with data bases: analyses of comprehensive data bases are not subject to publication biases.) Such a system is both ethically and scientifically sound.

CONCLUSIONS

1. Randomization is not essential for scientific inference.
2. Randomized clinical trials are inherently unethical. They are not appropriate for life-threatening conditions.
3. Clinical equipoise is an invention used to avoid difficult ethical questions.
4. Randomized consent is unethical by its nature.
5. It is possible to learn in a clinical setting and still deliver good medicine.
6. Analysis of clinical trials should use all available information, including historical controls.
7. Analysis of clinical trial data should use all available covariates, whether or not the trial was randomized.
8. Neyman-Pearson inference, in which the analysis is tied irrevocably to the design, is impractical and sometimes unworkable.
9. Bayesian inferences apply at any time during or after a study; the course of a study can be dictated by interim Bayesian calculations which weigh the costs and benefits (in terms of good medical treatment) of the various options.
10. Medical research should move away from randomized trials and toward establishing comprehensive patient registries.

ACKNOWLEDGMENTS

My discussion benefited from the input of many people; I especially thank David Lane, Tom Louis and Janis Hardwick for their suggestions.

ADDITIONAL REFERENCES

BERRY, D. A. (1989). Monitoring accumulating data in a clinical trial. *Biometrics.* To appear.

BERRY, D. A. and FRISTEDT, B. (1985). *Bandit Problems: Sequential Allocation of Experiments.* Chapman and Hall, New York.

ELLWOOD, P. M. (1988). Shattuck lecture—Outcomes management: A technology of patient experience. *New England J. Med.* **318** 1549–1556.

TOOMASIAN, J. M., SNEDECOR, S. M., CORNELL, R. G., CILLEY, R. E. and BARTLETT, R. H. (1988). National experience with extracorporeal membrane oxygenation for newborn respiratory failure: Data from 715 cases. *ASAIO Trans.* **34** 140–147.

Comment: A Bayesian Perspective

Robert E. Kass and Joel B. Greenhouse

Ever since the first modern randomized clinical trial (RCT), clinicians and statisticians have struggled with the question of whether it is proper to deny a patient some possibly beneficial treatment for the sake of conducting an experiment. Even as Sir A. Bradford Hill made his influential arguments in favor of RCTs, he emphasized the importance of ethical considerations. They are, Hill (1951) said, ". . . paramount and must never, on any scientific grounds whatever, be lost sight of. If a treatment cannot ethically be withheld then clearly no controlled trial can be instituted." The problem, however, is to define the circumstances under which "a treatment cannot ethically be withheld." Hill (1951, 1953) distinguished the "dramatic" situations, in which a treatment might offer a cure for an otherwise invariably fatal disease, from the "more mundane" in which a treatment might produce a decline in mortality from, say, 15 to 10 per cent. The dramatic cases might not require a concurrent control group, but, he argued, the more common investigations could provide reliable information only through the use of RCTs.

As Professor Ware has clearly shown in the case of ECMO, the most difficult situation involves a disease that is not invariably fatal, yet the therapy is potentially of great benefit. The basic issue is whether such cases should be considered to be like the "dramatic" ones, or like the "more mundane," or whether, perhaps, there is an intermediate classification in which some third method of study, such as adaptive allocation, should be used.

In some respects, the trial Ware describes is like another that raised considerable debate by using an RCT to examine the effectiveness of Ara-A, an antiviral agent, in the treatment of herpes simplex viral encephalitis, a disease that had a historical fatality rate of around 70%. In that case, McCartney (1978) argued that none of the usual justifications for RCTs

Robert E. Kass is Associate Professor and Joel B. Greenhouse is Associate Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-2717.

applied, and effectively put that situation into Hill's "dramatic" classification. Even with the apparent 80% historical fatality rate of PPHN, however, the case of ECMO differs from Ara-A in one very important way: Ware expressed grave concern about the potential morbidity of ECMO treatment, especially brain hemorrhage and subsequent severe impairment. Thus, it seemed quite plausible to Ware and his colleagues that there was a significant risk of permanent brain damage from the ECMO therapy. From this position, they concluded that the historical data "were not sufficient to justify routine use of ECMO in the treatment of PPHN ..." and they were "... uneasy about rapid acceptance of a new and potentially dangerous technology ..." without good information from a well-designed RCT.

