respiratory failure. *Trans. Amer. Soc. Artificial Internal Organs* **25** 459–461.

WUNG, J.-T., JAMES, L. J., KILCHEVSKY, E. and JAMES, E. (1985). Management of severe respiratory failure with persistence of the fetal circulation, without hyperventilation. *Pediatrics* **76** 488–494.

ZAPOL, W. M., SNIDER, M. T., JOHNSON, F. W. et al. (1976). Extracorporeal membrane oxygenation in severe acute respiratory failure: A randomized prospective study. *J. Amer. Med. Assoc.* **242** 2193–2196.

ZELEN, M. (1979). A new design for randomized clinical trials. *New England J. Med.* **300** 1242–1245.

Comment: Ethics and ECMO

Donald A. Berry

I will address several general issues that the Ware paper raises. These include the use of historical controls, the ethics of randomized trials, the impracticality of Neyman-Pearson inference, and optimal adaptive design. I will also suggest a more ethical and perhaps more scientific approach to medical research than that of randomized clinical trials.

RANDOMIZED CLINICAL TRIALS: THE EMPEROR'S NEW CLOTHES

Randomization has achieved hallowed status in biostatistics. Some biostatisticians and clinicians refuse to believe that a treatment has an effect unless it has been shown in a "properly conducted" randomized clinical trial. A report of a randomized clinical trial takes for granted that the trial provides the conclusive answer: if its conclusion is the same as the prevailing wisdom that is based on historical data, the authors tell us that we can finally believe this wisdom; if it differs, they chide historical data and extol the virtues of randomized studies. In the case of ECMO, there was a substantial amount of historical data that, in my view, not only carry more weight than the Ware study, but suggest that randomizing patients to non-ECMO therapy as in the Ware study was unethical.

Ware refers to several previous studies concerning ECMO. The Bartlett et al. (1985) study included 12 patients in its play-the-winner phase; all 11 ECMO patients survived and the conventional therapy patient died. Bartlett et al. also reported on 10 patients who met their entry criteria but were treated after the study was completed: all 8 patients assigned to ECMO survived and the 2 assigned to conventional therapy died (though the authors do not indicate the reasons for different therapy assignments—one possibility unrelated to prognosis is the availability of ECMO machines). Bartlett et al. say they admitted only patients who had at least an 80% chance of dying on conventional therapy. I am currently examining historical controls provided by Dr. Bartlett to verify this mortality rate, and so far I have no reason to doubt it. The 40% (4 of 10) death rate on CMT in the Ware study is somewhat inconsistent with an 80% mortality rate, but patients in the Bartlett et al. study generally had worse prognoses than those in the Ware study.

Commenting on the Bartlett et al. study, Ware and Epstein (1985) lament its 50% false-positive rate (or type I error level) since "in trials comparing equally effective innovative and standard therapies, the innovation would be identified as superior therapy in 50% of the trials." Type I error levels do not depend on the data; they are unconditional measures of inference. In particular, they average over data that might have occurred but did not. So the significance level of ½ would apply even if it happened that equal numbers had been assigned to the two therapies with all failures on one therapy and all successes on the other (this is unlikely but possible when using randomized play-the-winner assignment). I will return to conditional versus unconditional inference below. Ware and Epstein conclude that "Further randomized clinical trials using concurrent controls and addressing the ethical aspects of consent, randomization, and optimal care will be difficult but remain necessary." Hence the current study.

I disagree with the conclusion of Ware and Epstein: there was ample evidence in the Bartlett et al. study and in other evidence available at the time to conclude that ECMO is beneficial. (And I felt as strongly about this before I became aware of the Ware study.) This is clear if one uses measures of inference that condition on the observed data. For example, a Bayesian analysis that takes into account historical controls and the differing prognoses of the patients shows a dramatic benefit for ECMO (Berry and Hardwick, manuscript in preparation). Historical controls are

Donald A. Berry is Professor, School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455.

much maligned by biostatisticians. There are many examples in which historical control data are "contradicted" by randomized studies. But there are also many examples in which randomized studies are contradicted by other randomized studies. No set of data is perfectly informative. Using historical controls is difficult and mistakes will be made. The most important concern is that historical controls may differ systematically from current patients, and measurements for judging comparability may not be available. It may be wrong to use historical controls as though they are exchangeable with current patients. Good statistical methods for using historical control data seem not to be available. This is not because such methods are impossible to develop, but because their use is not a priority in medical research.

