American Journal of Epidemiology

Vol. 153, No. 12
Printed in U.S.A.

PRACTICE OF EPIDEMIOLOGY

Commentary: Practical Advantages of Bayesian Analysis of Epidemiologic Data

David B. Dunson

In the past decade, there have been enormous advances in the use of Bayesian methodology for analysis of epidemiologic data, and there are now many practical advantages to the Bayesian approach. Bayesian models can easily accommodate unobserved variables such as an individual's true disease status in the presence of diagnostic error. The use of prior probability distributions represents a powerful mechanism for incorporating information from previous studies and for controlling confounding. Posterior probabilities can be used as easily interpretable alternatives to *p* values. Recent developments in Markov chain Monte Carlo methodology facilitate the implementation of Bayesian analyses of complex data sets containing missing observations and multidimensional outcomes. Tools are now available that allow epidemiologists to take advantage of this powerful approach to assessment of exposure-disease relations. *Am J Epidemiol* 2001;153:1222–6.

Bayes theorem; epidemiologic methods; hierarchical Bayes; latent variable; Markov chain Monte Carlo; posterior probability; prior distribution

Several years ago my wife noticed a lump in her breast. Although she was very young to have developed breast cancer (in her early 20's), she has a family history of the disease and believed that the lump might be malignant. She estimated the risk at 5–10 percent. Her physician had a much lower expectation of her risk, having a knowledge of the medical literature and having seen numerous young women with benign breast tumors. However, he did not report a negative diagnosis until receiving confirmation of the lump's benign status from a biopsy.

The process of updating a patient's and physician's prior beliefs about whether the individual has a disease (in this case, breast cancer) by using a diagnostic test (e.g., a biopsy) is inherently Bayesian (1). In fact, in making the final diagnosis, my wife's physician was essentially applying Bayes' theorem, which in this case can be expressed as

$$P(D = 1|T = 0) = P(T = 0|D = 1)P(D = 1)/$$

$$\{P(T = 0|D = 1)P(D = 1) + P(T = 0|D = 0)P(D = 0)\},$$

where D indicates true disease status ($D = 1$ if disease, $D = 0$ if no disease), T indicates the result of the diagnostic test ($T = 1$ if positive, $T = 0$ if negative), $P(D = d)$ is the prior probability that an individual has disease status d, $P(T = t|D = d)$ is the likelihood of the test result conditional on disease status, and the quantity of interest is the posterior probability of disease conditional on the test result $P(D = 1|T = t)$. From the physician's perspective, my wife's *prior probability* of breast cancer $P(D = 1)$ was low. The biopsy, which has sensitivity $P(T = 1|D = 1)$ and specificity $P(T = 0|D = 0)$, was used to update his prior. The resulting *posterior probability* $P(D = 1|T = 0)$ formed the basis for the physician's diagnosis. Typically, physicians will order more tests (i.e., collect more data) until their posterior probability $P(D = 1|X)$ is close to 0 or 1, where X denotes all of the data collected for an individual.

The application of Bayesian ideas to diagnostic testing is familiar to physicians and epidemiologists. What is much less familiar is the extension of the Bayesian framework to the analysis of data from epidemiologic studies. To illustrate such an extension, let us consider the breast cancer application further. It is well known that carriers of mutations in the *BRCA1* and *BRCA2* genes are at increased risk of breast cancer. In fact, using Bayesian methodology, one can estimate the posterior probability that a woman carries one of these genes conditional on her family history and on prior information about mutation frequencies in the general population and the age-specific incidence rates of breast and ovarian cancer in carriers and noncarriers of the mutations (2, 3). Such posterior probabilities are very useful to physicians, who may otherwise have had to rely on a subjective assimilation of the evidence in making recommendations for genetic testing. In addition, in the absence of genotyping data, the posterior probabilities of a mutation can be used in

Received for publication December 26, 2000, and accepted for publication March 12, 2001.

Abbreviations: ADHD, attention deficit hyperactivity disorder; MCMC, Markov chain Monte Carlo.

From the Biostatistics Branch, MD A3-03, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709 (e-mail: dunson1@niehs.nih.gov). (Reprint requests to Dr. David B. Dunson at this address).

Bayesian analyses of epidemiologic data to assess genetic main effects and gene × environment interactions. Iversen et al. (4) used such an approach to assess the difference in survival after breast cancer onset between carriers and noncarriers of *BRCA1* and *BRCA2* mutations, even though genotyping data were not available.