Ware's paper raises many important issues and will undoubtedly produce extensive discussion. We do three things in our commentary. First, in Section 1, we suggest a definition of conditions under which randomization is ethically justifiable, and we indicate how the definition may be implemented from a Bayesian point of view. Next, we follow Ware in assuming that randomization was ethically justifiable and, in Section 2, we consider the information available at the completion of the first stage. An interesting feature of Ware's discussion is his use of Bayesian methods to assess the evidence at that point. One important reason for using Bayesian methods in this context is that they do not require accounting for the design employed in obtaining the data. Although Ware considers the Bayesian testing problem, we prefer to analyze the data via estimation. We consider prior distributions suggested by the assumption that randomization was ethically justifiable, and then compute relevant marginal posterior probabilities. We then go on to note the subtlety of the Bayesian testing problem, and we illustrate what we consider to be an appropriate approach to it. Finally, in Section 3, we mention our hope that methodology for examining historical information could be developed and more widely applied to problems in which there are substantial ethical difficulties with RCTs.

1. ETHICAL BASIS FOR RANDOMIZATION

In this section, we offer a perspective on the fundamental problem of defining the circumstances under which randomization is ethically justifiable. Part of our purpose is to provide a framework for Bayesianly oriented methodology. We should note right away that we will not discuss the deep and difficult problem of explaining convincingly why randomization should be used to ensure comparability among treatment groups. (See Kadane and Seidenfeld, 1989, for a recent discussion and references.) Instead, we take up the problem of defining conditions under which ethical concerns would not preclude randomization. This is what we will mean when we say that randomization is "ethically justifiable." For simplicity, we assume the trials we discuss are, like the ECMO trial, designed to compare a "treatment" with a "control," without consideration of covariates. Also, in the ECMO trial there are important issues involving cost and the allocation of scarce resources, but we ignore such considerations here.

We will try to motivate our suggestion by linking it with an observation made by Hill (1953) that the difficulty of the dilemma depends on two things, the severity of the disease and the state of uncertainty about the effectiveness of the treatment. "Where life and death (or serious after-effects) are not at issue the problem is clearly eased. It is also eased, more often than not, by the state of our ignorance. For, frequently, we have no scientific evidence that a particular treatment will benefit the patients and ... we are often, willy-nilly, experimenting upon them." In our statement of conditions under which randomization is ethically justifiable, we incorporate these two features of the ethical problem, referring to "cautiousness" about the decision to apply a treatment. When the disease is less severe or the evidence is poor, it becomes more acceptable to wait for better information, which, in the sense we use the word, will mean it becomes appropriate to exercise greater caution in selecting the treatment.

Our basic conception is motivated by the presumption that the purpose of a trial is to collect data that bring to conclusive consensus at termination opinions that had been diverse and indecisive at the outset. When there are diverse opinions among knowledgeable and thoughtful observers, however, it is because different people attach different degrees of importance to various pieces of information concerning the merits of the treatment. In this situation, we may articulate the position of a reasonable skeptic who may recognize that some historical evidence is relevant and may consider plausible certain theoretical arguments on behalf of the treatment, but at the same time appreciates the dangers of adverse reactions and knows that medical history is littered with many false claims of success. We suppose that this reasonable skeptic is cautious with regard to coming to conclusions about the superiority of the treatment or the control. That is, the cautious reasonable skeptic will recommend action on the basis of fairly firm knowledge, but not otherwise, with the degree of caution exhibited depending on the quality of available information and the seriousness of the disease. We then arrive at our understanding of what constitutes the basis for an ethical trial: *Randomization is ethically justifiable when a cautious reasonable skeptic would be unwilling*

to state a preference in favor of either the treatment or the control.

This formulation is intentionally vague. Perhaps the simplest way to make it precise would be to represent the skeptic's beliefs by a probability distribution and then suppose that no preference will be stated unless the probability that the treatment is superior is either less than p_* or greater than p^* for some specified values of p_* and p^* with $0 < p_* < .5 < p^* < 1$. Alternative formulations could be given using decision theory, along the lines of Anscombe (1963), upper and lower probability (e.g., Smith, 1961), or belief functions (e.g., Dempster, 1967). The general features of this conception, however, do not depend on the details of its implementation, which would ultimately lie in the hands of the investigators and the planning advisory board. Indeed, *the judgment itself*, of whether randomization is ethically justifiable, is necessarily subjective and is the responsibility of the investigators and the planning board. Thus, if the skeptic's beliefs are represented by a probability distribution, it must be recognized at the outset that the choices of the distribution, p_*, and p^* will be somewhat arbitrary. We believe that good choices can be made, and in Section 2 we briefly indicate the sort of thinking that is involved, but we do not wish to give the impression that they would be uniquely specified somehow by the nature of the problem.

