

## Betting Odds and Genetic Associations

*Duncan C. Thomas, David G. Clayton*

“Multiple Comparisons? No Problem!” read the title of an editorial in an issue of *Epidemiology* in 1991 (1). Attitudes toward multiple comparisons have swung like a pendulum, the view quoted above representing one end of a cycle that has been reversed in recent years. Few classical risk factor epidemiology studies test more than a few hundred associations between exposures and diseases, even in an exploratory mode. In molecular epidemiology, however, the growing feasibility of testing many thousands, or even millions, of single nucleotide polymorphisms (SNPs) associations in genome-wide studies have forced even the most liberal investigators to re-examine their positions on multiple hypothesis testing (2,3). These concerns have been reinforced by the observation that an alarming proportion of reported associations between genetic variants and diseases are not replicated (4–6).

In the classical theory of multiple comparisons for analysis of variance (developed in the mid-twentieth century), it was recognized that if  $g$  groups are compared, a total of  $g(g-1)/2$  two-way comparisons are possible (7). This classical approach provides investigators with a means to report all possible comparisons without spuriously inflating their type I error rate. The probability of making at least one false-positive statistical significance claim (i.e., the experiment-wise type I error rate) increases with each additional test performed, even though the probability of any one particular hypothesis being rejected is unaffected by the number of further tests carried out. In analysis of variance, the multiple comparison issue is complicated by the fact that the comparisons are not mutually independent. In a simpler case where  $k$  independent hypotheses are tested at a statistical significance level  $\alpha$ , the experiment-wise type I error rate becomes  $1-(1-\alpha)^k$ , leading to the familiar Bonferroni correction in which one must apply a level  $\alpha/k$  to each test to maintain a fixed experiment-wise statistical significance level of  $\alpha$ .

In the mid-1970s, epidemiologists began to doubt that the evidence in support of a hypothesis should depend on the number of tests carried out in the same study (8–11). Indeed, investigators questioned the relevance of the experiment-wise type I error rate and argued that it is not the number of tests performed but rather the prior credibility of the hypotheses that is important for interpreting a set of observed associations. That is, when a hypothesis is unlikely to be true, *a priori*, we should require stronger evidence to be convinced of its truth. In the Bayesian theory of statistics, this argument leads to a formal numerical rule via Bayes theorem; in comparing two rival hypotheses,  $H_1$  and  $H_0$ , the posterior odds for  $H_1:H_0$  are obtained by multiplying the prior odds by the likelihood ratio, the ratio of the probabilities of observing the data if  $H_1$  or  $H_0$  were true. This likelihood ratio measures the weight of evidence for  $H_1$  over  $H_0$ .

In a valuable contribution to this issue of the Journal, Wacholder et al. (12) argued that the Bayesian paradigm is more

relevant to molecular epidemiologic studies than are classical ideas of multiple testing. Few would argue that the simple theory of multiple testing applies naturally in the context of gene association studies, in part because it is often nearly impossible to define how to correct for the appropriate number of tests when using the Bonferroni method; e.g., is it the number of tests we have already carried out in a study, the number of tests we intend to carry out, or the number of tests that we and others might together intend to carry out? Likewise, it makes little sense to require two investigators faced with identical data to come to different conclusions simply because one has carried out more tests than the other. These difficulties in the Bonferroni approach are avoided by focusing attention on the prior credibility of hypotheses.

Wacholder et al. (12) combine these Bayesian ideas with classical notions of statistical significance by reducing the data provided by an experiment to the single observation of whether a statistically significant association was found. If this binary observation is the only one available, then the likelihood ratio is the ratio of the statistical power of the test to the statistical significance level applied. Multiplication of the prior odds in favor of a true association by the likelihood ratio gives the posterior odds. For example, if the prior odds are 1000:1 against a true association that we have 90% statistical power to detect at the 5% statistical significance level, the posterior odds, after observing a statistically significant finding, are still 1000:18 ( $\approx 50:1$ ) in favor of the null hypothesis. That is, the probability that such a positive finding will be false is .982. Wacholder et al. term this probability the false-positive report probability (FPRP) and provide spreadsheet programs for carrying out these simple calculations. It is easy to see from the above example that, if the prior odds of a hypothesis are overwhelmingly in favor of no association, we must use very small  $\alpha$  levels in statistical significance testing, if most positive findings are not to be truly false.

