Statistical Science
1990, Vol. 5, No. 3, 269–298

Biostatistics and Bayes

Norman Breslow

Abstract. Attitudes of biostatisticians toward implementation of the Bayesian paradigm have changed during the past decade due to the increased availability of computational tools for realistic problems. Empirical Bayes' methods, already widely used in the analysis of longitudinal data, promise to improve cancer incidence maps by accounting for overdispersion and spatial correlation. Hierarchical Bayes' methods offer a natural framework in which to demonstrate the bioequivalence of pharmacologic compounds. Their use for quantitative risk assessment and carcinogenesis bioassay is more controversial, however, due to uncertainty regarding specification of informative priors. Bayesian methods simplify the analysis of data from sequential clinical trials and avoid certain paradoxes of frequentist inference. They offer a natural setting for the synthesis of expert opinion in deciding policy matters. Both frequentist and Bayes' methods have a place in biostatistical practice.

Key words and phrases: Bioequivalence, sequential clinical trials, longitudinal data, model uncertainty, multiple comparisons, risk assessment.

1. INTRODUCTION

As a graduate student during the mid 1960s, I attended a series of lectures on the theory of games and statistical decisions that provided some familiarity with formal aspects of Bayesian inference: how to convert prior into posterior probabilities by selecting special conjugate priors; how to combine posterior distributions with loss functions so as to take optimal actions; and how to keep one's objectivity by pretending to be uninformed at the beginning of the investigation. An elementary text was available (Chernoff and Moses, 1959) but the more advanced material was taught from scattered articles and notes since the textbooks in current use had not yet appeared. There were few realistic examples of applications, and those that I came in contact with emphasized decision making in social contexts that appeared far removed from my growing interest in medical science.

Once out into the "real world" of biostatistical practice, it was not long before this mild exposure to Bayesian ideas came to be regarded as one of those intellectual exercises one conducts as a student that has little apparent relevance to the task at hand. This attitude was reinforced by a postgraduate year spent in England, where statistics was taught more as an applied science and where skepticism regarding the decision-theoretic approach to statistical inference was openly expressed. Since the Bayesian outlook was at that time firmly coupled to formal decision theory, at least in my mind, it seemed less helpful than informal, frequentist methods of inference for biomedical applications. *P*-values and standard errors served just fine when it came time to analyze clinical or laboratory data, or to instruct medical students on how to interpret the results of statistical studies. A major goal in communicating with medical colleagues was simply to get them to recognize the importance of statistical power for planning the next investigation and of confidence intervals for interpreting the uncertainties associated with the present one. The computations associated with these standard tools were simpler than those required for a Bayesian analysis and, more importantly, they were implemented in standard computer packages. There were far too many studies to plan and too much data to analyze to worry seriously about what the *p*-values and confidence coefficients produced by the packages actually meant. One consoled oneself for this lack of philosophical rigor with the vague notion that, in many standard problems, frequentist and Bayes' inferences with suitably flat priors led to more or less the same conclusions about whether a treatment effect had been

Norman Breslow is Professor and Chairman, Department of Biostatistics SC-32, University of Washington, Seattle, Washington, 98195. This paper was invited by the Biometrics Section for presentation in the session on Historical Perspectives and New Directions in Biometry at the Sesquicentennial Meeting of the American Statistical Association, held in Washington, D.C., August 6–10, 1989.

demonstrated or a new risk factor discovered. Furthermore, the research community was unlikely to accept any new finding until it had been replicated in a variety of circumstances, regardless of what the statistician inferred from the particular study in which he or she had been involved.

In later years this experience with clinical trials and therapeutic medicine led to new interests in epidemiology, disease control and the regulation of hazardous substances. The statistical advances that seemed most relevant were those extending the tools of multiple regression analysis for use with censored observations on survival times (Kalbfleisch and Prentice, 1980; Cox and Oakes, 1984) and with the outcome-dependent sampling schemes known to epidemiologists as the "case-control" study (Breslow and Day, 1980). The development of the generalized linear model (McCullagh and Nelder, 1983) neatly synthesized the most valuable analytic tools for regression modeling of continuous and categorical data into a common conceptual framework and provided increased flexibility for dealing with "overdispersion" and other more complicated variance structures. Recent developments that hold great promise for medical applications include computationally intensive methods for exploratory data analysis (Breiman, Friedman, Olshen and Stone, 1984; Silverman, 1986; O'Sullivan, Yandell and Raynor, 1986; Hastie and Tibshirani, 1986), the generalized estimating equation approach to the analysis of discrete longitudinal data (Liang and Zeger, 1986; Zeger and Liang, 1986) and resampling methods such as the bootstrap (Efron, 1982) for realistic assessment of the uncertainty inherent in complicated inference procedures.

What, then, is the role of Bayesian statistics in biomedical applications? As preparation for this talk, I was invited to review the March 1979, Special Issue of Biometrics, dedicated to Gertrude Cox, which contains 24 articles on diverse aspects of biometry written by the leading statisticians in the field. Not one title, one summary, nor a single keyword referred to Bayesian statistics nor to prior, posterior or predictive distributions. A couple of articles mentioned a Bayesian rationale or derivation for a particular result, but no prominence was accorded this viewpoint. Prior to his death in 1979, Cornfield (1969) was essentially alone among biostatisticians in giving serious attention to Bayesian ideas.

The current decade has witnessed a substantial change in the attitudes of many biostatisticians and other applied statisticians toward implementation of the Bayesian paradigm. In spite of my early experiences in England, a major stimulus has come from there, notably Lindley's (1965) vigorous, philosophical advocacy and, perhaps more importantly, the work of Smith, Skene, Shaw, Naylor and Dransfield (1985) and Smith, Skene, Shaw and Naylor (1987) to develop computational tools that bring Bayesian treatment of realistic, complex problems within the reach of practicing statisticians. This work has inspired the statistical research group at the Swiss pharmaceutical firm CIBA-GEIGY to develop realistic Bayesian approaches to a number of important biostatistical problems including bioassay, crossover trials, bioequivalence studies and pharmacokinetics (Racine, Grieve, Flüher and Smith, 1986). In the United States, Dempster and Rubin and their students have led the way with Bayesian analyses applied, among other things, to historical control data in toxicology and to longitudinal studies of both human and animal populations. Some of the work of both these groups is mentioned further below. There has been a welcome reversal of the unfortunate tendency of Bayesian statisticians to publish their papers in specialized journals or conference proceedings that were read only by other Bayesians. Recent advances in statistical computing (Tierney and Kadane, 1986; Tanner and Wong, 1987; Gelfand and Smith, 1990) promise to make Bayesian analyses even more feasible for those who wish to experiment with them.

The remainder of this paper strives to identify areas of contemporary biostatistics where Bayesian methods, broadly defined, offer promise of improvement on past practice. Although my practical experience with this methodology is essentially nil, I venture to comment on it because of my belief that frequentist inference, in spite of generally great success, fails to deal adequately with a number of important problems. I hope that the Bayesians will excuse my relatively superficial understanding of their subject and that the biostatisticians do not dismiss all of my remarks as due to the naiveté of the nonpractitioner.

2. BAYES' THEOREM

Biostatisticians are of course well acquainted with Bayes' theorem. Elementary courses taught to health professionals often introduce the theorem in the context of the diagnostic tests used in screening programs. If ζ denotes the sensitivity of the test, θ the specificity and π the prior probability that someone has the disease, we know that the predictive value of a positive test (PVP) is given by

$$\begin{aligned}\text{PVP} &= \frac{\pi\zeta}{\pi\zeta + (1-\pi)(1-\theta)}\\ &= \frac{\pi\zeta}{\pi(\zeta+\theta-1)+(1-\theta)},\end{aligned}$$

and thus that even very accurate tests will identify a large number of false positives when the disease is rare. This knowledge has helped to discourage the well-meaning but counterproductive proposals for routine HIV testing of populations at low risk for AIDS. Gastwirth (1987) investigates the sampling error of the PVP when ζ and θ are estimated from laboratory studies while π is estimated from the screening program itself. Since a major source of error in estimation of the PVP is contributed by the standard error of the estimate of θ, he recommends that many more disease-free individuals need to be tested in order to better evaluate its usefulness. This message could well be heeded also by toxicologists who use the Ames/salmonella and other low-cost, short-term test procedures in an attempt to predict the carcinogenicity of chemicals (Chankong, Haimes, Rosencrantz and Pet-Edwards, 1985). More testing of "noncarcinogens" is needed.

Bayes' theorem is used in therapeutic medicine to evaluate patient prognosis and thus contribute to clinical decisions. The journal *Medical Decision Making*, now over 10 years old, contains a section, "Clinical Decision Making Rounds," that features case studies in which a prior probability of success with a particular medical procedure is converted into a posterior probability after appropriate consideration of individual patient characteristics.

In epidemiology, Bayes' theorem constitutes the essence of Cornfield's (1951) demonstration that the exposure odds ratio for diseased (D) and nondiseased ($\sim D$) persons equals the disease odds ratio for exposed (E) versus nonexposed ($\sim E$),

$$\frac{\mathrm{pr}(E \mid D)\mathrm{pr}(\sim E \mid \sim D)}{\mathrm{pr}(\sim E \mid D)\mathrm{pr}(E \mid \sim D)} = \frac{\mathrm{pr}(D \mid E)\mathrm{pr}(\sim D \mid \sim E)}{\mathrm{pr}(\sim D \mid E)\mathrm{pr}(D \mid \sim E)},$$

and thus that the relative risk of a rare disease is estimable from a case-control study. It is now recognized that this result holds much more generally, provided that "relative risk" is understood as the ratio of disease incidence rates and that controls are sampled at random throughout the period in which incident cases are diagnosed (Greenland and Thomas, 1982). Modern methods of analysis of case-control data derive largely from this key relationship (Prentice and Pyke, 1979). Bayes' theorem plays a similarly prominent role in statistical genetics. However, these commonplace applications of Bayes' theorem, where the "prior probabilities" often are based on objective data, do not constitute Bayesian statistics per se and need not concern us further.