We speak of a reasonable skeptic in part to emphasize that the beliefs we specify need not be our own, nor need they be the beliefs of any actual person we happen to know, nor derived in some way from any group of "experts." Instead, we think it is possible to imagine ourselves sufficiently cautious and skeptical that we would think the trial ethical. Having formally articulated the beliefs we would have in that situation, we may then examine them. If we, personally, find those beliefs not only in disagreement with our own but, in our judgment, untenable by any reasonable person, then we would find the trial unethical. On the other hand, having gone through this exercise, we would also be able to use these formal representations of beliefs to analyze data coming from the trial. We illustrate with a Bayesian analysis in Section 2.

We began with the presumption that an RCT is supposed to bring differing opinions to consensus. Our reduction of the problem to a single agent, our "cautious reasonable skeptic," is largely for convenience. As suggested earlier, a single distribution could be used and its evolution with increasing data, according to Bayes' Theorem, could be followed. It might then be assumed that this evolution, toward increased precision, would be sufficiently informative about the ability to reach consensus that specification of diverse opinion is unnecessary. (There are, we think, interesting technical issues raised by the latter assumption, though it is well known that different Bayesians will eventually reach consensus with sufficiently much data; see Savage, 1954, Sections 3.6 and 3.7.) Alternatively, the skeptic could be considered "to be of two minds," which represent extreme opinions among the group, and this could be formalized using two probability distributions. Our initial presumption is consistent with the point of view of Freedman (1987), who, as Ware mentions, calls the state in which no consensus exists, "clinical equipoise." Kadane and his colleagues in ongoing work (Kadane, 1986) have taken much the same position as Freedman, and they have implemented the suggestion using elicitation of expert opinion. (In a simple trial with no covariates, their scheme will allocate to achieve balance between the treatment and control as long as at least one expert opinion favors each; when all experts agree, the favored therapy will be allocated until it starts to do so poorly that opinion changes.) We propose an alternative framework because we think progress can be made without having to define a set of "experts," and without having to formally elicit their beliefs. The articulation of the beliefs of the skeptic remains somewhat arbitrary, but it has the advantage of being simple and, we think, relatively easy to understand. We note that our proposal, like that of Kadane and his colleagues, provides a single criterion that may be applied both at the beginning of and *during* a trial, to assess whether there is sufficient evidence to preclude use, or further use, of randomization.

2. BAYESIAN METHODOLOGY

The ECMO trial is again a good example for raising another old subject of debate: whether to test or to estimate. The issue is of great consequence from the Bayesian point of view, thought it is less so within the non-Bayesian framework. In our view there exist important situations in which a sharp null hypothesis should be taken seriously. These are cases in which it is believed possible, for all practical purposes, to have the null hypothesis hold exactly. Occasionally, clinical trials may have genuinely sharp nulls, but usually it may be assumed that the treatment and control will differ, the question being in which direction and by how much. In the large majority of clinical trials it would seem difficult to argue, as one must in adopting a testing methodology, that small differences, if detected, would be interesting. Thus, we believe that in the case of ECMO, as in most RCTs, estimation would be more appropriate than testing.

2.1 Marginal Inference

Inference about one parameter in the presence of another remains a major outstanding problem in statistics. In his analysis, Ware uses profile likelihood

together with first-order asymptotics to produce a confidence interval. Profile likelihood, however, has been criticized because it can give misleadingly precise inferences about the parameter of interest. In addition, first-order asymptotics are suspect in this small-sample case (see Cox and Reid, 1987, and the accompanying discussion, for non-Bayesian alternatives).

The Bayesian approach to this problem offers two advantages. First, posterior distributions are not affected by the design or stopping rule used in obtaining the data. Second, in contrast to the non-Bayesian approach, which presents various ways of obtaining confidence intervals, from the Bayesian point of view, once the prior is chosen, marginal inference about a parameter is based on a single well-defined entity, the marginal posterior distribution. That is, the arbitrariness in Bayesian marginal inference is put in one conspicuous place: the prior. Thus, from the Bayesian point of view, one may try to better understand what the data say about the parameter of interest by introducing alternative priors and determining resulting inferences. We illustrate by briefly considering what we might make of the data collected at the first stage, that is, the 6 survivals among 10 controls and 9 survivals among 9 ECMO patients. *We assume that it was appropriate to conduct an RCT*, i.e., that the first phase of the trial was ethical. This means we will use various priors that we feel would represent opinions of a cautious reasonable skeptic who was initially unwilling to state a preference for ECMO or the control therapy.

We will phrase our discussion in terms of a prior probability distribution on (δ,γ), the parameters δ and γ being defined by $\delta = \eta_2 - \eta_1$, and $\gamma = (\eta_1 + \eta_2)/2$, where $\eta_i = \log\{p_i/(1 - p_i)\}$, for $i = 1, 2$ ($i = 1, 2$ corresponding, respectively, to "control" and "ECMO"). There is nothing special about this particular parameterization and, as we have thought about the problem, we have found ourselves moving back and forth between the scales of probability, odds, and log odds. Ware uses $\delta = p_2 - p_1$. The latter may be easy to understand when the values of p_1 and p_2 are near .5, but we think the distance between probabilities of .75 and .999 should be far greater than that between probabilities of .5 and .749. Since probabilities close to 1 are being contemplated in this problem, we prefer the scale of log odds.