In such analyses, the presence of the mutation for a given woman is a *latent variable*; that is, it is not observed directly. Other examples of latent variables include an individual's true ferritin value in the presence of measurement error (5), an individual's true disease status in the presence of diagnostic error (6), and the true day of ovulation within a menstrual cycle in fertility studies that use error-prone markers (7). Latent variables can also be more abstract quantities, such as the amount by which an individual's log odds of disease varies from the population mean. Although latent variables can sometimes be incorporated into *frequentist* (i.e., non-Bayesian) models, Bayesian approaches tend to be a more natural statistical formalization of the normal scientific process of evaluating evidence. In addition, it is often the case that a more complex and biologically realistic model can be fitted using Bayesian methods than would have been possible following a frequentist approach.

The goal of this article is to highlight some of the advantages and distinct features of Bayesian analysis of epidemiologic data to encourage epidemiologists to take advantage of this powerful approach to assessing exposure-disease relations. Other recent articles on Bayesian statistics can be found in the epidemiologic and medical literature (8–15).

Bayesian analytical framework

There are no fundamental conceptual differences between the use of Bayes' theorem to obtain a posterior probability of disease for a patient and the general application of Bayesian methods to the analysis of epidemiologic data. In the diagnostic setting one wishes to predict the unknown disease status of an individual, while in analyzing data one wishes to perform inferences on a set of unknowns, which may consist of both latent variables and population parameters (e.g., the regression coefficients in a logistic model). In the diagnostic case, the physician first chooses a prior probability of disease based on the evidence available for the patient. He or she then updates this prior (by ordering appropriate diagnostic tests) to obtain a posterior probability, by plugging the prior and the likelihood of the diagnostic test result (conditional on the latent disease status) into Bayes' theorem. In the general case, the investigator first chooses a *prior probability distribution* for the unknowns in the model (i.e., parameters and latent variables) and then updates this prior distribution to obtain a *posterior distribution* for the unknowns by plugging the prior and the likelihood of the data (conditional on the unknowns) into Bayes' theorem.

According to Bayes' theorem, the posterior distribution is proportional to the product of the prior and the likelihood function, with the likelihood receiving more and more weight as the sample size increases. The posterior distribution summarizes the state of knowledge about an unknown (e.g., the odds ratio for an exposed group) conditional on the prior and current data in the same manner that the posterior probability of disease summarizes the information about an individual's disease status conditional on physician expectation and diagnostic test results. Bayesians base inferences about exposure-disease relations and other hypotheses of interest on the posterior distribution and not on the maximized likelihood or a p value. However, both Bayesian and frequentist statistics incorporate the likelihood of the data from a current study. The Bayesian approach is distinct with respect to both the flexibility with which prior information can be incorporated and the use of posterior probability.

Prior probability distributions

Although most researchers would agree that it is appealing to consider data and information from previous studies in interpreting data from a current study, there is considerable disagreement about whether this prior evidence should be incorporated formally through a Bayesian prior distribution or informally through an investigator's assimilation of the prior and current evidence. Subjective Bayesians advocate choosing informative priors that quantify one's prior beliefs about the likely values for the unknowns independently of the data from the current study. The use of subjective priors has been the most controversial aspect of Bayesian statistics. Many researchers believe that such priors can compromise the integrity of the study results and can even lead to conclusions that are driven not by the data but by a prior representing the unconfirmed beliefs of a possibly overenthusiastic or overskeptical investigator. This criticism is not entirely unfounded, since the choice of the prior certainly contributes to the posterior and therefore to inference. However, responsible subjectivists will conduct sensitivity analyses to evaluate the robustness of their results to the prior choice. Furthermore, priors can often be chosen objectively on the basis of previous data, and investigators who wish to avoid incorporating prior information about an exposure-disease relation can certainly choose a *vague* prior (i.e., one that assigns equal or close to equal probability to a wide range of plausible values) for the regression coefficients of interest. For simple models, Bayesian analyses using vague priors often (but not always) yield results that are quite similar to maximum likelihood-based inferences, at least in large samples.