3. PROBLEMS OF MULTIPLICITY

Multiple inferences abound in biostatistics. Clinical trials statisticians may attempt to estimate treatment effects separately for different categories of patients defined by age, stage of disease or prior treatment (Simon, 1982). Geneticists contend with a myriad of associations between different diseases and distinct genotypes, defined for example by the multiple alleles of the HLA system (Takasugi et al., 1973). Occupational epidemiologists undertake the simultaneous assessment of relationships between multiple cancer sites and multiple job/exposure categories (Siemiatycki et al., 1982). Toxicologists working to identify chemical carcinogens must examine the data on occurrence of cancer at multiple sites in multiple strains and species of animals (Haseman, 1984). Environmental epidemiologists contend with rates of cancer at multiple sites in multiple geographic areas in order to construct disease maps (Mason and McKay, 1973). Orthodontists use serial measurements of the ramus bone height in order to establish the degree and rate of bone growth and thus plan treatment for a multiplicity of children (Elston and Grizzle, 1962).

Each of the areas mentioned above involves the estimation of an unknown quantity θ_i of interest, for example a treatment difference or a relative disease risk, under a large number of different but possibly related conditions. Individual estimates of each θ_i are likely to be highly variable and virtually useless unless the corresponding sample sizes are large. Simultaneous testing procedures that compare each individual estimate against a null value θ_0 in such a way as to control the probability of reaching any false positive conclusions tend to be overly conservative since they fail to account for the relatedness of the θ_i. In order to make progress, it is essential to reduce the degree of multiplicity by imposing some reasonable structure that expresses this relatedness. Empirical Bayes (EB) methods (Efron and Morris, 1973) were developed expressly for such problems.

In some circumstances, it is feasible to develop a regression model of the form $g(\theta_i) = x_i\alpha$ where the x_i are covariables associated with each estimate, g is a suitable "link" function and α denotes the regression coefficients. Thus in clinical trials we study interactions between treatment and prognostic factors (Byar and Corle, 1977), while in epidemiology we model variations in relative risk according to age and calendar year (Breslow, 1976). However, a regression approach is not always feasible and, even where it is, the observed variation in the individual estimates of θ_i about the regression surface is often larger than reasonably can be ascribed to known sampling errors. This suggests that we regard the unknown parameters θ_i as drawn from a (prior) probability distribution, one that depends on the available covariables though the regression coefficients and on one or more additional parameters ϕ that represent variances and covariances among the random error terms. After estimation of these parameters from the data, the mean or mode of the posterior distribution gives an improved estimate that "borrows strength" from information on related parameters. Some examples will illustrate the point.

3.1 Longitudinal Data Analysis

The analysis of serial or repeated measurements on individuals has a long tradition in biometry. Such data may be collected for the purpose of constructing curves of normal growth, investigating the effects of air pollution on respiratory function or predicting future bone loss from osteoporosis. Some key contributions are those of Potthoff and Roy (1964), Rao (1965) and Grizzle and Allen (1969). Because of computational constraints, most practical applications were to complete sets of data where measurements were recorded at fixed times for each individual.

An influential paper by Laird and Ware (1982) demonstrated that Patterson and Thompson's (1971) theory of REML estimation in variance component models (see also Harville, 1977) could be exploited to treat unbalanced, incomplete data in a common conceptual framework that subsumed much of the earlier work. Their starting point is the general linear mixed model for the vector y_i of (continuous) measurements on the ith individual, namely

$$y_i = X_i\alpha + Z_ib_i + \varepsilon_i,$$

where X_i and Z_i are design matrices linking the responses (y) to fixed (α) and random (b_i) effects. The b_i and ε_i are assumed to be drawn from multivariate, normal populations with 0 means and covariance matrices D and R_i, respectively, that are determined by a vector ϕ of variance components. Following work of Dempster, Rubin and Tsutakawa (1981), Laird and Ware emphasize the connection between REML and Bayesian estimation of α and ϕ in order to provide a unified treatment of estimation and computation via the EM algorithm (Dempster, Laird and Rubin, 1977). The essential step is the EB estimation of the random effect via

$$\begin{aligned}\hat{b}_i(\hat{\phi}_R) &= E\{b_i \mid y_i, \hat{\alpha}(\hat{\phi}_R), \hat{\phi}_R\}\\ &= (Z_i^TR_i^{-1}(\hat{\phi}_R)Z_i + D_i^{-1}(\hat{\phi}_R))^{-1}\\ &\quad \cdot Z_i^TR_i^{-1}(\hat{\phi}_R)(y_i - X_i\hat{\alpha}(\hat{\phi}_R)),\end{aligned}$$

where $\hat{\phi}_R$ is the REML estimate of the variance components. By "borrowing strength" from the entire sample, the individual estimates of b_i are pulled back from the ordinary least-squares estimates, obtained by treating the b_i as fixed effects, towards the common mean of 0. Rao (1975, 1987) develops the EB approach to prediction of future observations on the ith individual, and compares it with other methods.

Hui and Berger's (1983) study of bone loss in postmenopausal women is a nice illustration of EB in practice with longitudinal data. Figure 1 shows the series of observations available for a subsample of women. They first reduce the data to the slopes s_i and estimated variances $d_i = \mathrm{Var}(s_i)$ that were obtained by fitting a linear regression model to each woman's series by ordinary least squares and pooling the error sums of squares. The s_i are modeled

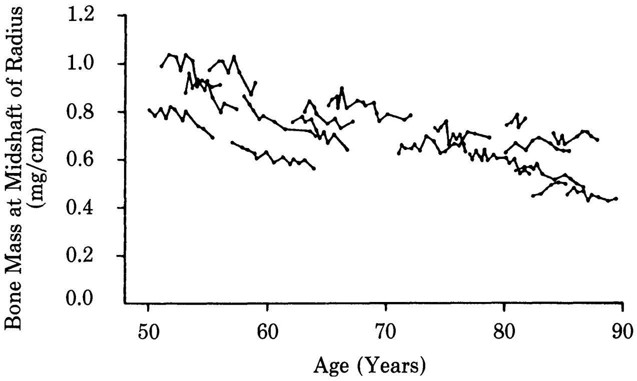

FIG. 1. *Random sample of data from a longitudinal study of bone mass in women. Consecutive measurements of each subject are joined by straight lines. Reproduced from Hui and Berger* (1983) *with permission from the American Statistical Association.*

$$s_i \mid \theta_i \sim \mathcal{N}(\theta_i, d_i),$$
$$\theta_i \sim \mathcal{N}(\alpha_0 + \alpha_1t_i, \phi),$$

where t_i represents a carefully chosen summary of the ages at which data were collected for the ith woman, α_0 and α_1 are parameters of the age specific population mean rate of bone loss (assumed linear) and ϕ is the variance component. Although some information may be lost in comparison with the more comprehensive modeling strategy of Laird and Ware, this approach is simple and robust.

The OLS estimates s_i display a high degree of variability due to the limited data available for some women (Figure 2). The corresponding EB estimates shown in Figure 3 are drawn in sharply towards the population average and appear much more reasonable on general grounds. Hui and Berger (1983) also consider EB estimation of the error variances d_i. This avoids the assumption of constant error variance about the separate regression lines, but the refinement was not needed for the example.

A complication that heretofore has been ignored in typical applications of EB methodology is the contribution of the errors of estimation of α and ϕ to the variability of the estimates $\hat{\theta}_i$. One line of attack on this problem has been via the delta method (Cox, 1975; Morris, 1983) and another via the bootstrap (Laird and Louis, 1987), while a third involves a hyperprior for (α, ϕ) in a fully Bayesian treatment (Lee and Geisser, 1975; Deely and Lindley, 1981). Recent work by Kass and Steffey (1989) combines

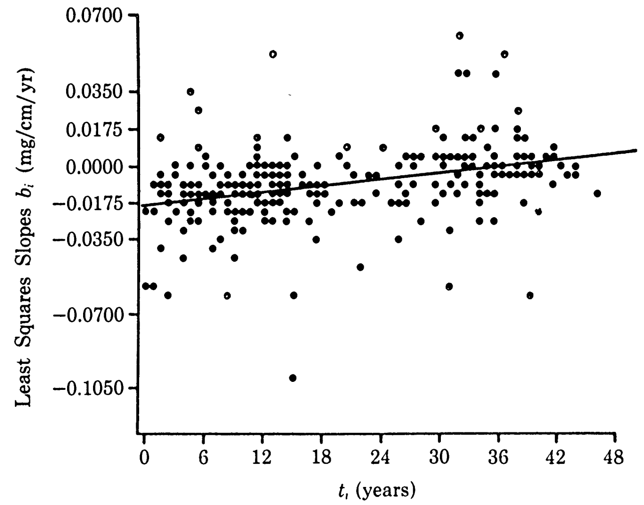

FIG. 2. *Individual least-squares estimates of rate of bone loss b_i versus t_i where the t_i are suitably chosen points in the follow-up intervals. Reproduced from Hui and Berger* (1983) *with permission from the American Statistical Association.*

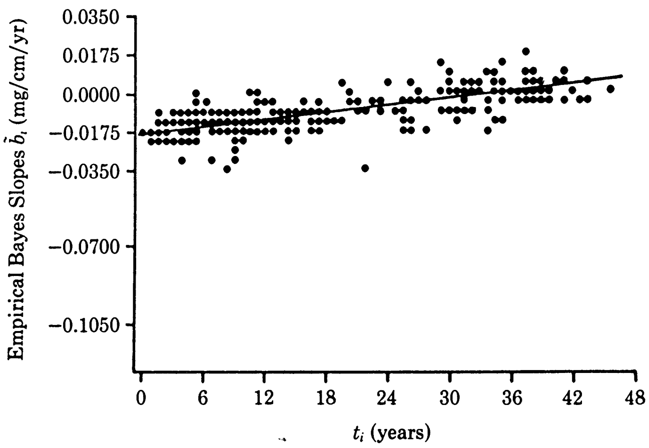

FIG. 3. *Individual empirical Bayes estimates of rate of bone loss b_i versus t_i. Reproduced from Hui and Berger* (1983) *with permission from the American Statistical Association.*

elements of the first and third approaches with Laplace's method for asymptotic integral approximation (Tierney and Kadane, 1986) to yield approximate expressions for var$(\hat{\theta}_i)$ that appear suitable for use in many practical situations.