Before discussing our choices of priors in quantitative terms, let us mention the general issue of using the historical controls, that is, the 2 of 13 survivals reported by Ware, which were obtained from the chart reviews of patients treated at CHMC or BWH in 1982 and 1983. Like Ware, we wish to use this information, but we do not want to use it as if the historical controls were simply a previous sample from the same population as the experimental controls. In subjective terms, we do not consider the historical and experimental controls to have been exchangeable. If they had been, there would have been no reason to have used a control group in the trial. Since we are doing this analysis under the assumption that the trial was appropriately designed, we conclude that we should downweight the historical control information.

In addition, the limited use of the historical control information may be applied directly to η_1 or, instead, to γ, depending on whether one thinks it informs only about control patients or, rather, about general features of survivability that would be common to both the control and ECMO groups. For simplicity, we have chosen priors of two forms, $\pi(\delta,\gamma) = \pi_\delta(\delta) \cdot \pi_\gamma(\gamma)$, and $\pi(\delta,\eta_1) = \pi_\delta(\delta) \cdot \pi_{\eta_1}(\eta_1)$. We find the independence assumption, and the inextricably related issue of whether to apply the historical control information to η_1 or γ, somewhat subtle. We like the form $\pi(\delta,\gamma) = \pi_\delta(\delta) \cdot \pi_\gamma(\gamma)$, and in the remainder of this comment we will focus primarily on analyses based on such priors; we are unable, however, to spend the time to consider the issue carefully, which would require a deeper knowledge of the details of the therapies involved. Indeed, the same may be said for our choices of distributions for the marginal priors. Thus, we issue the disclaimer that our analysis is illustrative only, and we would not yet feel comfortable drawing firm conclusions from it.

We used a total of 42 priors of the form $\pi_\delta(\delta) \cdot \pi_\gamma(\gamma)$, and 42 of the form $\pi_\delta(\delta) \cdot \pi_{\eta_1}(\eta_1)$. We obtained the 42 by choosing 6 marginal densities $\pi_\delta(\delta)$ and, for each of these, selecting 7 marginal densities which became either $\pi_\gamma(\gamma)$ or $\pi_{\eta_1}(\eta_1)$ (that is, the same set of 7 densities was used for both γ and η_1). For each of the 84 priors, we computed the posterior probabilities $P\{\delta > 0 \mid y\}$ and $P\{\delta > .4 \mid y\}$ (where y is used to denote the data). The latter probability was chosen because we presume that an odds ratio of 3:2 would be substantial enough to be of great interest and $\log(3/2) = .405 \doteq .4$. That is, we use $P\{\delta > .4 \mid y\}$ to represent the posterior probability of a substantial superiority of ECMO, based on the first-stage data and the given prior.

We present here only an abbreviated explanation and summary of our analysis, giving results based on just a few of the priors we tried. (Further details may be obtained from us.) We used five priors on δ that were centered at $\delta = 0$. Two were Cauchy and three were Normal. We began with the intention of selecting a prior that would be fairly liberal, in the sense of assigning a nontrivial probability to a very large effect, and at the same time would assign most of the probability to more moderate effects. Thinking in terms of the odds ratio e^δ, the Cauchy$(0,\sigma_C^2)$ prior on δ with $\sigma_C = 1.099$ has 75th, 90th and 95th percentiles at odds

ratios of roughly 3:1, 30:1 and 1000:1. Thus, this prior, for example, assigns probability ½ to $\{\frac{1}{3} < e^{\delta} < 3\}$. This seemed quite suitable as a starting point. We also considered a Cauchy with $\sigma_C = .405$, which assigns probability ½ to $\{\frac{2}{3} < e^{\delta} < \frac{3}{2}\}$ instead of $\{\frac{1}{3} < e^{\delta} < 3\}$ and is a considerably tighter distribution on the odds ratio scale. The three Normal distributions had standard deviations of $\sigma_N = .60$, 2, and 10. (The first of these was chosen to match the Cauchy with $\sigma_C = .405$ according to the probability assigned to the interval $(-\sigma_C, \sigma_C)$.)

