exposure and disease incidence. Controversies abound in the applications of the case–control methodology due to questions on the comparability of cases and controls. These studies are easy to fault.

The logic of Bayesian methods for case–control studies is overwhelming. Aspects of this methodology and point-of-view are discussed in Zelen and Parker (1986).

Rejoinder

Norman Breslow

My choice of the topic "Biostatistics and Bayes" for the ASA session on historical perspectives and new directions in biometry was motivated by several concerns: to attempt to avoid the tedium that often accompanies "past, present and future" overviews; to use the opportunity to undertake some remedial self education in an area that I had long neglected; and to simulate both myself and my colleagues to glance up from the project work that often so completely absorbs us to consider some broader issues. Having anticipated criticism from both biostatisticians and Bayesians for daring to comment on a controversial topic to which I had personally contributed no research work, I was pleased with the generally supportive remarks of the discussants and perhaps even a little disappointed that I was not called more severely to task. I appreciate their efforts in contributing to the discussion and I thank the editors for their indulgence and encouragement.

PROBLEMS WITH MANY PARAMETERS

Professor Armitage and Dr. Jennison, while generally sympathetic to the use of (empirical) Bayes methods for problems involving large numbers of unknown but related parameters, both remark that classical random effects models are available outside the Bayesian framework. Reduction of the estimation problem to one involving a small number of parameters that describe means and variance components indeed is a necessary first step. However, in the standard set-up, the "random effects" often are considered more of a nuisance that complicate inferences on the main effects. The appealing feature of the Bayes approach is its focus on joint shrinkage estimation of the original parameters, i.e., the random effects, for presentation in maps or reconstructed images. Use of the Bayesian paradigm as a technical tool for sharpening estimates of a large number of parameters considered in the aggregate seems well established and noncontroversial. Viewed from this perspective, Bayesian methods may have an even greater role to play in exploratory data analysis than Dr. Flühler would acknowledge. It is only when one needs to single out one of the estimated parameters for special inference and decision that greater caution in assessing model assumptions, including specification of the prior, is called for. This underlies my own hesitation in fully embracing the use of Bayes techniques for cancer risk assessment and, evidently, Professor Armitage's hesitation in applying them to bioassays of pharmaceutical compounds.

BIOEQUIVALENCE

I am grateful to Drs. Spiegelhalter and Freedman for helping, both in their commentary and in several published articles, to clarify my thinking about bioequivalence, especially in the context of clinical trials. They correctly emphasize the need to keep the specification of the indifference region, or range of equivalence, quite separate from consideration of prior beliefs about the actual difference between treatments. Westlake, Armitage and others have shown how the use of confidence intervals superimposed on the line graph of preference and indifference regions provides a reasonable frequentist solution to the problem of inferring equivalence in simple problems. In reply to Dr. Jennison, I did not mean to imply that adequate frequentist solutions could not be found more generally. I agree with Drs. Spiegelhalter and Freedman, however, that the Bayesian interpretation in terms of the posterior probability of equivalence is more natural and, knowing how they are inclined to think of confidence intervals, feel nearly certain that my medical associates would agree also.

To amplify on Dr. Flühler's remarks about the role of Bayesian thinking at the design stage, I would again like to borrow from Spiegelhalter and Freedman's commentary and suggest to Dr. Jennison that his OC curve be averaged with respect to a plausible prior distribution for the treatment differences in order to decide if the proposed study has any hope of achieving its goal. The prior used for this calculation, which would indicate the pharmaceutical company's belief in the efficacy of its new drug, might well differ

substantially from the flat prior selected by the regulatory agency for use in a bioequivalence evaluation protocol.

CASE–CONTROL STUDIES

I particularly appreciated Professor Zelen's contribution on a topic not mentioned by the other discussants: case–control studies. Having long worked to develop statistical methods for case–control and cohort studies in epidemiology, admittedly from a frequentist perspective but with heavy emphasis on likelihood concepts, I was keenly interested in his observation that Bayesian methods had a greater role to play. On balance, the case he makes seems not unlike that presented by Dempster and colleagues in advocating the incorporation of historical control information into analysis of the rodent bioassay. Results of the case–control study are strengthened when the exposure frequency in the controls turns out to be similar to that observed in other surveys conducted in the general population, whereas the exposure frequency in the cases is elevated. The extreme example used by Zelen and Parker to illustrate their method occurred when investigators set out to determine the cause of adenocarcinoma of the vagina in young women, a disease that was virtually unknown before a cluster of cases occurred during the late 1960s (Herbst, Ulfelder and Poskanzer, 1971). Interviews established that the mothers of seven of the eight cases had been treated for infertility with the hormone diethylstilbestrol (DES) during the first trimester of pregnancy, whereas none of the mothers of 32 matched controls were so treated. Such treatment was known to be rare in the general population. While one could argue, as Professor Zelen has, that controls are not needed in such circumstances, this was only obvious in retrospect, after the mothers of both cases and controls had been interviewed and the data compiled. Had some other more common exposure turned out to have caused the disease, although this is unlikely since the agent would have had to be introduced rapidly on a large scale in order to explain the jump in incidence, the controls would have had greater relevance to the inference.

