) = 2m\,\mathrm{var}(\theta)/(\pi\theta_A^2), \tag{5}$$

where θ_A is the expected value of the effect under the alternative hypothesis, if there is an effect in the positive direction (relative risk (RR) $>$ 1.00). An advantage of this approach is that it allows for the ratio of m/n_0 to be expressed as a simple function of the observed $\mathrm{var}(\theta)$ and a specified alternative effect θ_A. Specifically, from equation 5, it follows that

$$m/n_0 = \pi\theta_A^2/2\,\mathrm{var}(\theta). \tag{6}$$

Sensitivity analyses can be used to examine whether conclusions are affected appreciably, depending on what effect is assumed under the alternative, that is, under different priors.

A Bayesian framework can also help an investigator examine the level of support for the null hypothesis versus all competing alternative hypotheses (e.g., as described by Goodman (8)). This has the disadvantage that large studies with minimal effects may seem to have strong support for the ‘‘generic’’ alternative hypothesis, although the effect is practically ‘‘null’’ (e.g., an odds ratio of 1.04 with very tight confidence intervals) (6). Therefore, the analyses presented here address each time a specific prior for the alternative: The overall expected value of the effect under the alternative is 0 (i.e., the same chance that an effect is in one direction or the other), while the expected value of the effect under the alternative in the positive direction is θ_A.

Databases of observational associations

The first database contained 389 studies published in 2004–2005 that presented relative risk estimates for diverse continuous risk factors in epidemiologic studies. The search strategy and eligibility criteria have been described in detail previously (13). In brief, the studies pertained to continuous risk factors that had been examined in contrasts using medians, tertiles, quartiles, or quintiles. In each study, the first reported eligible relative risk had been recorded. For the current analysis, the data set was further restricted to studies in which the relative risk was nominally statistically significant at the $p = 0.05$ threshold without any adjustments for multiple comparisons and the 95 percent confidence interval was also available, so as to calculate the variance of the natural logarithm of the relative risk under normality assumptions. Overall, 272 studies fulfilled these additional criteria.

The second database contained 50 meta-analyses of gene-disease associations (obtained from a total of 752 combined studies) published through February 2004 for which investigators had claimed a nominally statistically significant association ($p < 0.05$) between a common genetic variant and a disease phenotype by random-effects calculations that account for between-study heterogeneity (14). Search strategy, eligibility criteria, and selection algorithms for the genetic contrasts are described in detail elsewhere (15–17).

Meta-analyses may be considered as single studies for estimation of their B.

Estimation of Bayes factors

For each association, estimation of B used the observed effect size and the variance thereof and different assumptions for the prior. B does not depend on $p(0)$, but it depends on the exact shape of the prior for the alternative hypothesis. Different values for θ_A were used in the analyses. These values reflect the expected magnitude of epidemiologic risks. Most relative risks in the current era are anticipated to be relatively small (18). In population genetics, in particular, small and very small effect sizes (RR = 1.1–1.6) are considered typical (17, 19). Most of the epidemiologic investigations analyzed did not include any sample size or power calculations. Therefore, it is unknown what the investigators themselves would specify up front as the alternative hypothesis for effect size. A wide range was considered, including θ_A of 0.1, 0.25, 0.5, 1.0, 1.5, and 2.0, which correspond respectively to relative risks of 1.11, 1.28, 1.65, 2.72, 4.48, and 7.39. Typical effect sizes may vary in different fields of observational research. Therefore, an additional analysis considered the median relative risk across all studies in the database that addressed the same type of risk factor after coining of all relative risks so that they were greater than or equal to 1.00 (i.e., relative risks less than 1.00 were inversed) for consistency. Categorization of risk factor types was performed as previously described (13).

The presented categories of B follow the traditional Jeffreys calibration (20): $B = 0.32$–1.00, ‘‘not worth more than a bare mention’’; $B = 0.10$–0.32, ‘‘substantial support’’ for the alternative hypothesis; $B = 0.032$–0.10, ‘‘strong support’’ for the alternative hypothesis; $B = 0.010$–0.032, ‘‘very strong support’’ for the alternative hypothesis; and $B < 0.010$, ‘‘decisive support’’ for the alternative hypothesis.

When B exceeds 1.00, the credibility of the probed association is worse after the study as compared with what it had been prior to the study. At first reading, it sounds paradoxical that a formally statistically significant result may sometimes *decrease* the credibility of a probed association. However, this makes perfect sense as a consequence of the Lindley paradox (21). When the observed formally statistically significant effect is small in magnitude (close to the null), the alternative hypothesis may be less likely than the null.

Evolution of evidence on associations

Evidence is not static but evolves over time (22, 23). An association may be revisited by the same investigator or other investigators. More recent studies may be included in a cumulative update, and eligibility criteria or genetic models may also be revisited. One may then ask whether associations without substantial support are more likely to lose nominal statistical significance in subsequently published meta-analyses compared with meta-analyses that had decisive support.

This evaluation focused on genetic meta-analyses in which the Bayes factor (for θ_A equal to the median relative risk in the field of genetic association meta-analyses) suggested either no substantial support or decisive support for the association. PubMed searches (March 2004–February 2007) were made to identify any subsequent meta-analyses on the same association that had been published at least one calendar year after the first. Ideally the same genetic contrast and eligibility criteria were preferred, if available; differences were allowed otherwise. When several more recent meta-analyses were identified, the most recent one was selected. For each selected more recent meta-analysis, the odds ratio and 95 percent confidence interval were extracted for the same genetic contrast as was made in the earlier meta-analysis. If information on the same exact genetic contrast was not provided, the odds ratio and 95 percent confidence interval were selected for the primary contrast reported in the more recent meta-analysis.