The approach of Wacholder et al. oversimplifies the analysis by assuming a simple binary choice between the null hypothesis of no effect and the hypothesis of an effect of known size. In a full Bayesian analysis, we would also need to specify a prior distribution for the size of effect, and the calculation of the posterior probability of association would then require the like-

---

*Affiliations of authors:* Department of Preventive Medicine, University of Southern California, Los Angeles, CA (DCT); Diabetes and Inflammation Laboratory, University of Cambridge, Cambridge, U.K. (DGC).

*Correspondence to:* Duncan C. Thomas, PhD, Department of Preventive Medicine, University of Southern California, 1540 Alcazar St., CHP-220, Los Angeles, CA 90089-9011 (e-mail: dthomas@usc.edu).

See “Notes” following “References.”

DOI: 10.1093/jnci/djh094

*Journal of the National Cancer Institute*, Vol. 96, No. 6, © Oxford University Press 2004, all rights reserved.

likelihood ratio to be averaged over this distribution. This averaged likelihood ratio, known as the Bayes factor, is advocated as the appropriate measure for the weight of evidence against the null hypothesis. However, we have found that this more formal approach, adopting prior distributions for the size of effect suggested by population genetics (13), gives results very similar to those that would be obtained by the method of Wacholder et al. (12) using a single plausible estimate of effect size.

The Wacholder et al. (12) approach is appropriate for the central problem the authors address, namely, the choice of an appropriate statistical significance level for judging when to describe a finding as noteworthy, which is an appropriate question to address from the standpoint of journal editors and the scientific community as a whole. However, this problem should not be confused with the one faced by a single investigator who has obtained a certain *P* value and wishes to assess the probability of a true association between a genetic variant and a disease. In this case, the datum is the observed *P* value itself, not just the observation of whether it falls above or below an agreed-upon statistical significance level. In this situation, the Bayes factor takes a different form, and we encounter a well-known paradox (14,15)—that is, a *P* value based on a large study has more impact on prior belief than the same *P* value based on a small study. This apparent paradox is not as unreasonable as it might seem, because obtaining a small *P* value from an underpowered study requires an implausibly large effect. Although Wacholder et al. advocate the use of their calculations only for *a priori* determination of reporting thresholds, we believe that these calculations are inappropriate for actual *P* values.

The rigorous basis of Bayesian arguments has rendered them attractive to mathematicians; however, they have been less attractive to scientists. This lack of enthusiasm stems from the requirement that scientists must state prior beliefs about a hypothesis, which seems to contradict the ideal of scientific objectivity. More recently, however, the utility of Bayesian methods in real scientific situations has begun to be debated. For example, Spiegelhalter et al. (16) have convincingly argued that the Bayesian perspective has much to contribute to the analysis of clinical trials. Wacholder et al. (12) have made a similarly convincing case for the application of Bayesian methods in molecular epidemiologic studies. There are, however, two ways to avoid the subjectivity of the Bayesian approach.

The first arises when the prior distribution of effect sizes has some physical basis. Such an argument was used in the landmark paper of Morton (17), who derived the criteria for declaring a linkage “significant” (lod score >3) and for “excluding” a region (lod score less than -2) for a Mendelian single-locus trait, based on the theory of sequential testing for accumulating evidence, combined with the assumption that the locus is, *a priori*, equally probable to lie anywhere in the genome.

Could similar arguments be used for studies of genetic association? Two types of genome-wide association studies can be considered: 1) direct association studies of candidate polymorphisms such as non-synonymous SNPs in exons and 2) indirect association studies, which rely on linkage disequilibrium between markers and unobserved causal polymorphisms. Whereas the size of the human genome is known and the extent of linkage disequilibrium is becoming clearer as the International HapMap Project (18) advances, the main barrier to using physical arguments to derive appropriate prior probabilities is that neither the number of genes that might be truly

involved in a complex disease nor the size of their effects are known at the present time. Indeed, this issue is controversial (19–21). However, such considerations would suggest that the prior odds against an association will usually exceed 1000:1, even for candidate genes, and may even exceed 10 000:1 for random polymorphisms. The arguments of Wacholder et al. (12) would then suggest the use of statistical significance levels in the range of  $10^{-4}$  to  $10^{-6}$ .