An alternative role for Bayesian methodology in the analysis of case–control data is in combining information on an exposure/disease association from several case–control studies. Standard approaches include log–linear and Mantel–Haenszel methodology, where the relative risk is assumed constant from study to study. However, it may be more realistic to posit some random variation in the relative risk, due for example to variation in unmeasured cofactors in the different study populations, in which case one will want a different weighting scheme for the combined estimate. My colleague, Dr. Raghunathan, who has been most helpful in giving me guidance on the Bayesian perspective, is developing new methods of combining information from 2×2 tables that incorporate such variation and that focus on realistic assessment of the uncertainty in the "random effect" estimated for each table.

Although Professor Zelen may be correct that many case–control studies do not employ random samples of cases and controls, this is not in my opinion true of the best and most influential studies. My epidemiology colleagues do their utmost to enroll *all* cases of disease diagnosed in a defined geographic area during a defined time period in their studies and to choose controls at random from this same population by random digit dialing or similar techniques. The rationale, as elaborated at length in my 1980 research monograph with Dr. N. Day, is to regard the defined population as a cohort that is being followed forward in time, and the case–control study as a method of outcome dependent sampling from the cohort. Of course, one must still contend with the issue of confounding, the fact that persons exposed to the risk factor in question may differ with respect to other causal factors (e.g., age) from those who are not exposed. Hence age adjustments are generally made.

In hospital-based studies, one tries to approximate the random sampling of controls by selecting patients with other diseases. Epidemiologists are generally very careful to select control diseases believed to be unrelated to the exposure under study, so that the age-specific exposure distribution for each control disease group may reasonably be assumed equal to the age-specific exposure distribution in the general population. Rosenbaum (1987) provides a nice formal proof that this assumption justifies application of standard age-adjustment procedures and he advocates testing the equality of the age-specific exposure distributions in the different control groups as a means of partially verifying it. The inference is strengthened when the control groups are known to differ with respect to suspected, unmeasured confounders. (Rosenbaum may have overreacted to my suggestion that many epidemiologists were already familiar with these basic principles.)

SEQUENTIAL CLINICAL TRIALS

It was perhaps inevitable that the commentaries would concentrate heavily on the issue of sequential trials since it is here that the lines between Bayesian and frequentist inference are most sharply drawn. Furthermore, the ethical dilemma brings an added emotional dimension to the debate (Ware, 1989). While human lives are also at stake in the approval of drug therapies for life-threatening disease or the

regulation of toxic chemicals in the workplace, decisions based on statistical analysis of medical data somehow seem more urgent when carried out in close proximity to the patient's bedside.

Dr. Jennison has elaborated on several issues that were only briefly alluded to in my paper. In particular, his equation (3) nicely illustrates the fact that nominal 5% significance tests, or their Bayesian analogs, yield high probability of rejection under the null hypothesis if carried out repeatedly. Indeed they will be certain to do so ("sampling to a forgone conclusion") if repeated indefinitely. Both Jennison and Armitage remark that frequency properties of the resulting sequential test are of considerable concern to them. Not even Spiegelhalter and Freedman are happy with my citation of Cornfield's response to this problem, namely assignment of a prior probability mass to the point null hypothesis. As I believe was evident from the statement that I "could do no better" than repeat such arguments, I must admit to some residual doubts of my own on this score. Nevertheless, re-reading of the literature on frequentist versus Bayesian p-values (e.g., Berger and Sellke, 1987) suggests to me that the Bayes factors (Cornfield's relative betting odds) that arise in converting such prior probabilities into posterior ones may have something to offer as an informal way of assessing the strength of the current evidence for or against the null hypothesis.