Prior distributions represent a powerful mechanism for the control of confounding that may even alter how epidemiologists view study design. Consider, for example, epidemiologic studies of infant mortality. It is common knowledge that cigarette smoking during pregnancy conveys a slightly increased risk of infant mortality (the odds ratio associated with smoking is approximately 1.3). In studying other risk factors for infant mortality, a standard analytical approach would be to fit a logistic regression model with exposure metrics for the risk factors of interest included in the model along with potential confounders such as smoking, body mass index, age, and race. Unless the study is large, some of the factors known to be associated with infant mortality (e.g., smoking) may not even be significant according to a likelihood ratio test and may have estimated

odds ratios inconsistent with prior knowledge (e.g., a nonsignificant odds ratio less than 1 associated with smoking). Certainly, many epidemiologists have encountered this common scenario, and some may have considered dropping known confounders from the model if the coefficient of confounding is unreasonable, or even fixing the coefficient (e.g., at the mean of the estimates from previous studies). Some investigators may even avoid conducting studies of small to moderate size if there is insufficient power to obtain good estimates of the coefficients of confounding.

Within a Bayesian analysis, information from previous studies (e.g., estimates of the regression coefficients) can easily be incorporated through an informative prior distribution. This can be done by simply placing prior restrictions on the possible values of the unknowns (e.g., smoking does not have a beneficial effect on infant mortality) or by assigning a prior probability distribution based on data or summary statistics from previous studies. For example, one could choose a prior distribution for the odds ratio associated with smoking that is centered on 1.3 (the approximate mean of the values estimated in previous infant mortality studies) and assigns zero prior probability for values less than 1. Such an approach can improve efficiency and limit bias in estimating the odds ratio for the exposure of interest compared with a frequentist multiple logistic regression analysis, which may produce unreasonable or overly noisy estimates of the coefficients of confounders in small to moderate-sized studies. As the sample size increases, the estimated Bayesian point and interval estimates for the odds ratio will be driven more and more by the observed data and less by the prior. The use of informative priors for the coefficients of confounding is appealing, since epidemiologists typically know something about the influence of commonly measured confounders and want to do the best job possible in controlling their influence.

Perhaps the best way to begin gaining intuition about the Bayesian approach is to choose a prior and to estimate the posterior for a simple application with help from a Bayesian statistician. My expectation is that most investigators will find it appealing to use a prior, particularly for the confounding coefficients, once they become familiar with the process. Note that one can place a vague prior on the parameters of interest to maintain objectivity even when choosing informative priors for the confounding coefficients. In addition, software is available for fitting of a wide variety of Bayesian models (16–18), including multiple logistic regression and even complex hierarchical models with random effects. Although most Bayesian analyses cannot be implemented in SAS, the software package WinBUGS (18) is freely available through the Internet. WinBUGS is easy to use and extremely flexible, and I encourage researchers interested in Bayesian statistics to work through some of the examples provided at the WinBUGS website (www.mrc-bsu.cam.ac.uk/bugs).

Computational advantages in complex models

In fact, although the ease and flexibility with which prior information can be incorporated are a major advantage of the Bayesian approach, the primary factors responsible for the increased use and visibility of Bayesian methods in recent years are the development of Markov chain Monte Carlo (MCMC) algorithms for Bayesian computation (17, 19–21) and the rapid improvements in computing speed that have facilitated implementation of these algorithms. Briefly, MCMC algorithms iteratively generate samples of the parameters in a statistical model. After convergence, these samples represent serially correlated draws from the joint posterior distribution of the model parameters. Based on a large number of iteratively generated samples, one can easily obtain estimates of the posterior distribution of any parameter or function of parameters in a model. Summaries of these posterior distributions may include, for example, posterior means and 95 percent credible intervals, which can be used as Bayesian alternatives to the maximum likelihood estimates and 95 percent confidence intervals, respectively. Unlike confidence intervals, which are typically calculated by assuming large sample approximations, Bayesian interval estimates obtained from MCMC procedures are appropriate in small samples. Bayesian interval estimates also have an intuitively appealing interpretation as the interval containing the true parameter with some probability (e.g., 95 percent). Most researchers prefer this interpretation to that of the $100(1 - \alpha)$ percent confidence interval, which is the range of values containing the true parameter $100(1 - \alpha)$ percent of the time in repeated sampling.