3.2 Small Area Estimation: Mapping Cancer Rates

Several research teams have applied EB techniques in order to estimate and map cancer mortality rates according to geographic area (Manton et al., 1987, 1989; Clayton and Kaldor, 1987; Tsutakawa, 1988). Clayton and Kaldor's (1987) approach appears particularly well suited to the task, since it incorporates internal estimation of age effects and, at least under one formulation, allows for spatial correlation of random area effects as well as their dependence on regression variables such as degree of urbanization. Briefly, the observed number O_{ij} of deaths in the jth age group and ith area is assumed to have a Poisson distribution with mean $\theta_i \zeta_i n_{ij}$, where the n_{ij} are known person-years denominators, the ζ_j denote fixed age effects and the θ_i are random area effects. Four different models are proposed for θ_i: (i) independent Gamma; (ii) log-normal, without or (iii) with spatial correlation; and (iv) iid nonparametric. Under model (i), one has $E(\theta_i) = \phi/\sigma_i = \exp(x_i\alpha)$ and $\mathrm{Var}(\theta_i) = \phi/\sigma_i^2$, whereby the scale parameters σ_i of the Gamma distributions are given a regression structure. Empirical Bayes' estimates are obtained from the equation

$$\hat{\theta}_i = E(\theta_i \mid O_{i.}, \sigma_i, \phi) = \frac{O_{i.} + \phi}{E_i + \sigma_i}$$

where $O_{i.} = \sum_j O_{ij}$ is the observed number of deaths in area i and the "expected" deaths E_i satisfy

$$E_i = \sum_j n_{ij}\zeta_j.$$

In some applications, the age-specific rates ζ_j are assumed known or proportional to standard rates determined from national vital statistics (Manton et al., 1987, 1989). Otherwise, estimates

$$\hat{\zeta}_j = \frac{\sum_i O_{ij}}{\sum_i n_{ij}\hat{\theta}_i}$$

are inserted at each step of the recursion used to estimate (α, ϕ), which for the log-normal and nonparametric models is easily accomplished via the EM algorithm. This method constitutes the empirical Bayes' equivalent of the epidemiologic technique of indirect standardization, whereby separate estimates of the standardized mortality ratio (SMR) are made for each geographic area (Mantel and Stark, 1968; Breslow and Day, 1975).

Table 1 presents a portion of Clayton and Kaldor's (1987) results for estimating lip cancer incidence in Scottish counties. The columns labeled i–iv correspond to the four models proposed above for the θ_i. All achieve the primary objective of moving the extreme SMRs based on small numbers of cases towards a local or global average. The nonparametric model, for which the estimate of the prior is concentrated on four masses at $\theta = 0.362$, 1.16, 3.08 and 3.89, involves the fewest assumptions and may be preferred for that reason. All the EB estimators appear more sensible for these sparse data than the wildly fluctuating individual SMRs.

3.3 Estimation of Relative Risks in a Case–Control Study

In another application, Thomas et al. (1985) tried to make sense of data from a large case–control study

TABLE 1

Lip cancer incidence in Scotland by county: Observed numbers (O), SMRs and empirical Bayes estimates of the relative risk

County	O	SMR	Empirical Bayes estimate[a]				Adjacent counties
			i	ii	iii	iv	
1	9	652.2	421.9	495.5	453.4	345.0	5, 9, 11, 19
2	39	450.3	414.6	424.5	415.7	367.2	7, 10
3	11	361.8	302.2	310.6	313.0	362.2	6, 12
4	9	355.7	289.7	298.1	271.8	320.5	18, 20, 28
5	15	352.1	308.0	313.9	328.7	320.7	1, 11, 12, 13, 19
21	16	153.0	152.0	149.2	152.5	117.4	16, 29, 50
22	31	136.7	137.1	135.7	146.9	116.5	10, 16
23	11	125.4	127.5	124.5	116.8	116.5	9, 29, 34, 36, 37, 39
24	7	124.6	127.7	123.6	83.5	116.8	27, 30, 31, 44, 47, 48, 55, 56
25	19	122.8	124.2	122.5	131.6	116.5	15, 26, 29
51	1	29.1	57.5	65.0	71.9	57.3	34, 38, 42, 54
52	1	27.6	55.4	63.5	64.9	55.1	34, 40, 49, 54
53	1	17.4	38.3	51.6	53.9	40.4	41, 46, 47, 49
54	1	14.2	32.3	47.1	50.3	37.8	34, 38, 49, 51, 52
55	0	0.0	30.9	58.5	92.7	61.2	18, 20, 24, 27, 56
56	0	0.0	56.4	70.4	72.6	40.9	18, 24, 30, 33, 45, 55

[a] Model (i) independent gamma; (ii) log-normal without or (iii) with spatial correlation; (iv) nonparametric. See text.
Reproduced from Clayton and Kaldor (1987) with permission from the Biometric Society.

designed to discover occupational carcinogens. After substantial preliminary analysis, they confronted a collection of 684 relative risk estimates RR_i, one for each combination of 12 cancer sites and 57 exposure categories. Each relative risk was estimated separately with adjustment for a priori confounders but not for other chemical exposures. They chose not to impose any regression or correlation structure on this table but rather simply assumed that, as a first approximation, the log RR_i were independently and normally distributed with means θ_i and variances S_i^2 that were determined from the data to account for sampling error. The θ_i were drawn from a mixture of two normals, $\lambda\mathcal{N}(0, \sigma_1^2) + (1-\lambda)\mathcal{N}(\mu, \sigma_2^2)$, where λ represents the proportion of "null" associations, $\sigma_1^2 > 0$ accounts for uncontrolled confounding or inappropriate reference groups, and $1 - \lambda$ represents the proportion of true "positive" associations distributed around a mean of μ.

Table 2 contrasts the EB estimates with ML estimates for a subset of 17 associations selected as "positive" by a stepwise variable selection algorithm applied with a polytomous logistic regression procedure (Thomas, Goldberg, Dewar and Siemiatycki, 1986). The EB estimates are pulled back sharply towards the estimated median relative risk of 1.30. Whether or not the EB estimates are closer to the "truth" depends on our judgment as to whether the assumed mixed normal prior is more reasonable than the flat, improper prior that is implicitly assumed by ML. The flat prior is clearly unreasonable in this example. However, our confidence in the EB estimates would be raised considerably if the authors had shown that they remained fairly stable under a variety of plausible prior distributions for the log relative risks.

3.4 Multiple Tumor Sites in a Toxicology Experiment

More structure is assumed by Meng and Dempster (1987) in a toxicological analysis of tumors occurring at 17 separate sites for Sprague-Dawley rats in one treatment group, one concurrent control group and six historical control groups. They postulate logit models

$$\text{logit } P_{ij}^{C} = b_i + u_j + d_{ij},$$

$$\text{logit } P_{0j}^{T} = \eta + b_0 + u_j + d_{0j} + t_j,$$

for the binomial probabilities P_{ij} of tumor occurrence at the jth site for animals in the treatment (T) or ith control (C) group, $i = 0, \cdots, k$. The b_i, t_j and d_{ij} are assumed to be independently and normally distributed with 0 means and variances σ_b^2, σ_c^2 and σ_d^2. The parameters η and u_j have flat priors, meaning they are regarded as fixed effects. Thus u_j represents the average (on the logit scale) "spontaneous" tumor occurrence at site j in the historical control series, while η represents the average treatment effect regardless of tumor type. The interexperimental variation is broken down into overall (b_i) and site-specific (d_{ij}) components. A critical assumption is that the treatment and concurrent control groups have the same assay effects b_0 and d_{0j}, in other words that there is no extraneous within assay variation (Board of Scientific Counselors, 1984).

TABLE 2

Maximum likelihood (ML) and empirical Bayes (EB) estimates of risk parameters for associations selected by stepwise analysis[a]

Total no. of cases	Proportion exposed	Expected no. of exposed cases	Log relative risk estimates			
			ML[b]	EB[c]	Prior[d]	Pull back %
246	0.198	48.7	1.55	1.43	1.30	47
212	0.108	22.9	1.89	1.52	1.30	62
100	0.198	19.8	1.85	1.48	1.30	67
246	0.039	9.6	3.88	1.86	1.30	78
69	0.306	21.1	1.58	1.39	1.30	67
187	0.091	17.0	1.88	1.45	1.30	74
255	0.051	13.0	2.22	1.54	1.30	74
100	0.167	16.7	1.53	1.37	1.30	69
40	0.306	12.2	1.90	1.42	1.30	80
187	0.038	7.1	1.97	1.50	1.30	70
255	0.013	3.3	4.71	1.96	1.30	81
212	0.027	5.7	2.01	1.42	1.30	83
36	0.108	3.9	3.12	1.51	1.30	88
69	0.043	3.0	2.85	1.47	1.30	89
69	0.032	2.2	2.81	1.45	1.30	90
40	0.041	1.6	4.76	1.56	1.30	92
246	0.005	1.2	3.61	1.41	1.30	95

[a] Entry criterion: $p < 0.10$ for the score statistic conditional on previously entered associations, no associations eliminated.