Few molecular epidemiology studies, with sample sizes in the hundreds that have been typical in the field, are likely to attain such levels of statistical significance. This lack of statistical power, together with the usual sources of bias (e.g., confounding, inappropriate controls, and measurement error), might account for most of the observed failures to replicate reported associations between genetic variants and diseases. Interestingly, in a recent editorial (22), the authors called for studies that “have large sample sizes, have small *P*-values, report associations that make biologic sense, and have alleles that affect the gene product in a physiologically meaningful way.” In addition, the authors asked for “either a replication in an independent sample or physiologically meaningful data supporting a functional role of the polymorphism in question.” The last of these requirements is particularly telling, because it addresses the issue of *a priori* plausibility, which is central to the Bayesian approach.

The second way in which the subjectivity of the Bayesian approach can be avoided is to simultaneously carry out a large number of tests. In this case, classical multiple testing approaches may be more easily justified than the Bayesian approach; however, even here, use of Bayesian approaches has been advocated in epidemiologic studies (23). Here, the “prior” distribution of true effects can be estimated from the data, and such approaches are consequently termed “empirical Bayes” methods (24). In molecular genetics studies, this approach has been used in the analysis of gene expression array experiments (25), in which the expression of thousands of genes is typically compared. Statistical methods are available to estimate the true percentage of positive findings and the distribution of test statistics among these, allowing calculation of posterior probabilities very similar to FPRPs. These posterior probabilities are termed *q* values (26), which are closely related to the “false discovery rates” proposed by Benjamini and Hochberg (27).

The use of empirical Bayes methods in the context of genome-wide association studies has been advocated (28). However, although direct genome-wide association studies of non-synonymous SNPs are already feasible (even if the coverage is incomplete) and indirect genome-wide studies will become feasible soon, it remains likely that, for the next few years, most association studies will be considered for only limited numbers of candidate genes. For these studies, the choice of appropriate standards of evidence will inevitably involve an element of subjectivity. Wacholder et al. (12) have done the molecular epidemiology community a great service by opening a debate about what prior probabilities might be considered reasonable for different categories of polymorphisms.

The most controversial part of the proposal by Wacholder et al. (12) concerns the reporting of association studies. Wacholder et al. suggest predetermining the FPRP to be used for calling a finding noteworthy. Despite the persistence of the traditional hypothesis-testing framework in basic textbooks, most scientists

today do not simply report hypothesis tests; they prefer to use  $P$  values as descriptive statistics that summarize the weight of evidence. We have already indicated that the relationship between  $P$  values and the Bayes factor is not straightforward. Furthermore, in epidemiologic studies, the emphasis on statistical significance testing has recently been criticized for ignoring the problem of false negatives (29,30). This problem would lead us to suspect that a false-negative reporting probability (FNRP) could also be part of the strategy. Moreover, if only statistically significant findings are reported, the literature will become distorted by reporting bias (31); however, there are now potential avenues for publication of statistically nonsignificant findings.

An emphasis on controlling the false-positive rate can also be challenged. Although there is certainly a cost of false-positive reports in terms of the resources and effort required for follow-up and the credibility of the discipline, the seriousness of failure to replicate findings can be exaggerated. The molecular epidemiology community can live with a moderate false-positive rate for initial reports, provided they are followed by definitive replication studies. This approach is preferable to a strategy that is so conservative at the initial stage as to preclude further discovery.

Finally, we are uncomfortable with the proposal by Wacholder et al. (12) to modify the prior probabilities in light of the literature. Most investigators do not have the resources to perform formal meta-analyses, let alone use them to derive prior probabilities for their own studies. Furthermore, a reporting strategy that leads to varying grounds for noteworthiness would make literature summary extremely difficult. For a discussion of similar issues in reporting clinical trials, see Spiegelhalter et al. (16).

In conclusion, as we move into the era of “genetic dissection of complex traits” (32), we must abandon statistical criteria based on the surely incorrect assumption that a single genetic mutation is the necessary and sufficient cause of disease. Instead, we must think of a “web of causation” (33,34) involving multiple and complex pathways, perhaps involving many genes and environmental substrates. We will need well-designed studies to elucidate many of the complex disease pathways, because the impact of bias and confounding may be serious in the context of small effects and large sample sizes. The fundamental issue of how associations between genetic variants and diseases should be reported in the modern genomic era remains to be resolved.