TABLE 1. Characteristics of studies included in an analysis of calculation of the Bayes factor under different assumptions

	Diverse epidemiologic studies (n = 272)	Genetic meta-analyses (n = 50)
Study design and metric		
Case-control, odds ratio	51	0
Cohort, odds ratio	65	0
Cohort, relative risk	156	0
Meta-analysis, odds ratio	0	50
Median relative risk*	1.96 (1.54–2.66)†	1.43 (1.29–1.65)
Median standard deviation	0.23 (0.14–0.34)	0.11 (0.08–0.17)

* Coined so as to be consistently greater than or equal to 1.00.
† Numbers in parentheses, interquartile range.

RESULTS

Characteristics of evaluated associations

The first data set included diverse types of risk factors: biologic markers (n = 107, median RR = 2.30), dietary factors (n = 74, median RR = 1.59), psychosocial factors (n = 26, median RR = 1.80), body characteristics (n = 21, median RR = 1.90), toxic exposures (n = 6, median RR = n = 2.97), physical activity (n = 6, median RR = 1.88), and various other factors (n = 32, median RR = 2.26) (table 1). Expectedly, the median relative risk in the 50 genetic meta-analyses was smaller (RR = 1.43) than in the other fields. The standard deviation of the observed effects was, on average, smaller in the 50 genetic meta-analyses than in single studies of other fields.

Bayes factors for single epidemiologic studies

The statistically significant results did not offer any substantial support to the probed association in many studies (28–62 percent, depending on the specification of the prior distribution) (table 2). The support was less than strong in 54–77 percent of the studies. Only 9–25 percent of the

TABLE 2. Categorization of Bayes factors (*B*) for 272 diverse epidemiologic studies under various prior assumptions

Estimated Bayes factor category	Assumption for the effect θ_A under the alternative hypothesis, given a positive effect													
	0.1		0.25		0.5		1		1.5		2		Field median*	
	No.†	%†	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
>1.00 (credibility worsened)	0		0		0		13	5	31	11	47	17	1	0
0.32–1.00 (bare mention)	169	62	97	36	77	28	84	31	80	29	80	29	76	28
0.10–0.32 (substantial support)	42	15	71	26	70	26	61	22	54	20	39	14	73	27
0.032–0.10 (strong support)	28	10	28	10	35	13	25	9	21	8	29	11	31	11
0.010–0.032 (very strong support)	9	3	21	8	25	9	24	9	25	9	19	7	23	9
<0.010 (decisive)	24	9	55	20	65	24	65	24	61	22	58	21	68	25

* Median effect size for studies in the same field.

† Number and percentage of studies.

studies yielded decisive support. The proportion of results offering less than strong support was very similar for a wide range of alternative priors (RRs = 1.28–7.39). Support for the observed associations was at its weakest when the alternative assumed a very small effect (RR = 1.11). The proportion of results that had less than strong support did not have obvious differences across different fields (e.g., 55–78 percent for dietary risk factors and 53–87 percent for biologic markers) or for different types of designs (e.g., 49–86 percent for case-control studies with odds ratios, 54–75 percent for cohort studies with relative risks, and 55–77 percent for cross-sectional cohort studies with odds ratios).

Figure 1 and figure 2 show *B* as a function of the observed *p* value for θ_A of 0.50 and 1.50, respectively (i.e., relative risks of 1.65 and 4.48). None of the 122 associations with *p* values of 0.01–0.05 had strong support ($B < 0.10$). Even among the 150 associations with $p < 0.01$, less than strong support was seen in 25 and 43 associations, respectively, depending on θ_A. Among the 91 associations with $p < 0.001$, 26 and 30 associations, respectively, did not have