[b] ML estimate derived from logistic regression coefficient, adjusted for the other associations selected by stepwise analysis and for a priori confounders.

[c] Based on the expected value of log RR given RR $\neq$ 1.

[d] Estimated prior mean, assuming common distribution for RRs, fitted to Mantel–Haenszel estimates of RR for all 684 associations.

Reproduced with permission of the publisher from Thomas et al. (1985).

The essential innovation over other random-effects models for categorical data is the representation of even the site-specific treatment effects $\eta + t_j$ as sampled from a normal population with mean η and variance σ_t^2. For the data they consider, the estimated mean treatment effect is $\hat{\eta} = -.10$, near zero, but the estimated variance is a large $\hat{\sigma}_t^2 = 0.41$, suggesting the presence of large (positive or negative) treatment effects at individual tumor sites. Meng and Dempster (1987) suggest using Bayesian p-values $\text{pr}[\eta + t_j < 0 \mid \text{data}]$ in order to identify individual tumor types that may be affected by treatment. These are contrasted in Table 3 with "chi-square" p-values that arise from a conventional analysis that uses data from each 2 × 2 table of concurrent control versus treated and tumor-bearing versus non–tumor-bearing animals, with consideration of neither the historical controls nor the multiple comparison issue. The advantage of the Bayesian approach is that the number of "borderline significant" findings for tumor sites with a highly variable control incidence is substantially reduced. The Bayesian p-values are generally pulled back toward the middle. However, the positive finding for testicular adenoma is strengthened by incorporation of historical control data that show a relatively low incidence ($^{19}/_{353} = 5.4\%$) compared to the concurrent control ($^{11}/_{60} = 18.3\%$) and treatment ($^{19}/_{60} = 31.7\%$) groups. The Bayesian p-values are more conservative than "partially pooled" p-values that are based on a separate analysis at each tumor site that incorporates the historical data (Dempster, Selwyn and Weeks, 1983). For larger values of σ_t^2, however, the Bayesian p-values are less conservative.

3.5 Remarks on Exchangeability

The key concept that allows progress to be made in all these examples is that of "exchangeability": exchangeability of growth parameters among individuals in the longitudinal study; exchangeability of treatment effects at different tumor sites in the bioassay; exchangeability of relative risks for different cancer/exposure combinations in the case–control study; and exchangeability of SMRs at different geographic areas in the cancer mapping study. One admits at the outset the futility of estimating each of these effects separately and instead determines a mean effect toward which individual estimates are displaced depending upon their intrinsic stability. Strict exchangeability may be relaxed when one has reason a priori to believe that certain effects are correlated, for example, because of the proximity or similarity of different tumor sites or tissues, or the structural similarity of different chemicals. Clayton and Kaldor (1987) illustrate this approach in their model (iii), whereby high spatial correlations result in the individual SMRs being displaced more towards local than global mean values.

TABLE 3
Posterior p-values of treatment effects

Tumor type	Current		Historical control incidence (pooled)	Bayesian p-value	Chi-square p-value	Partial pooling p-value
	Control	Treated				
1. Thyroid-adenoma	3/60	4/60	23/391	0.0545	0.0348	0.0464
2. Skin-fibroma	1/60	4/60	19/414	0.0347	0.0085	0.0187
3. Skin-lipoma	1/60	2/60	3/414	0.0348	0.0279	0.0126
4. Skin-squamous cell carcinoma	0/60	1/60	5/414	0.0513	0.0158	0.0346
5. Skin-fibrosarcoma	2/60	1/60	7/414	0.0624	0.0721	0.0534
6. Skin-papilloma	2/60	0/60	4/414	0.0700	0.0923	0.0539
7. Testis-adenoma	11/60	19/60	30/413	0.0038	0.0046	0.0020
8. Kidney-adenoma	0/60	1/60	1/414	0.0411	0.0158	0.0266
9. Pituitary-adenoma	34/60	18/59	196/405	0.0995	0.0998	0.0996
10. Adrenal-adenoma	5/60	6/60	21/405	0.0381	0.0376	0.0287
11. Adrenal-ganglioneuroma	0/60	1/60	1/405	0.0412	0.0158	0.0270
12. Liver-hepatocarcinoma	2/60	2/60	3/415	0.0383	0.0500	0.0218
13. Systemic-reticular cell sarcoma	1/60	0/60	5/409	0.0698	0.0842	0.0529
14. Pancreas-adenoma	8/60	4/60	39/406	0.0828	0.0888	0.0848
15. Brain-glioma	1/59	0/60	7/412	0.0725	0.0844	0.0529
16. Lungs-adenoma	3/60	0/60	5/413	0.0747	0.0960	0.0548
17. Parathyroid-adenoma	2/54	0/57	7/371	0.0758	0.0929	0.0541

Adapted and reproduced from Meng and Dempster (1987), with permission from the Biometric Society.

The major question raised for discussion is the reasonableness of this assumption of a common prior distribution for unknown parameters of interest and the extent to which the often quite arbitrary specification of the shape of this distribution influences the scientific conclusions. In the context of clinical trials, Cornfield (1976) remarks that the alternative to specification of a common prior is the assignment of different priors, which leads to each parameter being treated separately in the analysis. He notes that the choice between these two alternatives ultimately depends on one's belief that the effects represented by the parameters "have something in common." Tukey (1977) argues that such assumptions "do not seem . . . to be near enough the real world to be a satisfactory and trustworthy basis for the careful assessment of strength of evidence to which the ethical issues . . . must dedicate us." This comment may apply with even greater force to the toxicology examples in both this and the next section. The concept of "borrowing strength" seems most appealing for longitudinal data analysis and for the mapping problem.

4. SPECIES TO SPECIES EXTRAPOLATION IN CANCER RISK ASSESSMENT

A vexing problem in cancer risk assessment is a relative abundance of data on multiple animal species exposed to multiple environmental agents, but a paucity of such data on humans. The relevance of the animal data for assessing human risk has been a subject of considerable debate (Freedman and Zeisel, 1988). DuMouchel and Harris (1983) ambitiously addressed this issue from a Bayesian perspective that made rather strong assumptions about the relatedness of the human and animal studies. Their methodology was adopted by the BEIR IV (1988) committee of the National Academy of Sciences to infer the carcinogenic potency of radioisotopes of plutonium deposited internally in humans from available data in dogs and rats (Table 4).

DuMouchel and Harris start with a series of observations y_{ij} on the carcinogenic potency of agent j in species i, expressed as the log transform of the slope of a linear dose–response function fitted to the available experimental or epidemiological data. The y_{ij} are assumed to be normally distributed with means θ_{ij} and standard errors C_{ij} that are estimated from the dose–response fit but treated as known in subsequent analysis. They further assume that, conditional on the values of parameters α_i, β_j, σ^2,

$$\theta_{ij} \mid (\alpha_i, \beta_j, \sigma^2) \sim \mathcal{N}(\alpha_i + \beta_j, \sigma^2).$$

Thus the prior expectation is that the ratio of carcinogenic potencies for agent j in distinct species i and h, namely $\exp\{E(\theta_{ij}) - E(\theta_{hj})\} = \exp(\beta_j)$, is constant. The variance parameter σ^2 measures how well the actual θ_{ij} conform to this expectation.

In applying this model, DuMouchel and Harris initially assume a flat prior on the α_i and β_j, so that these quantities are effectively estimated from the data. They specify two proper priors for $\log \sigma$, both of which restrict only slightly the plausible range of multiplicative factors by which any individual potency deviates from the constant relative potency model. Posterior distributions of θ and σ are then used for inference. BEIR IV chooses a prior for σ based on the stronger assumption that the (multiplicative) interaction factor is unlikely (has prior probability less than 0.05) to be greater than 10 or less than 0.1. BEIR IV also specifies proper prior distributions for $\{\alpha_i\}$ and $\{\beta_j\}$, namely correlated normals that specify a close relationship between the average potencies in beagle dogs, regardless of mode of administration, and a somewhat less close relationship between the potencies of the two radium isotopes, on the one hand, and the two plutonium isotopes, on the other. Table 5 shows the moments of the posterior distributions of the log relative potencies θ_{1j} in humans, and Bayesian medians and confidence intervals for the potencies $\lambda = \exp(\theta)$ assuming the posterior distribution of θ to be normal.

The Bayesian analysis of the limited data in Table 4 reduces substantially the uncertainty in the potency estimate for ^{228}Ra, for which some human data were available, and yields extrapolated potency estimates for the two plutonium isotopes, for which human data were lacking altogether. Clearly one needs to consider very carefully the validity of the "extra information" provided by the assumed model and the assumed prior before accepting these results at face value. In his discussion of the DuMouchel and Harris paper, Kass (1983) remarks that the normal exchangeable prior assumed for the θ_{ij} is an important cause for concern since it expresses the idea of relatedness in a very strong manner. He concludes "I would hesitate to apply the model without additional theoretical or empirical knowledge." Kass and other discussants also point out that the precision of the Bayes' estimates of potency are overstated because of the failure to consider the uncertainty in other model assumptions, particularly the linearity of the dose–response relationships.

TABLE 5

Posterior distributions of carcinogenic potencies in man following Bayesian analysis of Table 4 data

Parameter of posterior distribution	Isotope			
	^{226}Ra	^{228}Ra	^{238}Pu	^{239}Pu
Moments of log potency θ				
$E(\theta)$	−3.22	−2.81	−1.11	−1.12
Standard deviation (θ)	0.30	0.46	0.65	0.65
Percentile of potency λ				
$\lambda_{.5}$	0.04	0.06	0.33	0.33
$\lambda_{.025}$	0.02	0.03	0.09	0.09
$\lambda_{.975}$	0.07	0.15	1.12	1.12

Reproduced with permission from *Health Risks of Radon and Other Internally Deposited Alpha Emitters*, © 1988 by the National Academy Press.