As far as the prior on γ or η_1 is concerned, we began with the Binomial likelihood $p^2(1-p)^{11}$ based on the historical controls. We then transformed to the log-odds scale $\log(p/(1-p))$ and applied the Normal approximation to the posterior based on a uniform prior, the location and precision of the Normal approximation being the posterior mode (the MLE) and the observed information. This gave Normal(μ,σ_N^2) with $\mu = -1.7$ and $\sigma_N = .769$. As explained above, we wished to "downweight" the historical control information, and we did so in several ways. The first was to use a 50:50 mixture of the Normal$(-1.7,(.769)^2)$ distribution with a uniform distribution. Since the resulting posterior is a mixture of the posteriors based on the Normal and uniform priors, we obtained the posterior based on the uniform prior as a by-product of the computation. We also tried Cauchy(μ,σ_C^2) distributions, first with $\mu = -1.7$ and $\sigma_C = .419$. We chose $\sigma_C = .419$ so that the resulting Cauchy distribution would assign the same probability to $(\mu - \sigma_N, \mu + \sigma_N)$ as did the Normal with $\sigma_N = .769$, i.e., probability .683. We then acknowledged that in most trials we would expect greater success than historical information might indicate, so we increased μ to $\mu = 0$, which seemed to us to be a substantial increase. (We did this for both Normal and Cauchy distributions; we also tried doubling the scale parameter.)

Selected results are given in Table 1, and the corresponding marginal densities appear in Figure 1. Based on our limited knowledge of the clinical situation, we follow the rationale sketched in the previous two paragraphs and focus on the posteriors labelled C and D. These were based on a Cauchy$(0,(1.099)^2)$ prior on δ and an independent prior on γ that was intended to use, yet downweight the historical control information. It may be seen that these two methods of downweighting the historical controls gave similar results: the first-stage data supplied sufficient information to alter the skeptic's opinion from $P\{\delta > 0\} = .5$ to $P\{\delta > 0 \mid y\} \doteq .95$, and to yield a probability of a substantial effect of ECMO, meaning an improved odds of survival of at least 3:2, of about .90.

To save space in this already-lengthy commentary, we refrain from making additional remarks about this analysis, except to say that we feel our choices of priors produced a satisfactory range of appropriate skeptical opinions, taking into account the historical controls. (We hope to make more extensive remarks available elsewhere.)

TABLE 1
Marginal posterior probabilities and Bayes factors based on selected priors

	Prior on γ	$P\{\delta > 0 \mid y\}$	$P\{\delta > .4 \mid y\}$	Bayes factor
A	$N(-1.7,(.769)^2)$	.91	.82	$(1.2)^{-1}$
B	uniform	.97	.93	$(3.7)^{-1}$
C	mixture	.95	.90	$(2.5)^{-1}$
D	$C(0,(.419)^2)$	.94	.88	$(2.1)^{-1}$
F	$N(-1.7,(.769)^2)$** Prior on η_1	.96	.93	2.1
E	$C(0,(.419)^2)$	.99	.97	$(15.7)^{-1}$

For priors labeled A–E the marginal prior on δ is Cauchy $(0,(1.099)^2)$, and the marginal prior on γ or η_1 is specified in the second column. The first column contains labels used to identify the densities in Figure 1. The ** at the prior labeled F indicates that for this distribution on γ, a Normal$(0,(10)^2)$ prior on δ was used.

We now return, again briefly, to the conception of a skeptic "being of two minds." We may consider two marginal prior distributions of δ, one centered to the left of $\delta = 0$, one centered to the right, and we will wish to see whether, according to both resulting posteriors, there is a large probability assigned to $\{\delta > 0\}$ or $\{\delta > .4\}$. For these data, clearly we need only examine the posterior based on the prior centered to the left of $\delta = 0$. We chose Cauchy(μ_C,σ_C^2), where $\mu_C = -.405$ and $\sigma_C = .405$. This distribution is centered at an odds ratio of 3:2 in favor of the control, and it assigns a probability of .75 that $\delta < 0$, and a probability of .5 that ECMO is worse by an odds ratio between 3:1 and 1:1. This seems, to us, to correspond roughly to a mild yet definite skepticism. We then used each of the seven priors on γ and η_1 discussed previously. We mention here only the results that were least favorable to ECMO, which were based on the Normal$(-1.7,(.769)^2)$ prior on γ. With this prior we obtained $P\{\delta > 0 \mid y\} \doteq .67$ and $P\{\delta > .4 \mid y\} \doteq .52$. Thus, even in this overly conservative case, the skeptic's prior probability from the more "pessimistic" of his "two minds" changed from .75 that ECMO would do more harm than good, to about .5 that ECMO would be substantially better.