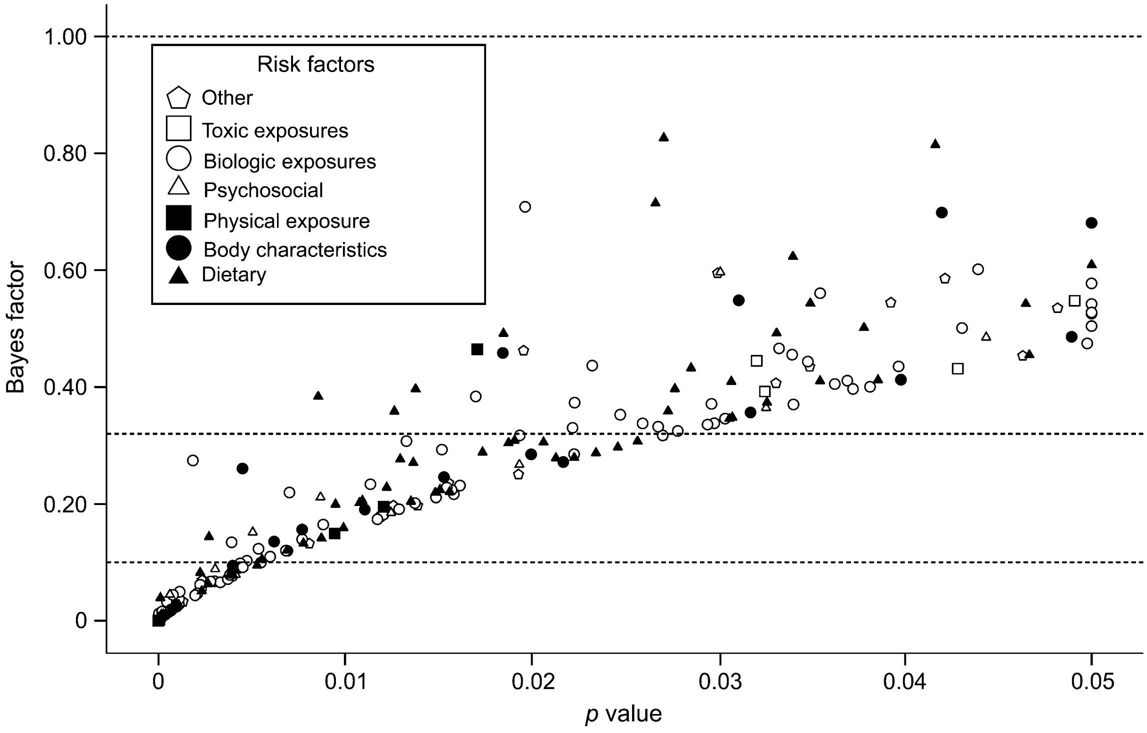

FIGURE 1. Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed *p* value in each study. Shown are calculations assuming θ_A of 0.50 (relative risk = 1.65). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.

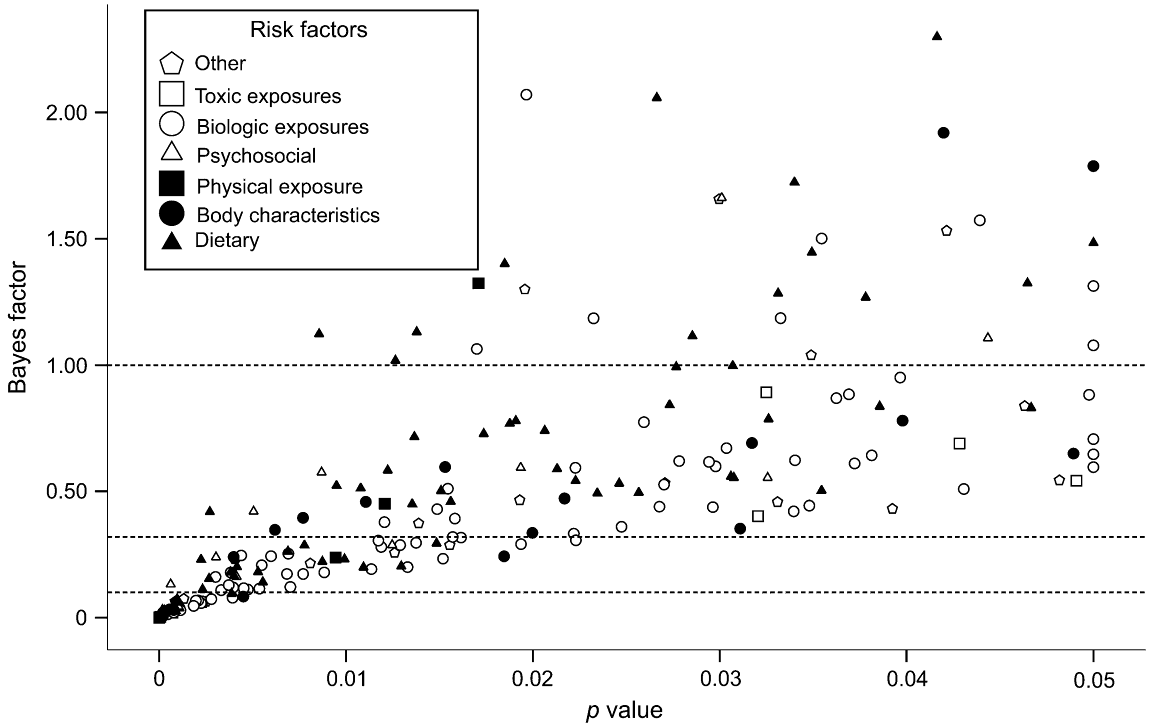

FIGURE 2. Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed *p* value in each study. Shown are calculations assuming θ_A of 1.50 (relative risk = 4.48). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.

decisive support. All associations with $p < 0.00001$ had decisive support, but only 37 associations passed this threshold.

For $\theta_A = 1.50$, in 11 percent of studies the credibility of the probed association actually decreased despite nominally statistically significant results.

Bayes factors for meta-analyses of genetic associations

The statistically significant results did not offer any substantial support to the probed association in 18–48 percent of the meta-analyses, depending on the prior (table 3). The support was less than strong in 44–70 percent of the meta-analyses. Only 12–22 percent of the meta-analyses had decisive support. However, for genetic associations, large effects θ_A are not very reasonable alternatives. After excluding θ_A values of 1.5 and 2.0 (corresponding to relative risks of 4.48 and 7.39, respectively), the proportion of results that did not offer substantial support to the probed association ranged from 18 percent to 32 percent, and the support was less than strong in 44–62 percent of the meta-analyses; again, 12–22 percent of the meta-analyses offered decisive support. Support for the observed associations was at its weakest when large effects were assumed under the alternative (RR = 7.39).