5. BIOEQUIVALENCE

The current protocol of the National Wilms' Tumor Study (NWTS), an effort with which I have been associated for more than 20 years, compares "standard" combination chemotherapy with "pulsed/intensive" administration of the same two agents. The intent is not so much to improve on the already excellent 90% cure rate for the majority of "low-risk" patients with this rare childhood kidney tumor, although this of course would be a welcome outcome, as it is to reduce the enormous financial and social burden borne by the patients' families, without compromising therapeutic effectiveness. Demonstration of such biological "equivalence" is also the goal of pharmaceutical firms trying to convince the FDA that their generic drug is absorbed into the bloodstream just as completely and rapidly as the brand name product. In view of the logical absurdity of trying to prove the null hypothesis, at least in a study of finite duration, it is well recognized that the usual formulation of testing the hypothesis of equality against an alternative of therapeutic efficacy is inappropriate (Westlake, 1979). Instead, one typically decides on a zone of "indifference," such that differences in biological activity or therapeutic efficacy within this zone are judged irrelevant in comparison with the anticipated savings in cost or toxicity. The study is designed so that larger differences are ruled out with high probability if the two treatments are in fact equivalent. Interval estimation is an obvious approach (Westlake, 1979). However, interval estimation may be complicated by the presence of a nuisance parameter, for example the unknown mean for the standard treatment when the criterion is the ratio of means. Accordingly, some authors (Dunnett and Gent, 1977; Hauck and Anderson, 1986) have advocated a hypothesis testing framework in which the null and alternative hypotheses are interchanged, such that the null hypothesis is specified as the limit of the indifference zone. This specification seems rather unnatural, however, a fact that was brought home to me when an NIH study section nearly refused to fund our trial because the statistical section of the protocol had adopted it.

TABLE 4

Log carcinogenic potencies (standard errors) of bone cancer by radioisotope and test species

Species (mode)	Isotope			
	^{226}Ra	^{228}Ra	^{238}Pu	^{239}Pu
Human	−3.30 (0.32)	−2.43 (0.72)	—	—
Beagle dog (injection)	−0.87 (0.21)	−0.32 (0.16)	—	1.69 (0.17)
Beagle dog (inhalation)	—	—	1.55 (0.15)	—
Rat	—	—	0.64 (0.34)	0.29 (0.24)

Reproduced with permission from *Health Risks of Radon and Other Internally Deposited Alpha Emitters*, © 1988 by the National Academy Press.

Bioequivalence is a perfectly natural concept for the Bayesians. Given appropriately diffuse prior distributions on parameters specifying the outcome distributions in treatment and control groups, they simply compute the posterior probability that the parameter of interest falls in the indifference zone and declare the two treatments "equivalent" if it is sufficiently large. For example, Flüher, Grieve, Mandallay, Mau and Moser (1983) consider a normal theory model involving means (μ_t, μ_c) of continuous outcome variables for treatment and control with constant unknown variance σ^2. The prior on (μ_t, μ_c, σ^2) is assumed proportional to σ^2. They define $\psi = \mu_t/\mu_c$ and declare "equivalence" if $\mathrm{pr}[0.8 \le \psi \le 1.2 \mid \text{data}]$ exceeds 0.95. Racine-Poon, Grieve, Flüher and Smith (1987) use the same model with a two stage sampling procedure to show that, given the first stage data, the predictive probability of establishing bioequivalence is a useful guide as to whether it is worth the effort to go on to the second stage.

Selwyn, Dempster and Hall (1981) utilize similar techniques to analyze data from crossover trials carried out in connection with Abbreviated New Drug Applications submitted to the FDA. In more recent work, Selwyn and Hall (1984) propose a Bayesian analog of the "75/75" rule, which stipulates that at least 75% of the subjects in a study must have blood concentration levels (measured by time-weighted average or maximum concentration) that are at least 75% of those for the standard formulation. Specifically, using paired observations (W_i, Z_i) on the blood levels measured under standard and new drugs for the ith of n subjects, they calculate the predictive probability $\mathrm{pr}[Z_{n+1} > 0.75W_{n+1} \mid \text{data}]$ that the outcome ratio meets the criterion, and determine whether it exceeds 0.75.

One reason the Bayesian formulation appears so natural in this context is undoubtedly the cost-benefit, decision nature of the problem. The goal of the pharmaceutical firms is to satisfy explicitly stated federal criteria for efficacy, safety and equivalence. Much of the work of specifying indifference zones and confidence levels, which often seems so arbitrary when carried out ad hoc, has been mandated already by government.

6. SEQUENTIAL CLINICAL TRIALS

Experimentation on human subjects poses a conflict between the desire for scientific proof of therapeutic effectiveness and the ethical imperative of providing patients with the best available treatment. Clinical trials that are carried out with a fixed, large sample size generally provide the most convincing evidence of therapeutic efficacy. However, adherence to a predetermined sample size may prove untenable when differences between regimens start to appear earlier than anticipated at the outset of the trial. Medical statisticians have developed sequential stopping rules that are explicitly designed to control Type I and II errors in such situations (Armitage, 1975; Whitehead, 1983). A problem with this approach is that the error probabilities, which are often rather arbitrarily selected when used to help plan fixed sample size studies, are now translated into inflexible decision procedures that cannot easily accommodate changes in trial goals and other unforeseen developments. My only attempt to employ a formal sequential design with the NWTS was frustrated when the study committee decided in the middle of the trial that the end point used for design purposes, namely, relapse to any site, was an inappropriate basis for the decision to stop. Since the trial was of radiation therapy, the committee ultimately concluded that the only relevant criterion was the much rarer end point of relapse to the (irradiated) tumor bed.

Part of the difficulty in specifying flexible sequential decision rules has been the technical problem of evaluating the operating characteristics of proposed sequential boundaries. The classical parallel line and wedge-shaped boundaries based on variations of Wald's sequential probability ratio test are used as much because their sampling properties are well understood as for their purported optimality in simple decision problems. Current practice emphasizes group sequential designs (Pocock, 1977; O'Brien and Fleming, 1979) in which a sequence of tests at prescribed significance levels is carried out at fixed time intervals or after a fixed number of patients have been entered. With one of the popular O'Brien–Fleming tests used for Phase II cancer studies, for example, a decision is supposed to be made after 20 patients have been enrolled as to whether to proceed with the trial. In multicenter trials, however, it frequently happens that more patients (say 25) actually have been entered by the time the decision to stop or continue is taken. Strict adherence to the sequential plan would require that data on five patients be discarded, for example, according to the date of arrival of the records in the statistical center. Fortunately, clinical trial statisticians behave more sensibly in actual practice (Crowley, personal communication).

Lan and DeMets (1983) proposed a new concept for sequential testing, namely to "spend" the Type I error at a prespecified rate throughout the trial. This lends greater flexibility to the group sequential approach since it accommodates multiple looks at arbitrary time points. However, inference is still dependent upon reaching a prespecified boundary. If the trial must be abandoned before a boundary is crossed, due to a loss in funding or a change in priorities, no inference is possible.

Jennison and Turnbull (1989) developed further the concept of spending Type I error in their construction of repeated confidence intervals for interim analysis. These allow informal exploration of the data at each time point and do not depend on a rigid stopping rule. If the trial has to be abandoned before a definitive outcome is reached and before all the Type I error has been spent, the only cost is conservatism in the confidence interval for the treatment difference observed at that point. However, the paradox that affects this and all other frequentist approaches to sequential analysis is evident from the fact that, once all the Type I error has been spent, "no use can be made of any future data" (Jennison and Turnbull, 1989).

The logical shortcomings of frequentist sequential inference have been articulated convincingly by Anscombe (1963), Cornfield (1966, 1976) and Berry (1987). At the heart of the matter is the fact that the statistical conclusions drawn from the data available at the end of the trial depend critically on the stopping rule selected at the outset. Faced with exactly the same set of data, two statisticians may report radically different p-values because one entertained the possibility of early stopping while the other did not. Interval estimates of parameters of interest following a sequentially stopped trial (Tsiatis, Rosner and Mehta, 1984) likewise depend on the stopping rule and thus on consideration of "data that might have been observed but were not" rather than on the relevant sufficient statistics. The repeated confidence interval approach is unable to accommodate new data once the prespecified Type I error has been "used up."

Bayesian statisticians have much greater flexibility in dealing with sequential clinical trials because their posterior and predictive distributions depend only on the likelihood of the observed data and not on the stopping rule. Work by Chaloner and Duncan (1983) and Freedman and Spiegelhalter (1983) should help to familiarize the medical community with the concept of prior distributions and their elicitation from those responsible for the conduct of the trial. It is entirely appropriate that the prior so determined is considered as part of the decision to stop. This decision is probably best made informally, by consideration of the posterior distribution of the treatment effect or the predictive probability that future observations will provide conclusive evidence of effectiveness to share with colleagues (Herson, 1979; Spiegelhalter, Freedman and Blackburn, 1986). Formal decision-theoretic treatments based on minimizing the number of patients assigned to the "wrong" treatment have not yet taken sufficient account of the fact that clinical trials are designed to provide scientific data to an international community of physicians, as well as to improve treatment in the specific population from which the subjects are drawn. This is a difficult arena in which to formalize the cost–benefit equation (Cornfield, 1976; Simon, 1977).