Why are randomized studies so generally accepted and believed? Didn't people learn before randomization? Don't people learn now without randomizing? Suppose I move to a new city. I drive route A to work during my first week and it takes between 28 and 32 minutes on each of the five days. The second week I use route B and it takes between 18 to 22 minutes on each of the five days. On the basis of these data I'm pretty convinced that B is better. Skeptics will (correctly) point out that there may be a time trend. But need I do a randomized study to convince you? All right, I confess that I randomized: I flipped a coin on each of the ten days and happened to get A's on the first five and B's on the second five! Does this change your interpretation of the experimental evidence? You might have been worried in the nonrandomized case that I had data available to me that influenced my decision and that was related to travel time (radio traffic reports, say). But assuming I can and do tell you everything I knew, you should analyze the data in the same way whether I randomized or not. And you should want to know what the traffic reports were even if you know that I randomized.

In H. C. Andersen's fairy tale, only fools failed to see the emperor's new "clothes." Nobody wants to be thought a fool, so everybody "saw" them. Nearly everyone praises randomization, even though once the data are at hand it is invisible: you can't tell that it was there. So maybe it doesn't matter whether it was there or not!

I want to make two qualifications to my suggestion that randomization has no value in clinical experiments. First, I may not be able to tell you everything I knew when I decided on a course of action. For example, suppose I am a physician who examined a patient before assigning therapy. I can tell you what measurements I made and the results. And I can give you a rough description of my perception of the patient's mental attitude, say. But I may not be able to quantify this perception. Nor can I precisely describe the feeling I had when looking into the patient's eyes. It may be difficult to satisfactorily overcome this problem, but I think it is possible to do so. (I hesitate to recommend having a clinician who has only seen the patient's clinical measurements assign treatment because this may also be unethical.)

The second qualification is that randomization can be used to force balance in treatment assignment. (A corollary is that randomization tends to minimize confounding of treatment assignment and time.) There are many circumstances in which I'd like such balance—even in some clinical trials. But not when the end point is serious, such as death. This leads to the question of whether randomized trials are ethical.

ETHICS AND RANDOMIZATION

In my view, the Ware study should not have been conducted. Randomizing patients to conventional therapy in the face of substantial evidence concerning the superiority of ECMO seems unethical. Ware makes it clear that the investigators favored ECMO before the study; he says that "there was a strong possibility that a randomized trial would show large differences in survival rates in the ECMO and CMT groups," and that the "unexpectedly high survival rate in the CMT group led to a larger study than was initially anticipated."

Ware invokes equipoise (genuine uncertainty) and clinical equipoise (lack of medical consensus) as sufficient reason for starting and continuing a study. No reasonable clinician should ever be certain, and there is seldom consensus about anything—not even about how much uncertainty there is! According to either of these two definitions, a state of equipoise concerning ECMO existed before the Ware study. And one still exists today. Would a new study be ethical? Neither a tad of uncertainty nor the existence of a minority opinion is sufficient reason to deliver medicine that the deliverer thinks is bad.

Ware strived "To balance ethical and scientific concerns ...." There should be no compromise here: Ethical concerns win. Scientific concerns are appropriate only insofar as they relate to good medical treatment. Is there good medical treatment to be gained by randomizing equally between ECMO and CMT? More may be learned about the comparison of the two therapies in a controlled, balanced study than in any other way. But there *are* other ways. The parents of the babies should have been given all available information and allowed to choose therapy. Most or all would have chosen ECMO. This actually would have given more information about ECMO than would a balanced trial, but much less about conventional therapy and hence less about the comparison. As I suggested above, this can be remedied at least in part

using historical controls. But if it can't be remedied sufficiently, we would have to make do with that insufficiency.