In summary, based on our analysis, a small part of which has been reported here, we find that although the information favoring ECMO from the first-stage data is by no means overwhelming, it does seem to us probably sufficient to terminate randomization. The main purpose of this exercise, however, has been to show how Bayesian sensitivity analysis may be used to achieve an understanding of the information pro-

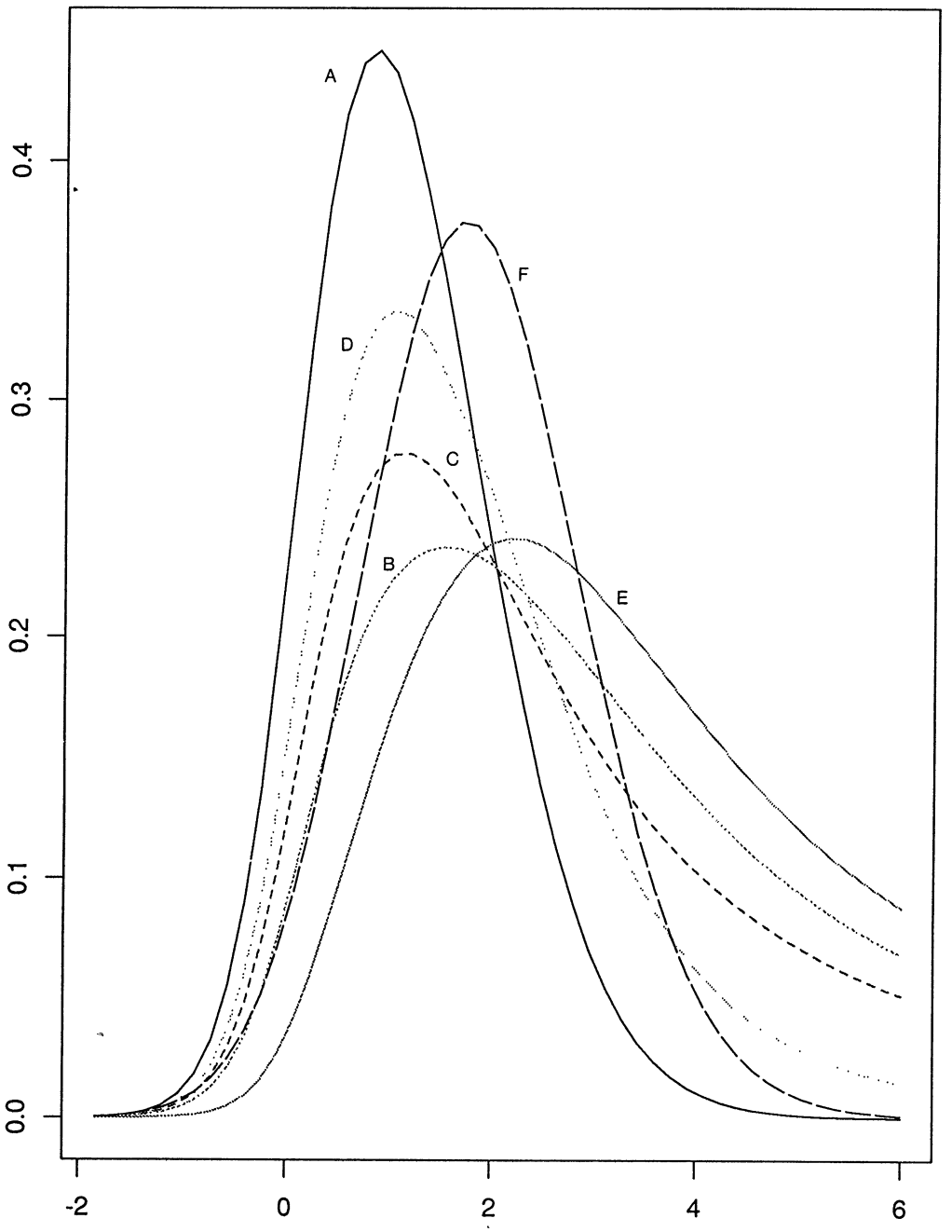

FIG. 1. *Marginal posterior densities of δ for priors specified in Table 1.*

vided by the data, even in small samples, without complications stemming from the design. In addition, we have tried to indicate how we might think about choosing priors when randomization is ethically justifiable.

2.2 Posterior Odds

From the Bayesian point of view, hypothesis testing is very different from estimation. Unlike many posterior interval probabilities, the posterior odds on a sharp null hypothesis is generally very sensitive to the choice of prior densities under the null and alternative hypotheses. This is because the Bayes factor based on data y is the ratio of two marginal densities of Y at the observed value y: the numerator is the marginal density at y under H_0, and the denominator is the marginal density at y under H_A. The observation y is often in the tail of each density, and the ratio is thus likely to vary substantially when the priors are changed. The two sets of priors that Ware has chosen do not really cover a "wide range." Assuming the testing problem were appropriate (and we have suggested above that in this case it is not), we think a

more thorough analysis would be needed before it would be appropriate to draw conclusions. We illustrate with a few calculations below.