A nice example of Bayesian sequential inference in practice is provided by Spiegelhalter, Freedman and Blackburn (1986). They study the use of the predictive distribution to answer a question frequently posed by data monitoring committees, namely: "What, given the data so far, are the chances of getting a conclusive result if we complete the trial?" They define a critical region R, based for example on a fixed sample size test, such that observations falling in R would be widely interpreted as conclusive evidence of a treatment difference. Given the data x_0 available at some interim point in the trial, they then calculate both the conditional power $\mathrm{pr}(R \mid \theta, x_0)$ and the posterior distribution $\mathrm{pr}(\theta \mid x_0)$ as functions of the parameter θ of interest. Averaging the conditional power with respect to the posterior produces the predictive probability of rejecting the null hypothesis. This has the advantage over stochastic curtailed sampling (Halperin, Lan, Ware, Johnson and Demets, 1982) of accounting in the posterior distribution for the accumulated information about θ. The conditional power function at a prespecified θ is not relevant for interim analysis if θ is no longer a plausible value.

Figures 4 and 5 illustrate this process with binomial observations on the probability of failure in a control (C) and treatment (T) group. At an interim point, with 887 patients already observed on each treatment arm and 592 more to come, 43 failures have occurred

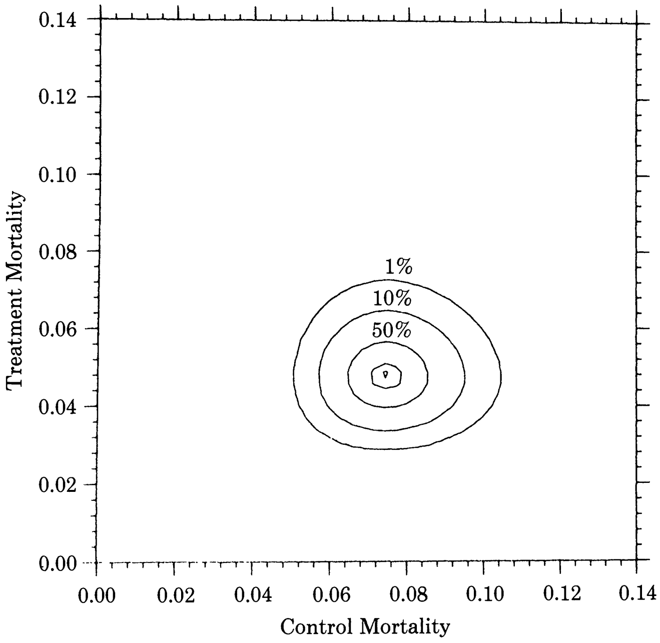

FIG. 4. *"Conditional power" contours showing the probability of rejecting H_0: $p_c = p_t$ in favor of $p_c > p_t$ given the data so far, as a function of true mortality rates p_c and p_t. Reprinted with permission of the publisher from Spiegelhalter, Freedman and Blackburn* (1986).

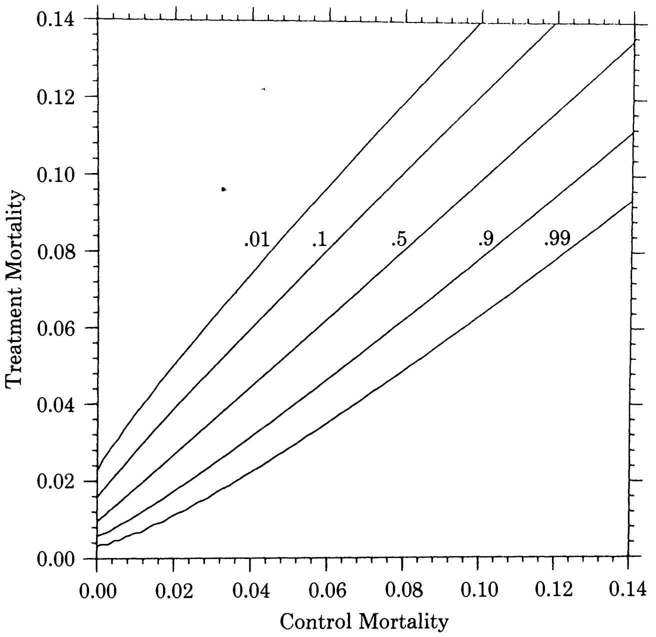

FIG. 5. *Current joint belief concerning p_c and p_t, expressed as contours of* 1%, 10%, 50%, 90% *and* 99% *of the maximum ordinate, assuming pretrial independence of p_c and p_t. Reprinted with permission of the publisher from Spiegelhalter, Freedman and Blackburn* (1986).

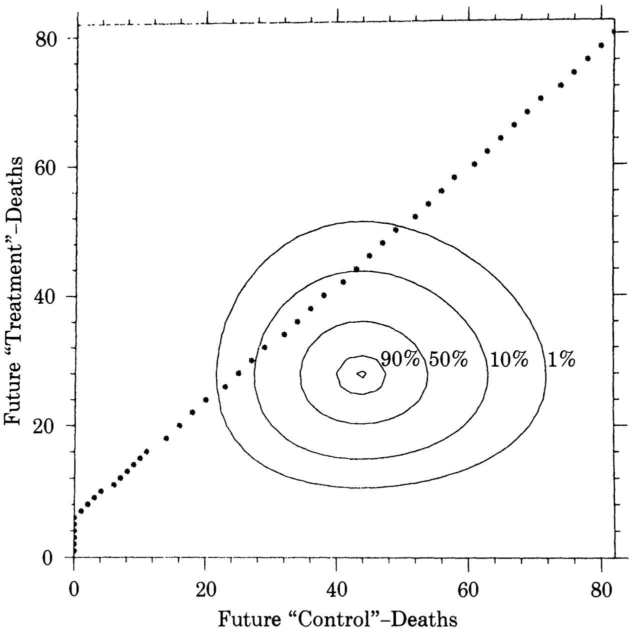

FIG. 6. *Joint predictive distribution of future number of deaths, superimposed on critical region R (lying to the right of the asterisks), where R includes values leading to the rejection of H_0 at the* 5% *level. Contours shown are* 1%, 10%, 50%, 90% *and* 99% *of the maximum ordinate. Reprinted with permission of the publisher from Spiegelhalter, Freedman and Blackburn* (1986).

on T and 67 on C. Figure 4 shows conditional power contours of eventually rejecting the null hypothesis as a function of (P_c, P_t), the true failure probabilities, while Figure 5 shows contours of the posterior distribution of (P_c, P_t) starting with a product of "noninformative" beta distributions as prior. The predictive probability of R is 0.950, which may be decomposed into 0.004 when integrating over the region $P_c < P_t$ (control superior) plus 0.946 when integrating over the region $P_c > P_t$ (treatment superior). The posterior probability that $P_c < P_t$ is 0.009. The predictive distribution of the numbers of future failures in each group (Figure 6) shows that the bulk of the plausible values would lead to rejection of the null hypothesis.

The Bayesian's choice of prior is particularly important in sequential applications. Rosenbaum and Rubin (1984) demonstrate that the sampling distributions of Bayesian coverage probabilities for interval estimation, which are concentrated on the nominal 0.95 if the model is correctly specified, are much more sensitive to misspecification of the prior when the sample size is sequentially determined rather than fixed in advance. Especially when the trial stops early, correct specification is critical if the inferences are to be reasonably well "calibrated."

This reservation notwithstanding, the Bayesian approach offers a considerable advantage in the context of sequential clinical trials by keeping the scientific inferences based on the observed data, as expressed in the likelihood and the posterior distribution, independent of the particular stopping rule used to determine the trial size. Once the data are reported, others are free to analyze them according to their own prior beliefs, which may or may not have led them to make the same stopping decision. This is precisely what happens in practice! For those who are concerned about the problem of sampling to a foregone conclusion (Armitage, 1963), I can do no better than repeat Cornfield's arguments. Such concern belies a prior that gives substantial weight to the null hypothesis. Once this is properly incorporated in the decision procedure, for example by specification of a prior that places a mass of probability at the null, the problem largely evaporates (Cornfield and Greenhouse, 1967).

7. ASSESSING MODEL UNCERTAINTY

A recent review of the literature on benzene and leukemia estimated that for every 1000 men exposed at 10 ppm for a working lifetime of 30 years, 50 would die from leukemia because of the benzene exposure, in addition to a baseline expectation of seven leukemia deaths (Austin, Delzell and Cole, 1988). The authors conclude: "However, this estimate is speculative and whether or not enough confidence can be placed in it to justify a lower occupational benzene standard remains a decision for policy makers." Regrettably, no help was offered the policy makers in assessing the uncertainty inherent in the risk assessment. Austin, Delzell and Cole acknowledged that the available epidemiological data were too sparse, or too subject to confounding, to substantiate a causal association at the low levels (1–10 ppm) of interest to them. Consequently, they followed standard guidelines by extrapolating from higher levels of exposure where there was a demonstrable effect, using a statistical model that is effectively linear at low doses.

These epidemiologists were fortunate to have had human data to work with. The pervasive public fear of cancer and birth defects resulting from exposure to radiation and toxic chemicals, coupled with advances in analytic chemistry that permit detection of trace amounts of chemicals in food, air and water, has led regulatory agencies to rely increasingly on quantitative risk assessments of many agents for which adequate epidemiology is lacking. Results in bioassays with a few hundred animals treated at "maximally tolerated" doses are extrapolated down the dose scale and across species to predict the results of long-term, low-level exposure in humans. It is well known that competing statistical models based on different concepts of toxicology and carcinogenesis can yield roughly equivalent fits to the data in the observable effect dose range, yet lead to low-level risk estimates that are orders of magnitude apart (Brown and Koziol, 1983). Unless a greater consensus can be achieved regarding their proper use and rationale, the shaky scientific foundation for quantitative risk assessments invites attack by critics who may well force their abandonment (Freedman and Zeisel, 1988).