Figure 3 shows B as a function of the observed p value, for θ_A equal to the median relative risk across the 50 associations (RR = 1.43). None of the 17 genetic associations with p values of 0.01–0.05 had strong support ($B < 0.10$), while even among the 33 associations with $p < 0.01$, less than strong support was seen for six associations. Among the 15 associations with $p < 0.001$, four did not have decisive support. Only four associations had $p < 0.00001$, and all of them had decisive support.

One nominally significant meta-analysis practically did not change at all the credibility of the association. As described below, an update of that meta-analysis resulted in loss of statistical significance (24).

Evolution of evidence

Twelve meta-analyses offered less than substantial support and 11 meta-analyses offered decisive support for the probed association. Six meta-analyses from the former group (25–30) and seven from the latter group (31–37) had been followed by subsequent meta-analyses (table 4) (24, 38–47). Different contrasts were selected in six subsequent meta-analyses, and eligibility criteria were considerably wider in two meta-analyses and considerably more restricted in another two than in their earlier published counterparts (table 4).

In five of the six meta-analyses that offered less than substantial support, evolution of the evidence resulted in a nonstatistically significant summary effect. Even in the one

TABLE 3. Categorization of Bayes factors (*B*) for 50 genetic meta-analyses under various prior assumptions

Estimated Bayes factor category	Assumption for the effect $\theta_A$under the alternative hypothesis, given a positive effect													
	0.1		0.25		0.5		1		1.5		2		Field median*	
	No.†	%†	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
>1.00 (credibility worsened)	0		0		1	2	6	12	12	24	13	26	0	
0.32–1.00 (bare mention)	16	32	9	18	13	26	10	20	8	16	11	22	12	24
0.10–0.32 (substantial support)	15	30	13	26	10	20	13	26	11	22	11	22	11	22
0.032–0.10 (strong support)	11	22	11	22	10	20	13	26	11	22	11	22	11	22
0.010–0.032 (very strong support)	2	4	6	12	6	12	3	6	4	8	4	8	6	12
<0.010 (decisive)	6	12	11	22	10	20	9	18	8	16	6	12	11	22

* Median effect size for all 50 meta-analyses of genetic associations.

† Number and percentage of studies.

association for which formal statistical significance was maintained, the Bayes factor from the newer meta-analysis barely crossed the threshold of offering substantial support ($B = 0.28$). Some changes in formal statistical significance occurred very rapidly. For example, in a meta-analysis of SerSer homozygosity for the Ser9Gly polymorphism of the dopamine receptor D3 (*DRD3*) gene across 40 studies with 8,761 subjects, Jonsson et al. (29) found a nominally statistically significant summary effect (odds ratio = 1.10, 95 percent confidence interval: 1.01, 1.21; $p = 0.031$) for schizophrenia. The same team revisited this association a year later after including four additional studies, for a total sample size of 11,066 subjects, and the association was no longer nominally significant (24).

Conversely, results remained formally statistically significant in all seven associations for which the original meta-analysis had offered decisive support. However, the Bayes factor from the new meta-analysis remained at the level of decisive support in only three cases, while support became weaker in the other four. For the postulated association between the apolipoprotein E (*APOE*) gene and ischemic stroke, the more recent meta-analysis had a borderline significant effect that provided practically no support for the association ($B = 0.92$).