## REFERENCES

- (1) Poole C. Multiple comparisons? No problem! *Epidemiology* 1991;2:241–3.
- (2) Dahlman I, Eaves IA, Kosoy R, Morrison VA, Heward J, Gough SC, et al. Parameters for reliable results in genetic association studies in common disease. *Nat Genet* 2002;30:149–50.
- (3) Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. *Lancet* 2003;361:865–72.
- (4) Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nat Genet* 2001;29:306–9.
- (5) Hirschhorn JN, Altshuler D. Once and again—issues surrounding replication in genetic association studies. *J Clin Endocrinol Metab* 2002;87:4438–41.
- (6) Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* 2002;4:45–61.
- (7) Hsu JC. Multiple comparisons: theory and methods. London (UK): Chapman and Hall; 1996.
- (8) Cole P. The evolving case-control study. *J Chronic Dis* 1979;32:15–27.
- (9) Miettinen OS. Theoretical epidemiology: principles of occurrence research in medicine. New York (NY): John Wiley & Sons; 1985. p. 113–7.
- (10) Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990;1:43–6.
- (11) Rothman KJ, Greenland S. Modern epidemiology. Philadelphia (PA): Lippincott-Raven; 1998. p. 225–9.
- (12) Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability of false-positive reports in molecular epidemiology studies. *J Natl Cancer Inst* 2004;96:434–42.
- (13) Zeng ZB. Correcting the bias of Wright’s estimates of the number of genes affecting a quantitative character: a further improved method. *Genetics* 1992;131:987–1001.
- (14) Lindley DV. Bayesian statistics, a review. Philadelphia (PA): Society for Industrial and Applied Mathematics; 1971.
- (15) Cox DR, Hinkley DV. Theoretical statistics. London (UK): Chapman and Hall; 1974. pp. 395–9.
- (16) Spiegelhalter DJ, Freedman LS, Parmar MK. Bayesian approaches to randomized trials (with discussion). *J Roy Stat Soc Ser A* 1994;157:357–416.
- (17) Morton NE. Sequential tests for the detection of linkage. *Am J Hum Genet* 1955;7:277–318.
- (18) The International HapMap Consortium. The International HapMap Project. *Nature* 2003;426:789–96.
- (19) Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003;33:177–82.
- (20) Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet* 2002;11:2417–23.
- (21) Tabor HK, Risch NJ, Myers RM. Opinion: candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* 2002;3:391–7.
- (22) Freely associating. *Nat Genet* 1999;22:1–2.
- (23) Thomas DC, Siemiatycki J, Dewar R, Robins J, Goldberg M, Armstrong BG. The problem of multiple inference in studies designed to generate hypotheses. *Am J Epidemiol* 1985;122:1080–95.
- (24) Greenland S, Robins JM. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology* 1991;2:244–51.
- (25) Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. *Genet Epidemiol* 2002;23:70–86.
- (26) Storey, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 2003;100:9440–5.
- (27) Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc Ser B* 1995;57:289–300.
- (28) Sabatti C, Service S, Freimer N. False discovery rate in linkage and association genome screens for complex disorders. *Genetics* 2003;164:829–33.
- (29) Lang JM, Rothman KJ, Cann CI. That confounded P-value. *Epidemiology* 1998;9:7–8.
- (30) The value of P. *Epidemiology* 2001;12:286.
- (31) Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. *J Roy Stat Soc Ser A* 1988;151:419–63.
- (32) Lander ES, Schork NJ. Genetic dissection of complex traits [published erratum appears in *Science* 1994;266:353]. *Science* 1994;265:2037–48.
- (33) MacMahon B, Pugh TF. Epidemiology: principles and methods. Boston (MA): Little, Brown and Co.; 1970. p. 23–5.
- (34) Rothman KJ. Causes. *Am J Epidemiol* 1976;104:587–92.

## NOTE

D. G. Clayton is funded by a Wellcome Trust/Juvenile Diabetes Research Foundation Principal Research Fellowship. Dr. Thomas is supported by grants CA52862 (from the National Cancer Institute), GM58897 (from the National Institute of General Medical Sciences), and P30 ES07048 (from the National Institute of Environmental Health Sciences), National Institutes of Health, Department of Health and Human Services.