The debate over quantitative risk assessment is a reminder of the subjectivity inherent in many forms of statistical modeling. As Robins and Greenland (1986) put it, "... all modelling strategies contain implicit prior beliefs about nature ... " The rationale for selection of a model for analysis is the belief that the savings in variance afforded by the model assumptions will offset the increase in bias that results from the assumptions being incorrect. The common practice of allowing the data to dominate the model selection process, whether by stepwise entry of variables into a regression equation or evaluation of goodness-of-fit to data in the observable effect range, is often inappropriate. A preferable strategy is to determine which of a broad class of models are reasonably consistent with the observed data and to select among these based on prior understanding of the subject matter. If the lack of such understanding precludes a definitive model choice, it is much wiser to admit this openly and accept the resultant uncertainty than it is to sweep the whole issue under the rug by presenting a single "best" model.

Current approaches to the assessment of model uncertainty involve the fitting of different plausible models in order to measure the "sensitivity" of the inferences, such as the prediction of low-level effects, to model assumptions. A drawback of this approach for decision-making purposes is that it does not provide a quantitative measure of the uncertainty of the prediction. Risk assessors have a hopeless task if faced with a wide range of predicted outcomes and no indication of their relative likelihood. Nor are they well served by a single estimate, based on the best-fitting model whose standard error accounts only for trivial sampling errors. What is needed is a framework that explicitly recognizes model uncertainty and that places greatest weight on those models that have the strongest subjective support among experts in the field, based on their understanding of fundamental mechanisms (Hattis and Smith, 1987).

The Bayesian paradigm provides a natural structure for the synthesis of expert opinion. Model uncertainty is expressed by the experts in the form of a prior distribution on a discrete set of models that are chosen to span a reasonably comprehensive model space. Additional specification of priors for the parameters in each model are required, but in many cases these could be assumed diffuse. Within this framework, Hodges (1987) emphasizes the propagation of model uncertainty through to model-based predictions by use of a predictive distribution that integrates over the model space as well as over the parameter space for each model. Thus, at least in theory, one could arrive at a distribution on the number of excess leukemia deaths observed among the 1000 men exposed to 10 ppm of benzene for 30 years that accounted for all the major elements of uncertainty: structural uncertainty about the model; estimation uncertainty about the parameters in the model; and prediction uncertainty about the future outcomes given the model and parameters (Draper, Hodges, Leamer, Morris and Rubin, 1987). This is obviously an extremely ambitious program that has not yet been employed in any risk assessment that I am aware of. Nonetheless, I believe it is important that some such approach be attempted lest the deficiencies in current guidelines for risk assessment lead to their total abandonment.

8. CONCLUSIONS

Bayes' or empirical Bayes' procedures are useful, or at least promise to become so, for estimation of a multiplicity of related effects, for demonstrating bioequivalence of pharmacologic compounds, for deciding when to stop a sequential clinical trial and for synthesizing expert opinion so as to more adequately express model uncertainty. With the exception of the multiplicity problem, each of these areas involves the explicit use of scientific data for decision-making or regulatory purposes where the introduction of prior beliefs is both natural and unavoidable. It is perfectly appropriate in a democratic society that a carefully quantified measure of expert opinion be used to assist decision makers with their task.

Bayesian methods have a more limited attraction when the goal is scientific description or explanation (Efron, 1986). Exploratory data analysis techniques where uncertainty is assessed via cross-validation or the bootstrap facilitate scientific inferences with a minimum of prior assumptions and a maximum of objectivity. However, more structure is likely to be needed when the statistician starts to confront large, complex datasets and multidimensional problems. The model assumptions that necessarily enter into the process of frequentist inference need to be examined just as critically as the model assumptions, including specification of the prior, that enter into Bayesian inference. Whether one subscribes to Box's (1983) ecumenism or Cox's (1978) eclecticism, it seems clear that Bayesian and frequentist approaches each will have a role to play in biostatistical applications in the years to come.

ACKNOWLEDGMENTS

I am indebted to a large number of individuals whose comments on an earlier draft of this paper led to substantial improvements. These include John Crowley, Susan Ellenberg, Mitchell Gail, Jim Grizzle, Nathan Mantel, Adrian Raftery, Trivellore Raghunathan, Nozer Singpurwalla, Robert Smythe, Michael Stoto, Duncan Thomas and Clare Weinberg. This work was supported in part by Grant CA40644 from the National Cancer Institute.

REFERENCES

Anscombe, F. J. (1963). Sequential medical trials. *J. Amer. Statist. Assoc.* **58** 365–383.

Armitage, P. (1963). Sequential medical trials: Some comments on F. J. Anscombe's paper. *J. Amer. Statist. Assoc.* **58** 384–387.

Armitage, P. (1975). *Sequential Medical Trials*, 2nd ed. Blackwell, Oxford.

Austin, H., Delzell, E. and Cole, P. (1988). Benzene and leukemia: A review of the literature and a risk assessment. *Amer. J. Epidemiol.* **127** 419–439.

BEIR IV (1988). *Health Risks of Radon and Other Internally Deposited Alpha Emitters.* Nat. Acad. Press, Washington.

Berry, D. A. (1987). Interim analysis in clinical trials. The role of the likelihood principle. *Amer. Statist.* **41** 117–122.

Board of Scientific Counselors (1984). Report of the NTP Ad-Hoc Committee on Chemical Carcinogenesis Testing and Evaluation. Dept. Health and Human Services, Washington.

Box, G. E. P. (1983). An apology for ecumenism in statistics. In *Scientific Inference, Data Analysis and Robustness* (G. E. P. Box, T. Leonard and C.-F. Wu, eds.) 51–84. Academic, New York.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, Calif.

Breslow, N. E. (1976). Regression analysis of the log odds ratio. A method for retrospective studies. *Biometrics* **32** 409–416.

Breslow, N. E. and Day, N. E. (1975). Indirect standardization and multiplicative models for rates with reference to the age adjustment of cancer incidence and relative frequency data. *J. Chronic Dis.* **28** 289–303.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research I: The Analysis of Case-Control Studies.* IARC, Lyon.

Brown, C. C. and Koziol, J. A. (1983). Statistical aspects of the estimation of known risk from suspected environmental carcinogens. *SIAM Rev.* **25** 151–181.

Byar, D. P. and Corle, D. K. (1977). Selecting optimal treatment in clinical trials using covariate information. *J. Chronic Dis.* **30** 445–459.

Chaloner, K. M. and Duncan, G. T. (1983). Assessment of a beta prior distribution: PM elicitation. *The Statistician* **32** 174–180.

Chankong, V., Haimes, Y. Y., Rosenkranz, H. S. and Pet-Edwards, J. (1985). The carcinogenicity prediction and battery selection (CPBS) method: A Bayesian approach. *Mutat. Res.* **153** 135–166.

Chernoff, H. and Moses, L. E. (1959). *Elementary Decision Theory.* Wiley, New York.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43** 671–681.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data; Application to cancer of the lung, breast and cervix. *J. Nat. Cancer Inst.* **11** 1269–1275.

Cornfield, J. (1966). Sequential trials, sequential analysis and the likelihood principle. *Amer. Statist.* **20** 18–23.

Cornfield, J. (1969). The Bayesian outlook and its applications (with discussion). *Biometrics* **25** 617–657.

Cornfield, J. (1976). Recent methodological contributions to clinical trials. *Amer. J. Epidemiol.* **104** 408–421.

Cornfield, J. and Greenhouse, S. W. (1967). On certain aspects of sequential clinical trials. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4** 813–829. Univ. California Press.

Cox, D. R. (1975). Prediction intervals and empirical Bayes confidence intervals. In *Perspectives in Probability and Statistics* (J. Gani, ed.) 47–55. Academic, New York.

Cox, D. R. (1978). Foundations of statistical inference: The case for eclecticism. *Austral. J. Statist.* **20** 43–59.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall, London.

Deely, J. J. and Lindley, D. V. (1981). Bayes empirical Bayes. *J. Amer. Statist. Assoc.* **76** 833–841.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

Dempster, A. P., Rubin, D. B. and Tsutakawa, R. K. (1981). Estimation in covariance component models. *J. Amer. Statist. Assoc.* **76** 341–353.

Dempster, A. P., Selwyn, M. R. and Weeks, B. J. (1983). Combining historical and randomized controls for assessing trends in proportions. *J. Amer. Statist. Assoc.* **78** 221–227.

Draper, D., Hodges, J. S., Leamer, E. E., Morris, C. N. and Rubin, D. B. (1987). A research agenda for assessment and propagation of model uncertainty. The Rand Corp., Santa Monica, Calif.

DuMouchel, W. H. and Harris, J. E. (1983). Bayes methods for combining the results of cancer studies in humans and other species (with discussion). *J. Amer. Statist. Assoc.* **78** 293–315.

Dunnett, C. W. and Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 × 2 tables. *Biometrics* **33** 593–602.

Efron, B. (1982). *The Jackknife, the Bootstrap, and other Resampling Plans.* SIAM, Philadelphia.

Efron, B. (1986). Why isn't everyone a Bayesian? *Amer. Statist.* **40** 1–11.

Efron, B. and Morris, C. (1973). Combining possibly related estimation problems (with discussion). *J. Roy. Statist. Soc. Ser. B* **35** 379–421.

Elston, R. C. and Grizzle, J. E. (1962). Estimation of time-response curves and confidence bands. *Biometrics* **18** 148–159.

Flüher, H., Grieve, A. P., Mandallay, D., Mau, J. and Moser, H. A. (1983). Bayesian approach to bioequivalence assessment: An example. *J. Pharm. Sci.* **72** 1178–1181.