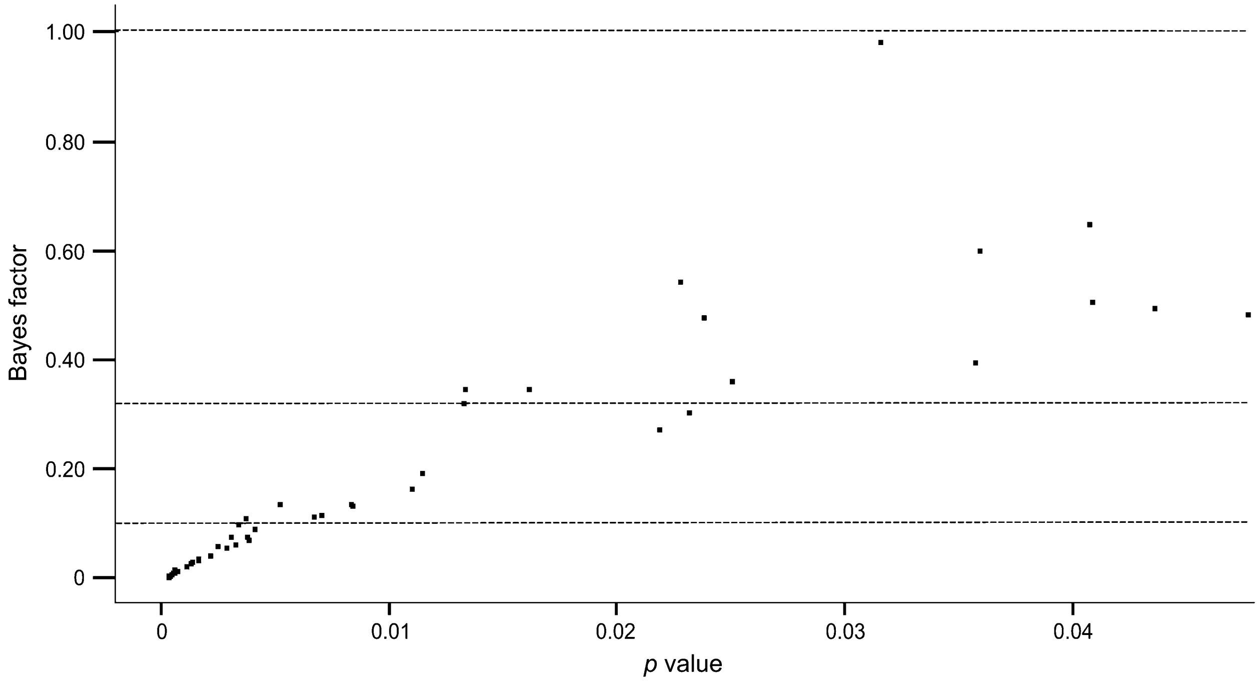

FIGURE 3. Estimated Bayes factors for 50 meta-analyses of genetic associations with formally statistically significant results. The Bayes factor is plotted against the observed *p* value in each meta-analysis. Calculations assume θ_A equal to the median relative risk observed in the 50 genetic associations (relative risk = 1.44). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.

TABLE 4. Comparison of subsequent versus earlier meta-analyses in genetic associations whose original meta-analysis offered less than substantial support and in those where it offered decisive support*

Earlier meta-analysis (ref. no.)	Gene (variant); contrast	Disease	Earlier meta-analysis		Subsequent meta-analysis		Subsequent meta-analysis (ref. no.)	Differences in contrast/eligibility	Bayes factor†
			Effect	95% CI‡	Effect	95% CI			
No substantial support§									
Boekholdt et al. (25)	*FGB*‡/*FGB* promoter (455G/A); AA vs. GG	Myocardial infarction	1.46	1.00, 2.13	1.12	0.90, 1.41	Smith et al. (38)	Allele/wider	0.48/NP‡
Maraganore et al. (26)	*UCH-L1*‡ (S18Y); S/S vs. other	Parkinson's disease	1.20	1.02, 1.40	0.96	0.86, 1.08	Healy et al. (39)	None/none	0.48/NP
Kosmas et al. (27)	*MTHFR*‡ (677C/T); TT vs. other	Preeclampsia	1.21	1.01, 1.45	1.01	0.79, 1.29	Lin et al. (40)	None/none	0.60/NP
Burzotta et al. (28)	*F2*‡ (20210G/A); other vs. GG	Myocardial infarction	1.32	1.01, 1.72	1.25	1.05, 1.50	Ye et al. (41)	Allele	0.51/0.28
Jonsson et al. (29)	*DRD3*‡ (Ser9Gly) SerSer vs. other	Schizophrenia	1.10	1.01, 1.21	1.05	0.97, 1.13	Jonsson et al. (24)	None/none	0.98/NP
Combarros et al. (30)	*IL1A*‡ (-889); 2/2 vs. other	Alzheimer's disease	2.35	1.03, 5.37	1.08	0.98, 1.18	Bertram et al. (42)	Allele/wider	0.49/NP
Decisive support¶									
Marcus et al. (31)	*NAT2*‡ (acetylation); slow/slow vs. other	Bladder cancer	1.43	1.20, 1.71	1.4	1.2, 1.6	Garcia-Closas et al. (43)	None/none	0.003/0.0002
McCarron et al. (32)	*APOE*‡ (epsilon 2/3/4); allele 4 vs. other	Ischemic cerebrovascular disease	1.69	1.37, 2.09	1.11	1.01, 1.22	Sudlow et al. (44)	Carriers/none	<0.0001/0.92
Golbe et al. (33)	*MAPT*‡ (allele A0); allele A0 vs. other	Parkinson's disease	1.52	1.22, 1.90	1.71	1.25, 2.36	Zhang et al. (45)	Haplotype H1H1/none	0.007/0.02
Johns et al. (34)	*GSTM1*‡ (gene deletion); null/null vs. other	Bladder cancer	1.54	1.27, 1.86	1.5	1.3, 1.6	Garcia-Closas et al. (43)	None/none	0.0003/<0.0001
Kosmas et al. (35)	Factor V (Leiden mutation); allele	Preeclampsia	2.22	1.46, 3.38	1.81	1.14, 2.87	Lin et al. (40)	Carriers/none	0.008/0.18
Sethi et al. (36)	*AGT*‡ (M235T); TT vs. MM	Essential hypertension	1.35	1.18, 1.55	1.30	1.10, 1.54	Mondry et al. (46)	None/restricted	0.0009/0.06
Hashibe et al. (37)	*GSTM1* (gene deletion); null/null vs. other	Head and neck cancers	1.32	1.14, 1.53	1.50	1.21, 1.87	Tripathy et al. (47)	None/restricted	0.009/0.008

* The study by Mondry et al. (46) was limited to populations of Caucasian descent, and Tripathy et al. (47) also had more restricted eligibility criteria, in comparison with their respective earlier meta-analyses. Conversely, Smith et al. (38) addressed all coronary artery disease and Bertram et al. (42) addressed all Alzheimer's disease, while Combarros et al. (30) addressed early onset in the presented estimate.