We also believe the probability on which Ware focuses his interpretation is inappropriate. He emphasizes the probability, say q, that ECMO is inferior, rather than the probability $1 - q^*$ that it is superior. Since the null hypothesis has non-negligible probability, q and q^* are quite different. Based on a uniform prior on p_1 and a conditionally uniform prior on p_2, Ware finds $q = .01$ and $q^* = .11$. The issue in emphasizing either q or q^* is not whether the test is one-sided or two-sided, but how one would view a truth that EMCO and conventional therapy had identical population survival rates. In our opinion, if this truth held, we would expect conventional therapy to be recommended. Thus, we think the number deserving attention in Ware's calculation is $q^* = .11$. It is, perhaps, tempting to emphasize $q = .01$ because of the natural psychological tendency to equate posterior odds with the more familiar p-values: since a p-value of .11 would not be considered small, one might think $q^* = .11$ was not very convincing. However, such an identification should be avoided. Posterior odds and p-values are very different probabilities, and there is no reason to think that intuitions about one will immediately translate to the other (at least without further adjustment for sample size and priors). In a horse race, 8:1 odds against winning might not quite qualify a horse as a "long shot," but by any bettor's reckoning, odds of 8:1 are large. (By Jeffreys' rule of thumb, 10:1 is the rough cutoff between "substantial" and "strong" evidence; Jeffreys, 1961, Appendix B.) Thus, if Ware's prior were considered appropriate, we would conclude from $q^* = .11$ that there was fairly clear evidence against the null hypothesis.

Our own calculations are based on the "odds factor" or "Bayes factor" for testing H_0: $\delta = 0$ against the alternative H_A: $\delta \neq 0$. Under H_0 we write $p = p_1 = p_2$ and then, with self-explanatory notation for the Binomial likelihoods, the Bayes factor becomes

$$\frac{\int p^{y_1+y_2}(1-p)^{n_1+n_2-y_1-y_2}\pi_\gamma(\gamma)\,d\gamma}{\iint p_1^{y_1}(1-p_1)^{n_1-y_1}p_2^{y_2}(1-p_2)^{n_2-y_2}\pi(\delta,\gamma)\,d\delta\,d\gamma}$$

where p, p_1, and p_2 are functions of γ and δ, which were defined in the previous subsection. Note that the Bayes factor is equal to the posterior odds of H_0, if we take the prior odds of H_0 to be 1:1.

Our purpose here is to show how we might obtain an assessment of the evidence over a realistically broad range of priors, if we believed that the issue was whether the population survival rates for ECMO and control therapy were exactly equal. (We repeat: we do not really believe this.)

For this analysis, we used three of the marginal priors used previously on δ and all of the priors used previously on γ or η_1 (for a total of 42 priors). The resulting Bayes factors for the priors discussed in the previous subsection are given in Table 1. Thus, for example, for prior D, if we assume even odds *a priori*, then there are odds of 2.5 against H_0 *a posteriori*. An additional result, not found in the table, is that a uniform prior on p under H_0 and on (p_1, p_2) under H_A produces a Bayes factor of $(3.6)^{-1}$. These results may be contrasted with the posterior odds of 8:1 against H_0 based on Ware's conditionally uniform prior. His prior and resulting odds are interesting, but they are clearly *not* what we would call "conservative," given that we wish to conform to the assumption that the trial was ethical.

Again we refrain from further detailed comments, except to call attention to prior F, which has a Normal$(0,(10)^2)$ marginal distribution on δ. The Bayes factor is 2:1 *in favor of the null.* This illustrates the more dramatic sensitivity of the Bayes factors, in contrast to the marginal interval probabilities.

3. ADDITIONAL COMMENTS

3.1 Alternatives to RCT's

In Section 1, we suggested a definition of conditions under which randomization would be ethically justifiable, and in Section 2.1 we illustrated the construction of probability distributions that could represent beliefs when the conditions were satisfied. But in matters of life and death there is, as we indicated, an increased willingness to proceed with whatever appears to be the better treatment, regardless of the trustworthiness of the current evidence. In such circumstances it becomes extremely important to make the best possible use of available information. We have in mind careful study of prognostic factors, so that important covariates could be identified and evaluated, possibly using a subsample of the historical records. Formal adjustment procedures could be used (see, for example, Cochran and Rubin, 1973). As Cornfield (1954) said,

> ... It is a good deal more difficult to control variables in observational than in experimental [studies] ... But there is no difference in principle. There are no such categories as first-class evidence and second-class evidence. There are merely associations, whether observational or experimental that, in a given state of knowledge, can be accounted for in only one way or in several different ways. If the latter, it is our obligation to state what the alternative explanations or variables might be and to see how their effects can be eliminated ...