Freedman, D. A. and Zeisel, H. (1988). From mouse to man: The quantitative assessment of cancer risks (with discussion). *Statist. Sci.* **3** 3–56.

Freedman, L. S. and Spiegelhalter, D. J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *The Statistician* **32** 153–160.

Gastwirth, J. L. (1987). The statistical precision of medical screening procedures: Application to polygraph and AIDS antibodies test data (with discussion). *Statist. Sci.* **2** 213–238.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

Greenland, S. and Thomas, D. C. (1982). On the need for the rare disease assumption in case-control studies. *Amer. J. Epidemiol.* **116** 547–553.

Grizzle, J. E. and Allen, D. M. (1969). Analysis of growth and dose response curves. *Biometrics* **25** 357–381.

Halperin, M., Lan, K. K. G., Ware, J. H., Johnson, N. J. and DeMets, D. L. (1982). An aid to data monitoring in long-term clinical trials. *Controlled Clin. Trials* **3** 311–323.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72** 320–340.

Haseman, J. K. (1984). Statistical considerations in the design, analysis and interpretation of animal carcinogenicity studies. *Environ. Health Perspect.* **58** 385–392.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models (with discussion). *Statist. Sci.* **1** 297–318.

Hattis, D. and Smith, J. A. (1987). What's wrong with quantitative risk assessment? In *Biomedical Ethics Review* (J. Humber and R. F. Almeder, eds.) 57–106. Humana, Clifton, N.J.

Hauck, W. W. and Anderson, S. (1986). A proposal for interpreting and reporting negative studies. *Statist. in Medicine* **5** 203–209.

Herson, J. (1979). Predictive probability early termination plans for phase II clinical trials. *Biometrics* **35** 775–783.

Hodges, J. S. (1987). Uncertainty, policy analysis and statistics (with discussion). *Statist. Sci.* **2** 259–291.

Hui, S. L. and Berger, J. O. (1983). Empirical Bayes estimation of rates in longitudinal studies. *J. Amer. Statist. Assoc.* **78** 753–760.

Jennison, C. and Turnbull, B. W. (1989). Interim analyses: The repeated confidence interval approach (with discussion). *J. Roy. Statist. Soc. Ser. B* **51** 305–361.

Kalbfleisch, J. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* Wiley, New York.

Kass, R. E. (1983). Comment on "Bayes methods for combining the results of cancer studies in humans and other species," by W. H. DuMouchel and J. E. Harris. *J. Amer. Statist. Assoc.* **78** 312–313.

Kass, R. E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Amer. Statist. Assoc.* **84** 717–726.

Laird, N. M. and Louis, T. A. (1987). Empirical Bayes' confidence intervals based on bootstrap samples (with discussion). *J. Amer. Statist. Assoc.* **82** 739–757.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38** 963–974.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70** 659–663.

Lee, J. C. and Geisser, S. (1975). Applications of growth curve prediction. *Sankhyā Ser. A* **37** 239–256.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22.

Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2, Inference.* Cambridge Univ. Press, Cambridge.

Mantel, N. and Stark, C. R. (1968). Computation of indirect adjusted rates in the presence of confounding. *Biometrics* **24** 997–1005.

Manton, K. G., Stallard, E., Woodbury, M. A., Riggan, W. B., Greason, J. P. and Mason, T. J. (1987). Statistically adjusted estimates of geographic mortality profiles. *J. Nat. Cancer Inst.* **78** 805–815.

Manton, K. G., Woodbury, M. A., Stallard, E., Riggan, W. B., Creason, J. P. and Pellam, A. C. (1989). Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates. *J. Amer. Statist. Assoc.* **84** 637–650.

Mason, T. J. and McKay, F. W. (1973). *U.S. Cancer Mortality by County: 1950–1969.* DHEW Publication No. (NIH) 74-615, Government Printing Office, Washington.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models.* Chapman and Hall, London.

Meng, C. Y. K. and Dempster, A. P. (1987). A Bayesian approach to the multiplicity problem for significance testing with binomial data. *Biometrics* **43** 301–311.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and application. *J. Amer. Statist. Assoc.* **78** 47–55.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35** 549–556.

O'Sullivan, F., Yandell, B. S. and Raynor, W. J. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81** 96–103.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58** 545–554.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64** 191–199.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* **51** 313–326.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–411.

Racine, A., Grieve, A. P., Flüher, H. and Smith, A. F. M. (1986). Bayesian methods in practice: Experiences in the pharmaceutical industry. *Appl. Statist.* **35** 93–120.

Racine-Poon, A., Grieve, A. P., Flüher, H. and Smith, A. F. M. (1987). A two-stage procedure for bioequivalence studies. *Biometrics* **43** 847–856.

Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika* **52** 447–458.

Rao, C. R. (1975). Simultaneous estimation of parameters in different linear models and applications to biometric problems. *Biometrics* **31** 545–554.

Rao, C. R. (1987). Prediction of future observations in growth curve models (with discussion). *Statist. Sci.* **2** 434–471.

Robins, J. M. and Greenland, S. (1986). The role of model selection in causal inference from nonexperimental data. *Amer. J. Epidemiol.* **123** 392–402.

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Sensitivity of Bayes inference with data-dependent stopping rules. *Amer. Statist.* **38** 106–109.

SELWYN, M. R., DEMPSTER, A. P. and HALL, N. R. (1981). A Bayesian approach to bioequivalence for the 2 × 2 changeover design. *Biometrics* **37** 11–21.

SELWYN, M. R. and HALL, N. R. (1984). On Bayesian models for bioequivalence. *Biometrics* **40** 1103–1108.

SIEMIATYCKI, J., GERIN, M., RICHARDSON, L., HUBERT, J. and KEMPER, H. (1982). Preliminary report of an exposure-based, case-control monitoring system for discovering occupational carcinogens. *Teratogenesis Carcinog. Mutagen* **2** 169–177.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

SIMON, R. (1977). Adaptive treatment assignment methods and clinical trials. *Biometrics* **33** 743–749.

SIMON, R. (1982). Patient subsets and variations in therapeutic efficacy. *Br. J. Clin. Pharmacol.* **14** 473–482.

SMITH, A. F. M., SKENE, A. M., SHAW, J. E. H. and NAYLOR, J. C. (1987). Progress with numerical and graphical methods for practical Bayesian statistics. *The Statistician* **36** 75–82.

SMITH, A. F. M., SKENE, A. M., SHAW, J. E. H., NAYLOR, J. C. and DRANSFIELD, M. (1985). The implementation of the Bayesian paradigm. *Comm. Statist. A—Theory Methods* **14** 1079–1102.

SPIEGELHALTER, D. J., FREEDMAN, L. S. and BLACKBURN, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clin. Trials* **7** 8–17.

TAKASUGI, M., TERASAKI, P. I., HENDERSON, B., MICKEY, M. R., MENCK, H. and THOMPSON, R. W. (1973). HL-A antigens in solid tumors. *Cancer Res.* **33** 648–650.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–540.

THOMAS, D. C., GOLDBERG, M., DEWAR, R. and SIEMIATYCKI, J. (1986). Statistical methods for relating several exposure factors to several diseases in case-heterogeneity studies. *Statist. in Medicine* **5** 49–60.

THOMAS, D. C., SIEMIATYCKI, J., DEWAR, R., ROBINS, J., GOLDBERG, M. and ARMSTRONG, B. G. (1985). The problem of multiple inference in studies designed to generate hypotheses. *Amer. J. Epidemiol.* **122** 1080–1095.

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86.

TSIATIS, A. A., ROSNER, G. L. and MEHTA, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40** 797–803.

TSUTAKAWA, R. K. (1988). Mixed model for analyzing geographic variability in mortality rates. *J. Amer. Statist. Assoc.* **83** 37–42.

TUKEY, J. W. (1977). Some thoughts on clinical trials, especially problems of multiplicity. *Science* **198** 679–684.

WESTLAKE, W. J. (1979). Statistical aspects of comparative bioavailability trials. *Biometrics* **35** 273–280.

WHITEHEAD, J. (1983). *The Design and Analysis of Sequential Clinical Trials.* Ellis Horwood, Chichester.

ZEGER, S. L. and LIANG, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 1–10.

Comment

Peter Armitage

This persuasive paper should be welcomed by all biostatisticians, not least because the author succeeds in conveying his enthusiasm for (although to some extent his reservations about) Bayesian analysis without indulging in the Messianic fervor so characteristic of some of its proponents. For my part I am convinced that Bayesian methods have a major role to play in the analysis of biomedical data, although I am as skeptical about claims that they provide an all-embracing "world-view" of statistics as I am about similar claims in the realms of politics, art or religion.

Since Dr. Breslow starts with some fascinating autobiographical detail, it may not be out of place to add a few personal comments. When I entered medical statistics in 1947, the discipline was still struggling to take on board the pre-war advances of Fisher and his contemporaries. In Britain, J. O. Irwin was, among biostatisticians, almost a lone representative of the Fisher–Neyman traditions; in the United States, Cochran had yet to enter biostatistics and take on his important leading role. (The developments of the 1920s and 1930s were, of course, more deeply established in agricultural research.) Jeffreys appeared as a lone figure of great stature but almost completely lacking in influence. Bayesian methods were propounded, in the U.K., by a few people, including W. Perks, an actuary, and I. G. Good, but to little effect, and it was not until the appearance of L. J. Savage's book in 1954 that more than a handful of statisticians took Bayes seriously.

In the gradual process of consolidating the use of "standard" methods, most of us gave little thought to the apparently more formidable task of introducing Bayesian inference and decision theory. I must have been one of the English statisticians, during Norman

Peter Armitage is Emeritus Professor of Applied Statistics at the University of Oxford. His mailing address is 71 High Street, Drayton, Abingdon, Oxon OX14 4JW, England.