† The pairs show the Bayes factors based on the earlier meta-analysis and the subsequently published meta-analysis. An effect designated ''not pertinent'' is one that is not formally significant.

‡ CI, confidence interval; FGB, fibrinogen beta chain; NP, not pertinent; UCH-L1, ubiquitin carboxyl-terminal esterase L1; MTHFR, methylenetetrahydrofolate reductase; F2, coagulation factor II; DRD3, dopamine receptor D3; IL1A, interleukin-1A; NAT2, *N*-acetyltransferase 2; APOE, apolipoprotein E; MAPT, microtubule-associated protein tau; GSTM1, glutathione *S*-transferase M1; AGT, angiotensinogen.

§ No substantial support for an effect in the earlier meta-analysis.

¶ Decisive support for an effect in the earlier meta-analysis.

DISCUSSION

Evaluation of a large number of observational associations demonstrates that most of the formally statistically significant results in this extensive literature did not convey strong support for the probed associations. This conclusion was relatively robust to different prior assumptions. Statistically significant results occasionally even decreased the level of support for an association in comparison with what was thought before the study was conducted. Moreover, with one exception, all examined meta-analyses with less than substantial support lost their formal significance when subsequent meta-analyses on the same association were published just 1–5 years later. Meta-analyses with decisive support did not lose formal statistical significance, but they often lost the decisiveness of the support in subsequent meta-analyses.

These findings suggest that a very cautious interpretation of nominally statistically significant findings is due in observational research. Most statistically significant findings do not markedly improve the credibility of the associations they probe. This applies to both single studies and meta-analyses where several studies may have already ''replicated'' a postulated association (48). One may thus understand why several observational associations are refuted upon further testing by either observational or randomized designs (1–3, 15, 49).

The empirical evaluation showed that none of the associations with p values between 0.01 and 0.05 had strong support. These associations accounted for almost half of the ''statistically significant'' associations. Using a more stringent threshold of statistical significance would dismiss many spuriously statistically significant claims that lacked strong support, but many associations with strong or even decisive support might also be dismissed, especially if the threshold were set too low. Associations with similar inferences based on statistical significance may have different inferences based on Bayes factors. Therefore, a single shift in the threshold of claiming statistical significance is unlikely to solve the problem. Bayes factors should be adopted routinely in interpreting observational results.

Conversely, one potential disadvantage of the Bayesian approach is the dependence on the specification of the prior. However, if Bayesian approaches are more widely adopted, it should be readily feasible to adopt a standard set of sensitivity analyses regarding prior specification, and this would allow comparability of results across studies. Moreover, the qualitative inferences usually remain quite robust under different prior assumptions (50). Furthermore, in many cases where a large body of research already exists in a specific field, the plausible alternatives could be potentially limited to a relatively narrow range. For example, based on a large body of studies on genetic associations for common variants, it is currently clear that large odds ratios are very uncommon; thus, these could be safely excluded from the typical consideration of alternatives. Conversely, large odds ratios may need to be considered for rare variants. Finally, as Berger et al. pointed out (51, 52), in fact frequentist methods may converge towards the Bayesian, if properly modeled as a conditional frequentist approach.

Through the use of a sample of meta-analyses, the present study shows empirically that statistically significant associations that did not have substantial support almost always also lost their nominal statistical significance when subsequent meta-analyses were performed with updated evidence. This sequential evaluation of evidence was limited to few topics and only one field of epidemiologic investigation. Further prospective studies should be encouraged to test the independent replication and validation of associations for which inferences have been qualitatively very different with the frequentist versus Bayesian approaches.

Some additional caveats must be discussed. First, the present evaluation focused on B without trying to estimate the poststudy odds for each association. Poststudy odds also depend on prestudy odds. One needs to define carefully the prestudy odds of observational research in each field and setting. Probably much observational research operates in low prestudy odds. This is increasingly common nowadays with massive testing of hypotheses through suitable biologic platforms (e.g., microarrays, proteomics, whole-genome association studies). When the tested biologic factors are enormous and only a few are expected to represent true associations, prestudy odds may be in the range of 10^{-4} to 10^{-8} or even lower. In such circumstances, even decisive support ($B < 0.01$) is insufficient; Bayes factors several log scales lower are needed to make a probed association credible. In some fields, such as molecular epidemiology, false-discovery rate and Bayesian approaches have already been widely adopted (53). For most traditional fields of epidemiology, much resistance towards such approaches may be exactly due to lack of consensus on the prestudy odds. Routine use of Bayes factors would help investigators avoid this problem.

Second, we have no guarantee that evidence always and continuously evolves towards the correct answer (54). Thus, some early meta-analyses with statistically significant results may have identified some true association, but statistical significance was lost in a subsequent meta-analysis because of chance or errors and biases. However, we have no empirical justification for this theoretical claim. Third, the presented analyses do not delve at all into the possibility of biases in the observational literature that may further decrease the credibility of specific associations. Finally, there is the question of how representative the two examined databases are for observational research at large. The first database used the term ''cohort'' in the search strategy (details presented in reference 14) but did not exclude case-control studies. It is thus considerably enriched in cohort studies compared with what one would expect from an entirely random sample of the epidemiologic literature. However, the distribution of p values is similar to what has been seen in other empirical evaluations of random samples of epidemiologic studies (55), where again a large portion of ''significant'' p values hover in the range of 0.01–0.05. If anything, the selection tilt towards cohorts would tend to promote the inclusion of larger and possibly better conducted studies, on average. Moreover, B values were largely similar between case-control studies and studies with cohort or cross-sectional designs in the analyzed data. Conversely, the database of genetic meta-analyses targeted, by default, a highly specialized field. The database is comprehensive for early identified candidate genes (15–17). In the last 2 years, genome-wide association studies have started yielding polymorphisms with extremely low statistical significance levels (19, 56), but these should also be appraised in the context of the background extreme multiple testing. Otherwise, similar approaches could be used. Hopefully the credibility of newer associations generated from a more systematic (rather than one risk factor at a time) approach will eventually be higher.