Where I part company with Jennison and with many of my colleagues is in the degree to which the specification of the statistical design, whether from a Bayesian or a frequentist viewpoint, is to be formalized as a statistical decision problem. As stated in the paper, I feel that the most ethical way to evaluate two medical treatments, whose long-term therapeutic and toxic effects are of unknown relative merit, is by means of a suitably large, randomized and controlled trial carried to conclusion in a representative sample of patients thought to be eligible for such treatment. The results of such a trial justifiably stand the greatest chance of being accepted as a basis for treatment selection by the medical community and thus, ultimately, of benefiting the larger population of patients. Interim trial data should, of course, be used by the monitoring committee together with all other available information to determine whether the trial should be curtailed because of large, unanticipated therapeutic or toxic effects, or for other reasons. Informal use of posterior distributions based on conservative (null biased) priors, as advocated by Spiegelhalter and Freedman, or of Bayes factors as advocated by Cornfield, should help in the decision as to whether the evidence in favor of one or the other treatment was so overwhelming as to justify early termination. In conducting such an informal assessment of the evidence, it would be useful to consider the sensitivity of the posterior distribution to a variety of priors (Kass and Greenhouse, 1989). As noted by Dr. Flühler, graphical presentation of such information is appreciated by all members of monitoring committees, not just by the statisticians. Generally one would demand stronger evidence for early termination than that needed to make a firm choice between treatments at the trial's conclusion, when follow-up was complete and full information about long-term toxicity as well as more immediate therapeutic effects was available. In this respect, the practical outcome might be little different from that of the widely used O'Brien–Fleming boundary (see Spiegelhalter and Freedman's second figure) where early Z statistics must be extreme to justify termination.

My viewpoint on these issues undoubtedly has been influenced by my involvement in long-term trials of chronic disease where the differences between treatments generally have not been very dramatic and where there has been ample time during the course of the trial for new information to modify one's thinking about the design specifications. A feature of the Wilms' tumor trial that I failed to mention earlier is that the ultimate decision to terminate the trial was not because of the proven efficacy of radiation therapy, as had been anticipated at the outset, but rather because increasing awareness of the long-term sequelae of such therapy led to the belief that some patients were being irradiated needlessly. While it certainly would be desirable to have all the details about choice of relevant end points and similar matters tidied up at the outset, in practice it is not always possible to achieve this. I started by research career working on frequentist sequential tests for medical trials (Breslow, 1969, 1970), but experiences with their application in the setting described (Breslow, 1978) eventually led me to abandom this effort in favor of other pursuits that seemed more relevant to the real scientific issues. I must thus plead guilty to Dr. Jennison's charge that my remarks on the difficulty of computing OC curves for sequential tests are outdated.

CONCLUSIONS

Having identified myself as someone who is sympathetic in principle with the application of Bayesian methods to problems that involve the use of scientific data for (informal) decision making, but who has not as yet actually put such methods into practice, should I be labeled a Bayesian? In view of the aphorism that actions speak louder than words, this question undoubtedly will be better answered by reference to future articles reporting clinical trial results and the like in the medical literature than by reference to methodological essays in the statistics literature. Like most statisticians, I am eager to try out new methods

on old and well understood problems provided that it is reasonably convenient to do so. My acceptance of those methods is likely to depend more on their performance in such real world settings than on philosophical dogma. From this perspective, I believe that progress in the application of Bayesian statistics depends critically on the development and distribution of well-documented, "user-friendly" computational tools that assist in the specification or elicitation of prior distributions and their conversion into posteriors, with ample provision for graphical display. Reference to work-in-progress along these lines by several research teams was made earlier, and I await the release of the final products with considerable anticipation.

ADDITIONAL REFERENCES

ARMITAGE, P. (1988). Some aspects of phase III trials. In *Biometry, Clinical Trials and Related Topics* (T. Okuno, ed.) 1–16. Excerpta Medica, Amsterdam.

ARMITAGE, P. (1989). Inference and decision in clinical trials. *J. Clin. Epidemiol.* **42** 293–299.

BERGER, J. O. and SELLKE, T. (1987). Testing a point null hypothesis: The irreconcilability of p-values and evidence (with discussion). *J. Amer. Statist. Assoc.* **82** 112–139.

BERRY, D. A. (1985). Interim analyses in clinical trials: Classical vs Bayesian approaches. *Statist. in Medicine* **4** 521–526.

BESAG, J. (1986). On the statistical analysis of dirty pictures (with discussion). *J. Roy. Statist. Soc. Ser. B* **48** 259–302.

BHAT RESEARCH GROUP (1984). The Beta-blocker heart attack trial: Design, methods and baseline results. *Controlled Clin. Trials* **5** 382–437.

BHAT RESEARCH GROUP (1987). The Beta-blocker heart attack trial: Recruitment experience. *Controlled Clin. Trials* **8** 79S–85S.

BRESLOW, N. (1969). On large sample sequential analysis with applications to survivorship data. *J. Appl. Probab.* **6** 261–274.

BRESLOW, N. (1970). Sequential modification of the UMP test for binomial probabilities. *J. Amer. Statist. Assoc.* **65** 639–648.