In designing a study, it is important to assess the possible benefits derived from the resulting information. Publishing a study that shows a substantial advantage for ECMO may have a beneficial effect on the treatment of patients outside the study—perhaps given current medical attitudes, even more beneficial if the study is randomized. On the other hand, when the Ware study was initiated (February, 1986), ECMO was becoming increasingly accepted and increasingly used. So the additional impact of this study is not clear.

What are the costs of randomization? The clear expectation was that more patients would die on conventional therapy than on ECMO. So the question is whether having an excess of deaths balances the worth of information gained. Since I believe such information is available without randomizing, my answer is a resounding "no." But suppose such information were available only by randomizing. I believe that an affirmative answer can only be made by society at large and not by an individual study group. As I suggest below, I don't believe that many people understand that randomization takes place in clinical trials, not even the patients in them. Any individual study group of course reflects the attitudes and mores of medical research generally, but medical researchers make up a tiny fraction of society—and they can hardly be regarded as unbiased because their livelihood as they know it is at stake. (When it comes time to vote for or against sacrificing a few individuals for the common good, I will vote against.)

Ware says that the study employed randomized consent. So parents of babies randomized to CMT were not told about the existence of ECMO. I agree with Ware that this raises "some difficult questions." But I disagree that "randomized consent was ethically justified." Here is a case where what people don't know hurts them a lot. How can it be ethical to keep potentially life saving information from someone who can use it?

And what of the Institutional Review Board that approved the study? Was it fully informed? Or was it simply told that the investigators did not *know* which therapy was better? At any rate, IRB's are part of the culture that thinks the emperor's new clothes are beautiful!

CONDITIONAL OR UNCONDITIONAL INFERENCE?

Bayesian inference is conditional in that it depends on the data actually observed and not on data that might have been observed but were not. Neyman-Pearson inference is unconditional since possibilities that did not occur also affect the conclusions. The Ware-Epstein (1985) calculation of 50% significance level in the Bartlett et al. (1985) study is an example.

The Ware study provides a revealing example of the impracticality of Neyman-Pearson inference. Unconditional inference depends greatly on the design; a given set of data can give strong evidence for an hypothesis if one design had been used, and strong evidence against it if another was used. The Ware study was based on inverse stopping and the negative binomial distribution. They achieved four deaths on CMT but only one (of 29 patients) on ECMO. So they couldn't use the negative binomial distribution and they couldn't make a Neyman-Pearson inference. (Actually, this is not too surprising; dropouts, protocol violations, and unforeseen circumstances make exact Neyman-Pearson inferences impossible in general. When calculating P-values, for example, statisticians usually pretend that the data collection procedure used was one that would have given what was actually observed.)

To patch things up Ware calls upon profile likelihood. This procedure uses the likelihoods after maximizing over nuisance parameters. Maximizing makes little sense to me, but averaging does. So I much prefer the Bayesian approach that Ware presents in his Discussion section.

Regarding the Bayesian approach, I will first comment on some specifics of Ware's analysis and then say how I think a Bayesian analysis should be carried out. First, Ware's prior distribution on (p_1, p_2) is very strange, with enormous discontinuities in the density near (0, 0) and (1, 1). In particular, given p_1 and $p_2 \neq p_1$, the probability of $p_2 > p_1$ is ½ *for all* p_1. I don't mean that this is bad, but I wonder if this really corresponded with anyone's prior opinion. Actually, this aspect of the prior has essentially no effect on the posterior when the likelihood of p_1 is concentrated away from 0 and 1, as it is in the Ware study.

In Ware's Bayesian calculations, he does not include the 20 ECMO patients that were in the nonrandomized phase of the trial. Curiously, he does include 13 historical CMT patients in one calculation as being exchangeable with the CMT patients in the randomized phase. Interestingly, his calculations make a strong case for the superiority of ECMO only when historical data are included. Suppose we include the 13 CMT historical patients and, consistent with my suggestions concerning the ethics of using conventional therapy, pretend that CMT had not been used at all in the study. This gives 11 deaths of the 13 patients on CMT as opposed to 1 death among 29 ECMO patients: a very compelling story!