In conclusion, while the dangers of simply focusing on nominal statistical significance have been repeatedly discussed (8, 57–59), the practice remains entrenched in the biomedical literature and beyond. This represents an overarching problem of interpreting research results regardless of study design (observational or randomized). Past discussions have focused on theoretical concerns and selected studies. The current large-scale evaluation provides additional empirical evidence favoring the routine use of Bayes factors in interpreting "significant" results. This may help us interpret more appropriately the otherwise useful insights we can glean from observational studies.

ACKNOWLEDGMENTS

The author is grateful to Drs. Thomas Trikalinos, Despina Contopoulos-Ioannidis, Evangelia Ntzani, Fotini Kavvoura, and George Liberopoulos for helping with the background work in previous projects that led to building some of the databases used in the current project.

Detailed information on the 272 studies included in the first database and the 50 meta-analyses included in the second database is available upon request.

Conflict of interest: none declared.

REFERENCES

1. Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet 2004;363:1728–31.
2. Lawlor DA, Davey Smith G, Kundu D, et al. Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? Lancet 2004;363:1724–7.
3. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.
4. Wacholder S, Chanock S, Garcia-Closas M, et al. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 2004;96: 434–42.
5. Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. Nat Rev Cancer 2005;5:142–9.
6. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. Chichester, United Kingdom: John Wiley and Sons Ltd, 2004.
7. Kass RE, Wasserman L. A reference Bayesian test for nested hypotheses and its relationship to the Schwartz criterion. J Am Stat Assoc 1995;90:773–95.
8. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. Ann Intern Med 1999;130:1005–13.
9. Cornfield J. The Bayesian outlook and its application. Biometrics 1969;25:617–57.
10. Sasahara A, Cole T, Ederer F, et al. Urokinase Pulmonary Embolism Trial: a national cooperative study. Circulation 1973;47(suppl):II1–108.
11. Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. BMJ 1996;313:603–7.
12. Hughes MD. Reporting Bayesian analyses of clinical trials. Stat Med 1993;12:1651–63.
13. Kavvoura FK, Liberopoulos G, Ioannidis JP. Selection in reported epidemiological risks: an empirical assessment. PLoS Med 2007;4:e79.
14. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88.
15. Ioannidis JP, Ntzani EE, Trikalinos TA, et al. Replication validity of genetic association studies. Nat Genet 2001;29: 306–9.
16. Ioannidis JP, Trikalinos TA, Ntzani EE, et al. Genetic associations in large versus small studies: an empirical assessment. Lancet 2003;361:567–71.
17. 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;164:609–14.
18. Shapiro S. Looking to the 21st century: have we learned from our mistakes, or are we doomed to compound them? Pharmacoepidemiol Drug Saf 2004;13:257–65.
19. Todd JA. Statistical false positive or true disease pathway? Nat Genet 2006;38:731–3.
20. Jeffreys H. Theory of probability. 3rd ed. New York, NY: Oxford University Press, 1961.
21. Lindley DV. A statistical paradox. Biometrika 1957;44:187–92.
22. Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. J Clin Epidemiol 1995;48:45–57.
23. Ioannidis J, Lau J. Evolution of treatment effects over time: empirical insight from recursive cumulative metaanalyses. Proc Natl Acad Sci U S A 2001;98:831–6.
24. Jonsson EG, Kaiser R, Brockmoller J, et al. Meta-analysis of the dopamine D3 receptor gene (*DRD3*) Ser9Gly variant and schizophrenia. Psychiatr Genet 2004;14:9–12.
25. Boekholdt SM, Bijsterveld NR, Moons AH, et al. Genetic variation in coagulation and fibrinolytic proteins and their relation with acute myocardial infarction: a systematic review. Circulation 2001;104:3063–8.
26. Maraganore DM, Lesnick TG, Elbaz A, et al. *UCHL1* is a Parkinson's disease susceptibility gene. Ann Neurol 2004; 55:512–21.
27. Kosmas IP, Tatsioni A, Ioannidis JP. Association of C677T polymorphism in the methylenetetrahydrofolate reductase gene with hypertension in pregnancy and pre-eclampsia: a meta-analysis. J Hypertens 2004;22:1655–62.
28. Burzotta F, Paciaroni K, De Stefano V, et al. G20210A prothrombin gene polymorphism and coronary ischaemic syndromes: a phenotype-specific meta-analysis of 12 034 subjects. Heart 2004;90:82–6.
29. Jonsson EG, Flyckt L, Burgert E, et al. Dopamine D3 receptor gene Ser9Gly variant and schizophrenia: association study and meta-analysis. Psychiatr Genet 2003;13:1–12.
30. Combarros O, Llorca J, Sanchez-Guerra M, et al. Age-dependent association between interleukin-1A (-889) genetic polymorphism and sporadic Alzheimer's disease. A meta-analysis. J Neurol 2003;250:987–9.
31. Marcus PM, Vineis P, Rothman N. *NAT2* slow acetylation and bladder cancer risk: a meta-analysis of 22 case-control studies

conducted in the general population. Pharmacogenetics 2000;10:115–22.