A major advantage of the Bayesian MCMC approach is its extreme flexibility. Using MCMC techniques, it is straightforward to fit realistic models to complex data sets with measurement error, censored or missing observations, multilevel or serial correlation structures, and multiple endpoints. It is typically much more difficult to develop and justify the theoretical properties of frequentist procedures for fitting such models. Consider, for example, studies of neurobehavioral conditions such as attention deficient hyperactivity disorder (ADHD). For such conditions, it is notoriously difficult to identify cases accurately and reliably in an epidemiologic study. Therefore, it is appealing to quantify the occurrence of ADHD using several test items that represent error-prone manifestations of a latent variable measuring the true ADHD status for an individual. Additional latent variables measuring sociologic factors, such as richness of educational environment or level of poverty, can be included in the model to adjust for confounding. Because of the high dimensional integration involved in fitting such models, maximum likelihood approaches are difficult to implement except under normality and linearity assumptions. However, using a Bayesian MCMC approach, a much broader variety of models can be fitted, including those with multilevel correlation structures, different measurement scales for the different test items (e.g., ordered categorical and continuous), and nonlinear regression frameworks (22, 23).

Bayesian hierarchical and latent variable models have been usefully applied in a broad variety of epidemiologic applications, including analyses of the natural history of disease based on interval-censored data (24), spatially correlated disease rates (25–27), measurement error (28), dietary exposures (29), high-dimensional gene expression arrays (30), human fertility (31, 32), and breast cancer susceptibility (2–4).

Posterior probability

In addition to the incorporation of prior information and the ease in computation of complex models, one of the primary advantages of the Bayesian approach is the use of posterior probability. For example, based on fitting of a logistic regression model using an MCMC algorithm, one can obtain estimates of the posterior distributions of not only the regression coefficients (β) but also any function of the regression coefficients (e.g., the odds ratio exp(β)). Posterior means and 95 percent credible intervals can be used to summarize these posteriors. One can also estimate the posterior probability that a regression coefficient is positive (or negative) or equivalently that the odds ratio is greater (or less) than 1. For example, consider a hypothetical study of the effect of lead intake on infant mortality. Suppose that β represents the regression coefficient for individuals with a high lead intake relative to those with a low intake and that 1.8 is the estimated posterior mean, [1.2, 2.3] is the 95 percent credible interval, and 0.02 is the posterior probability of a value less than 1 for the odds exp(β). The posterior probability of an odds ratio less than 1 (0.02) can be used in place of the *p* value. This posterior probability is more intuitive than the *p* value, which is the chance of observing a value as extreme as the observed value given repeated sampling under the null hypothesis. Numerous articles have been published discussing the limitations of *p* values and the advantages of Bayesian approaches to hypothesis testing (33–36). For a brief review of the debate, the reader can refer to a recent paper by Marden (37).

Conclusions

Philosophical issues aside, Bayesian approaches to the analysis of epidemiologic data represent a powerful tool for interpretation of study results and evaluation of hypotheses about exposure-disease relations. This tool allows one to consider a much broader class of conceptual and mathematical models than would have been possible using non-Bayesian approaches. For example, if one wishes to incorporate prior information, this can be done in a flexible manner and inferences can be compared under different priors for the parameters, the latent variables, and even the statistical model itself (e.g., using Bayesian model averaging (38)). In addition, even if vague priors are specified, Bayesian MCMC methods can be used to fit highly realistic models that account for complicating features of an epidemiologic study such as measurement error, multiple endpoints, highly multidimensional data, and spatial correlation. In many cases, these models can be easily fitted using Bayesian software, such as WinBUGS (18). However, unlike routine analyses (e.g., logistic regression in SAS), the subtleties involved in implementing and interpreting Bayesian analyses in current software require some degree of sophistication or collaboration with a statistician. Interested epidemiologists should refer to the book *Bayesian Biostatistics* (39) for a more detailed overview of Bayesian approaches to epidemiology.