BRESLOW, N. (1978). Perspectives on the statistician's role in cooperative clinical research. *Cancer* **41** 326–332.

BROWN, L. D., COHEN, A. and STRAWDERMAN, W. E. (1980). Complete classes for sequential test of hypotheses. *Ann. Statist.* **8** 377–398.

EALES, J. D. and JENNISON, C. (1990). An improved method for deriving optimal one-sided group sequential tests. In preparation.

FAIRBANKS, K. and MADSEN, R. (1982). P values for tests using a repeated significance testing design. *Biometrika* **69** 69–74.

FREEDMAN, L. S. and SPIEGELHALTER, D. J. (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled Clin. Trials* **10** 357–367.

FREI, A., COTTIER, C., WUNDERLICH, P. and LUDIN, E. (1987). Glycerol and Dextran combined in the therapy of acute stroke. *Stroke* **18** 373–379.

GELFAND, A. E., HILLS, S. E., RACINE, A. and SMITH, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. Unpublished manuscript.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis Machine Intell.* **6** 721–741.

GREEN, P. J., JENNISON, C. and SEHEULT, A. H. (1985). Analysis of field experiments by least squares smoothing. *J. Roy. Statist. Soc. Ser. B* **47** 299–315.

GRIEVE, A. P. (1988). Some uses of predictive distributions in pharmaceutical research. In *Biometry, Clinical Trials and Related Topics* (T. Okuno, ed.). Excerpta Medica, Amsterdam.

HEDAYAT, A. S., JACROUX, M. and MAJUNDAR, D. (1988). Optimal designs for comparing test treatments with controls (with discussion). *Statist. Sci.* **3** 462–491.

HERBST, A. L., ULFELDER, H. and POSKANZER, D. C. (1971). Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women. *New England J. Med.* **284** 878–881.

JENNISON, C. (1987). Efficient group sequential tests with unpredictable group sizes. *Biometrika* **74** 155–165.

JENNISON, C. and TURNBULL, B. W. (1983). Confidence intervals for a binomial parameter following a multistage test with applications to MIL-STD 105D and medical trials. *Technometrics* **25** 49–58.

KASS, R. E. and GREENHOUSE, J. B. (1989). Comment: A Bayesian perspective on "Investigating therapies of potentially great benefit: ECMO," by J. H. Ware. *Statist. Sci.* **4** 310–317.

LEONARD, T. (1983). Some philosophies of inference and modelling. In *Scientific Inference, Data Analysis, and Robustness* (G. E. P. Box, T. Leonard and C. F. Wu, eds.). Academic, New York.

MANDALLAZ, D. and MAU, J. (1981). Comparison of different methods for decision-making in bioequivalence assessment. *Biometrics* **37** 213–222.

MEHTA, C. R. and CAIN, K. C. (1984). Charts for early stopping of pilot studies. *J. Clin. Oncology* **2** 676–682.

O'QUIGLEY, J. and BAUDOIN, C. (1988). General approaches to the problem of bioequivalence. *The Statistician* **37** 51–58.

POCOCK, S. J. and HUGHES, M. D. (1990). Stopping rules, estimation problems, and reporting bias in clinical trials. *Controlled Clin. Trials*. To appear.

RACINE-POON, A. (1985). A Bayesian approach to nonlinear random effects models. *Biometrics* **41** 1015–1023.

RACINE-POON, A. and SMITH, A. F. M. (1990). Population models. In *Statistical Methodology in the Pharmaceutical Sciences* (D. A. Berry, ed.). Dekker, New York.

ROSENBAUM, P. R. (1987). The role of a second control group in an observational study (with discussion). *Statist. Sci.* **2** 292–316.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

SMITH, A. F. M. (1986). Some Bayesian thoughts on modelling and model choice. *The Statistician* **35** 97–102.

SPIEGELHALTER, D. J. and FREEDMAN, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective opinion. *Statist. in Medicine* **5** 1–13.

SPIEGELHALTER, D. J. and FREEDMAN, L. S. (1988). Bayesian approaches to clinical trials. In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 453–477. Oxford Univ. Press, Oxford.

WARE, J. H. (1989). Investigating therapies of potentially great benefit: ECMO (with discussion). *Statist. Sci.* **4** 298–340.

WEIHS, C. and SCHMIDLI, H. (1990). OMEGA—Online Multivariate Exploratory Graphical Analysis: Routine searching for structure (with discussion). *Statist. Sci.* **5** 175–226.

WESTLAKE, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics* **32** 741–744.

ZELEN, M. and PARKER, R. A. (1986). Case-control studies and Bayesian inference. *Statist. in Medicine* **5** 261–269.