We understand that it took a lot of work for Ware and his colleagues to define the entry criteria and to

find 13 historical controls. Nonetheless, this is in large part a question of resources and, at this point, it becomes difficult to discuss the ethical problems without considering the larger societal question of resource allocation. We believe there is room for development and application of alternative methodology so that, with sufficient effort, it would be possible to provide a more satisfactory basis for the ethical judgment required in commencing an RCT. In assessing treatments of potentially great benefit, we think we might, in many cases, conclude that RCT's would not be ethically justifiable, and that information sufficient to persuade even a cautious skeptic would be available from observational studies.

3.2 Two-Stage Designs

In the case of ECMO, Ware and his medical associates believed that the historical data did not justify the use of ECMO without an RCT, and in an attempt to balance ethical and scientific concerns, they designed the ECMO trial using an adaptive two-stage treatment assignment procedure. It is interesting to note that one of the first RCTs done at the NIH (in 1953) was a trial investigating prevention of blindness in premature infants who were receiving high concentrations of oxygen therapy to reduce the incidence of brain damage and death. In an attempt to steer a middle course between the need to minimize the possible increase in mortality for those on the new curtailed-oxygen treatment, and the concern that continued use of a high concentration of oxygen might result in an unnecessarily high incidence of blindness, Sir A. Bradford Hill implemented an adaptive two-stage treatment assignment procedure similar to the one used by Ware. The outcome indicated that the relative risk of blindness for a baby receiving high oxygen was three times that for a baby with curtailed oxygen and, as a result, the practice of giving premature infants a high concentration of oxygen was widely modified (Greenhouse, 1989).

The virtue of an adaptive two-stage procedure in which all the patients receive the same therapy during the second stage is that the second-stage patients would receive great scrutiny and might be considered quite comparable to the first-stage patients. However, because of possible changes over time in patients studied and, perhaps, the treatment administered, comparability remains an important concern. Such two-stage designs would be most helpful if used in conjunction with thorough assessment of covariate information of the kind mentioned above for the analysis of trials with nonconcurrent controls.

4. CONCLUSION

In describing the design and analysis of the ECMO trial, Jim Ware has presented a case study on the ethics of clinical trials to which statisticians, clinicians and ethicists will refer for a long time to come. Reactions to experiments on sick babies are highly emotional, and Ware is to be greatly commended for sharing his struggle and his attempt to conduct an ethically and scientifically sound investigation of ECMO. His paper forces all of us to think harder about the ethical basis of RCTs and helps us to understand more deeply the issues involved. For this we thank him.

ACKNOWLEDGMENTS

This work was supported in part by Grant DMS-87-05646 from the National Science Foundation, and by Clinical Research Grant #30915 from the National Institute of Mental Health. The authors thank Ruth Douglas, Mark Segal and Teddy Seidenfeld for comments on an earler draft, and Suresh Vaidyanathan both for his comments and for his assistance with the numerical work.

ADDITIONAL REFERENCES

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

COCHRAN, W. G. and RUBIN, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā Ser. A* **35** 417–446.

CORNFIELD, J. (1954). Statistical relationship and proof in medicine. *Amer. Statist.* **8** 19–21.

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **49** 1–39.

DEMPSTER, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Statist.* **38** 325–339.

GREENHOUSE, S. W. (1989). Some historical and methodological developments in early clinical trials at the National Institutes of Health. *Statist. in Medicine.* To appear.

HILL, A. B. (1951). The clinical trial. *Brit. Med. Bull.* **7** 278–282.

HILL, A. B. (1953). The philosophy of the clinical trial. The National Institutes of Health Annual Lectures, Washington. (Reprinted in Hill, A. B., *Statistical Methods in Clinical and Preventive Medicine* 3-14. Oxford Univ. Press, New York, 1962.)

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford Univ. Press, Oxford.

KADANE, J. B. (1986). Progress toward a more ethical method for clinical trials. *J. Med. Philos.* **11** 385–404.

KADANE, J. B. and SEIDENFELD, T. (1989). Randomization in a Bayesian perspective. *J. Statist. Plann. Inference.* To appear.

MCCARTNEY, J. M. (1978). Encephalitis and Ara-A: An ethical case study. *Hastings Center Report* **8** 5–7.

SAVAGE, L. J. (1954). *The Foundations of Statistics.* Wiley, New York.

SMITH, C. A. B. (1961). Consistency in statistical inference and decision (with discussion). *J. Roy. Statist. Soc. Ser. B* **23** 1–37.