REFERENCES

1. Sackett DL, Haynes RB, Guyatt GH, et al. Clinical epidemiology: a basic science for clinical medicine. 2nd ed. Boston, MA: Little, Brown and Company, 1991.
2. Berry DA, Parmigiani G, Sanchez J, et al. Probability of carrying a mutation of breast-ovarian cancer gene *BRCA1* based on family history. J Natl Cancer Inst 1997;89:227–38.
3. Parmigiani G, Berry D, Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes *BRCA1* and *BRCA2*. Am J Hum Gen 1998;62:145–58.
4. Iversen ES, Parmigiani G, Berry DA, et al. Genetic susceptibility and survival: applications to breast cancer. J Am Stat Assoc 2000;95:28–42.
5. Pilote L, Joseph L, Belisle P, et al. Iron stores and coronary artery disease: a clinical application of a method to incorporate measurement error of the exposure in a logistic regression model. J Clin Epidemiol 2000;53:809–16.
6. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol 1995;141:263–72.
7. Dunson DB, Weinberg CR. Modeling human fertility in the presence of measurement error. Biometrics 2000;56:288–92.
8. Davidoff F. Standing statistics right side up. Ann Intern Med 1999;130:1019–21.
9. Etzioni RD, Kadane JB. Bayesian statistical methods in public health and medicine. Annu Rev Public Health 1995;16:23–41.
10. Freedman L. Bayesian statistical methods: a natural way to assess clinical evidence. BMJ 1996;313:569–70.
11. Gurrin LC, Kurinczuk JJ, Burton PR. Bayesian statistics in medical research: an intuitive alternative to conventional data analysis. J Eval Clin Pract 2000;6:193–204.
12. Kadane JB. Prime time for Bayes. Controlled Clin Trials 1995; 16:313–18.
13. Lilford RJ, Braunholtz D. Who's afraid of Thomas Bayes? J Epidemiol Community Health 2000;54:731–9.
14. Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. BMJ 1996;313:603–7.
15. Spiegelhalter DJ, Myles JP, Jones DR, et al. An introduction to Bayesian methods in health technology assessment. BMJ 1999;319:508–12.
16. Best NG, Spiegelhalter DJ, Thomas A, et al. Bayesian analysis of realistically complex models. J R Stat Soc A 1996;159: 323–42.
17. Gilks WR, Richardson S, Spiegelhalter DJ, eds. Markov chain Monte Carlo in practice. Boca Raton, FL: CRC Press, 1996.
18. Lunn DJ, Thomas A, Best N, et al. WinBUGS—a Bayesian modeling framework: concepts, structure, and extensibility. Stat Comput 2000;10:325–37.
19. Gelfand AE, Smith AF. Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 1990;85:398–409.
20. Smith AF, Gelfand AE. Bayesian statistics without tears: a sampling-resampling perspective. Am Stat 1992;46:84–8.
21. Tierney L. Markov chains for exploring posterior distributions. Ann Stat 1994;22:1701–62.
22. Arminger G, Muthen BO. A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. Psychometrika 1998;63: 271–300.
23. Dunson DB. Bayesian latent variable models for clustered mixed outcomes. J R Stat Soc B 2000;62:355–66.
24. Craig BA, Fryback DG, Klein R, et al. A Bayesian approach to modeling the natural history of a chronic condition from observations with intervention. Stat Med 1999;18:1355–71.
25. Carlin BP, Xia H. Assessing environmental justice using Bayesian hierarchical models: two case studies. J Exp Anal Environ Epidemiol 1999;9:66–78.
26. Devine OJ, Louis TA, Halloran ME. Empirical Bayes methods for stabilizing incidence rates before mapping. Epidemiology 1994;5:622–30.
27. Pascutto C, Wakefield JC, Best NG, et al. Statistical issues in the analysis of disease mapping data. Stat Med 2000;19:2493–519.
28. Richardson S, Gilks WR. A Bayesian approach to measure-

ment error problems in epidemiology using conditional independence models. Am J Epidemiol 1993;138:430–42.
29. Witte JS, Greenland S, Haile RW, et al. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. Epidemiology 1994;5:612–21.
30. West M, Nevins JR, Marks JR, et al. DNA microarray data analysis and regression modeling for genetic expression profiling. (ISDS discussion paper 00-15). Durham, NC: Duke University, 2000.
31. Dunson DB, Weinberg CR, Wilcox AJ, et al. Day-specific probabilities of clinical pregnancy based on two studies with imperfect measures of ovulation. Hum Reprod 1999;14:1835–9.
32. Dunson DB, Zhou H. A Bayesian model for fecundability and sterility. J Am Stat Assoc 2000;95:1054–62.
33. Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of *p* values and evidence (with discussion). J Am Stat Assoc 1987;82:112–22.
34. Goodman SN. Toward evidence-based medical statistics. 1: the *p*-value fallacy. Ann Intern Med 1999;130:995–1004.
35. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. Ann Intern Med 1999;130:1005–13.
36. Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc 1995;90:773–95.
37. Marden JI. Hypothesis testing: from *p* values to Bayes factors. J Am Stat Assoc 2000;95:1316–20.
38. Hoeting JA, Madigan D, Raftery AE, et al. Bayesian model averaging: a tutorial. Stat Sci 1999;14:382–401.
39. Ashby D, Hutton JL. Bayesian epidemiology. In: Berry DA, Stangl D, eds. Bayesian biostatistics. New York, NY: Marcel Dekker, Inc, 1996:109–38.