Bayesian inferences should depend on everything that is known. This includes information about other patients as well as other information about the current

patients. Regarding the latter, randomization may tend to balance treatments over covariates, but it does a random job: sometimes good and sometimes not so good. Toomasian et al. (1988) developed a logit model for the probability of survival based on 715 ECMO cases. Professor Ware provided me with various covariates of the patients in his study so I could calculate the Toomasian et al. logit for these patients. In Berry (1989), I carry out a Bayesian analysis of the Ware study data that accounts for differences in prognosis. While I used individual prognoses, the averages in the two groups are interesting. The average survival probability of the nine ECMO patients in the randomization phase of the Ware study was 87.4%. The average survival probability (had ECMO been used) of the 10 CMT patients was 81.1%. So the control patients tended to have worse prognoses. But these differences in prognosis are not large enough to have a dramatic effect on the conclusion of Ware's Bayesian analysis. (Incidentally, the average survival probability for all 29 ECMO patients in the study was 84.2%. That their actual survival rate was 96.6% suggests more effective use of ECMO in the Ware study than in the national registry as a whole.)

Using all available information includes using historical controls at other institutions as well as at the same institution. In the case of ECMO, this includes the data on both ECMO and CMT available at Bartlett's center in Michigan. I indicated earlier that historical controls may not be exchangeable with current patients, especially when they are from other institutions, so I don't know how to best use them. As a first calculation I would take into account all known covariates and regard all patients with the same covariate values as exchangeable. But this is an area in which good applied research is necessary.

STUDY DESIGN

A good adaptive design in a randomized trial allows for early stopping should the accumulating data warrant it. As I suggested earlier, I think this happened before the Ware study started. But I will discuss its design in the abstract.

The stopping rule used in the Ware study has some undesirable characteristics. Following this rule, a trial might continue when the evidence is fairly strongly in favor of one of the therapies (perhaps very strongly when accounting for covariates). Yet it might stop when there is essentially no difference. It can even stop with four failures on one therapy when the evidence only slightly favors that therapy. For example, 6 of the first 10 patients in the Ware study were randomized to CMT. Had four of these six died and three of the four ECMO patients died, the "loser" in the randomized phase would have had the greater success rate: $\frac{2}{6} > \frac{1}{4}$. No one would claim that either therapy is better in this case, and that's the point: the trial may stop with very little evidence and no conclusion.

Bayesian analyses are not tied to the design (though designs are very important in the Bayesian approach). In particular, probability distributions of the various unknowns can be updated continually during a trial. This allows for fully informed decisions about continuing the trial or not by weighing what may be gained with what may be lost. Stopping rules should be based on such cost-benefit analyses and need not be completely specified in advance of the trial. One possibility is to attempt to maximize effective treatment over the patient horizon (see Berry and Fristedt, 1985, especially its bibliography). Optimal stopping rules do not allow for correct Neyman-Pearson tests of hypotheses or confidence intervals but, as I suggested earlier, strictly speaking we lose this ability in essentially every trial anyway!

THE FUTURE OF MEDICAL RESEARCH

Randomized trials are not perfect inferential tools. Many clinicians refuse to participate in randomized trials. What bias does this create? But even if randomized trials were perfect, ethical issues dictate their demise, at least for life-threatening conditions. The fact that randomization goes on in clinical trials is not understood by the vast majority of people. Few patients in clinical trials understand that their therapy is decided by flipping a coin, despite what it says on the informed consent form. This is changing. For example, federal requirements for randomized trials to obtain marketing approval for drugs have come under attack by AIDS activists concerning the placebo-controlled AZT randomized trial. Medical researchers will have to come up with other ways of doing business.

There is a valid alternative to randomized trials that is perfectly ethical: comprehensive data bases (or patient registries). Each MD could have a computer that is part of a national network. (This is similar to a recommendation of Ellwood, 1988.) Each patient's characteristics, diagnosis, treatment and follow-up visits would be entered into a national data base. These data bases would be open to the public. Medical journals would publish periodic summaries and analyses of the data bases. When a new therapy is introduced, control data of good quality will be accessible in the data base. Information concerning comparability with current patients will also be available. The many problems in drawing inferences can be overcome. They provide significant challenges for statisticians in developing the necessary methodology. (One problem with using literature controls is not present

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.