32. McCarron MO, Delong D, Alberts MJ. *APOE* genotype as a risk factor for ischemic cerebrovascular disease: a meta-analysis. Neurology 1999;53:1308–11.
33. Golbe LI, Lazzarini AM, Spychala JR, et al. The tau A0 allele in Parkinson's disease. Mov Disord 2001;16:442–7.
34. Johns LE, Houlston RS. Glutathione *S*-transferase μ1 (*GSTM1*) status and bladder cancer risk: a meta-analysis. Mutagenesis 2000;15:399–404.
35. Kosmas IP, Tatsioni A, Ioannidis JP. Association of Leiden mutation in factor V gene with hypertension in pregnancy and pre-eclampsia: a meta-analysis. J Hypertens 2003;21: 1221–8.
36. Sethi AA, Nordestgaard BG, Tybjaerg-Hansen A. Angiotensinogen gene polymorphism, plasma angiotensinogen, and risk of hypertension and ischemic heart disease: a meta-analysis. Arterioscler Thromb Vasc Biol 2003;23:1269–75.
37. Hashibe M, Brennan P, Strange RC, et al. Meta- and pooled analyses of *GSTM1*, *GSTT1*, *GSTP1*, and *CYP1A1* genotypes and risk of head and neck cancer. Cancer Epidemiol Biomarkers Prev 2003;12:1509–17.
38. Smith GD, Harbord R, Milton J, et al. Does elevated plasma fibrinogen increase the risk of coronary heart disease? Evidence from a meta-analysis of genetic association studies. Arterioscler Thromb Vasc Biol 2005;25:2228–33.
39. Healy DG, Abou-Sleiman PM, Casas JP, et al. *UCHL-1* is not a Parkinson's disease susceptibility gene. Ann Neurol 2006; 59:627–33.
40. Lin J, August P. Genetic thrombophilias and preeclampsia: a meta-analysis. Obstet Gynecol 2005;105:182–92.
41. Ye Z, Liu EH, Higgins JP, et al. Seven haemostatic gene polymorphisms in coronary disease: meta-analysis of 66,155 cases and 91,307 controls. Lancet 2006;367:651–8.
42. Bertram L, McQueen MB, Mullin K, et al. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet 2007;39:17–23.
43. Garcia-Closas M, Malats N, Silverman D, et al. *NAT2* slow acetylation, *GSTM1* null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. Lancet 2005;366:649–59.
44. Sudlow C, Martinez Gonzalez NA, Kim J, et al. Does apolipoprotein E genotype influence the risk of ischemic stroke, intracerebral hemorrhage, or subarachnoid hemorrhage? Systematic review and meta-analyses of 31 studies among 5961 cases and 17,965 controls. Stroke 2006;37:364–70.
45. Zhang J, Song Y, Chen H, et al. The tau gene haplotype h1 confers a susceptibility to Parkinson's disease. Eur Neurol 2005;53:15–21.
46. Mondry A, Loh M, Liu P, et al. Polymorphisms of the insertion/deletion *ACE* and M235T *AGT* genes and hypertension: surprising new findings and meta-analysis of data. BMC Nephrol 2005;6:1.
47. Tripathy CB, Roy N. Meta-analysis of glutathione *S*-transferase M1 genotype and risk toward head and neck cancer. Head Neck 2006;28:217–24.
48. Moonesinghe R, Khoury MJ, Janssens AC. Most published research findings are false—but a little replication goes a long way. PLoS Med 2007;4:e28.
49. Hauben M, Reich L, Van Puijenbroek EP, et al. Data mining in pharmacovigilance: lessons from phantom ships. Eur J Clin Pharmacol 2006;62:967–70.
50. Berger JO. Robust Bayesian analysis: sensitivity to the prior. J Stat Plann Inference 1990;25:303–28.
51. Berger JO, Boukai B, Wang Y. Unified frequentist and Bayesian testing of a precise hypothesis. Stat Sci 1997;12: 133–48.
52. Berger J, Brown L, Wolpert R. A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. Ann Stat 1994;22:1787–807.
53. Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet 2007;81:208–27.
54. Rosenbaum PR. Replicating effects and biases. Am Stat 2001; 55:223–7.
55. Pocock SJ, Collier TJ, Dandreo KJ, et al. Issues in the reporting of epidemiological studies: a survey of recent practice. BMJ 2004;329:883.
56. Chanock SJ, Manolio T, Boehnke M, et al. Replicating genotype-phenotype associations. NCI-NHGRI Working Group on Replication in Association Studies. Nature 2007;447:655–60.
57. Goodman SN. Toward evidence-based medical statistics. 1: the *P* value fallacy. Ann Intern Med 1999;130:995–1004.
58. Goodman S. Commentary: the *P* value, devalued. Int J Epidemiol 2003;32:699–702.
59. Goodman SN. *p* values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol 1993;137:485–96.