Comment

Colin B. Begg

I would like to preface my remarks by clarifying that I think the ECMO study was a very carefully constructed and thoughtful scientific study, and I complement Jim Ware and his colleagues on their efforts. My remarks, while they might be interpreted as being critical, are intended to be constructive, by sounding a cautionary note on the very understandable tendency of investigators to be more convinced by their own data than their professional colleagues are likely to be. Having said this, I fully appreciate that the author and his colleagues faced some very difficult decisions in this study due to the acute nature of the condition and, even though I feel that the randomized portion of the trial was terminated prematurely, I am not at all sure I would have acted differently had the decision been on my own shoulders.

My comments are designed to set the results of this particular medical issue in a historical context, and also to discuss the strength of the evidence from the randomized portion of the trial, since I think the key issue was the decision to terminate randomization.

Why do we do randomized trials? The consensus for randomization in medical research developed during the middle of this century in recognition of the empirical evidence that alternative, less structured, research designs are typically unreliable. There are many reasons for this. Uncontrolled studies are, for example, especially susceptible to variation in the case-mix of the study population due to patient heterogeneity. The consequent variation in outcomes can be very large especially if small sample sizes are employed. This fact, coupled with the various incentives for selective publication of favorable results (Begg and Berlin, 1988), in addition to the advocacy style of statistical analysis employed by many medical researchers, has produced a medical literature the credibility of which is continually being challenged. The randomized trial does not necessarily solve all these problems, but it provides a gold standard for judging new medical treatments, and for effectively debunking the more egregious claims of breakthroughs that frequently surface in the literature. In other words, the major value of the randomized trial is in a *confirmatory* role. That is, new breakthroughs are not discovered in a randomized trial. The new ideas are developed in pilot studies and other uncontrolled settings. The role of the randomized trial is to refute or confirm, as was the case in the ECMO study.

A consequence of these facts, in my opinion, is that it is desirable that the results of the trial be conclusive in their own right, insofar as this is possible. To be sure, if there are several trials being conducted, the confirmation may involve the formal or informal aggregation of data, as in meta-analysis. However, in the case of the ECMO study, this appears to be the only legitimate trial, and possibly the only one that will ever be conducted. So it is especially important in this trial that the results be conclusive and convincing. The fact that the author has resorted to the use of Bayesian analysis, incorporating results from uncontrolled studies in the prior, is a demonstration in and of itself that the results of the trial are not convincing in their own right. That is not to say that the use of data aggregation, or indeed Bayesian inference, is generally wrong. Rather it is an affirmation that only the randomized trial contains high quality data, and our historical experience tells us that it is desirable that our major conclusions be supported by such high quality data.

How strong is the evidence from the (randomized portion of the) ECMO study? The author has quoted a p-value of 0.054. However, the more conventional, two-sided, test has a p-value of 0.09 (Fisher's exact test). A more serious problem, however, is the potential for covariate imbalance between the treatment groups. In large studies, we can be confident that randomization distributes the poor risk and good risk patients in an evenhanded way. However, in small studies like this, serious covariate imbalance is quite likely and may well explain unusual results. A glance at Table 2 shows that the distribution of covariates is not especially balanced in this study, especially for age at entry and diagnosis. Without performing a stratified analysis it is hard to gauge the effect of the imbalance. I do not believe it is meaningful or appropriate to perform significance tests to compare the distributions of covariates, as a device to dismiss their potential importance. I feel that a minimal robustness analysis is to consider the effect on the results of any one patient's outcome being changed. If there is a radical change in the conclusions then we should be very concerned abut the believability of the study. There are two possible changes to consider. Suppose

Colin B. Begg is Chairman, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York 10021.

that 7 of the 10 CMT patients survived rather than 6 of 10, compared with 9 of 9 on ECMO. In this case the p-value is 0.21. Similarly, if one of the 9 ECMO patients had failed, we would be comparing 8 of 9 successes versus 6 of 10, giving a p-value of 0.31. Clearly, the nominal conclusion of the trial, as represented by the p-value , is very sensitive to changing the result of a single observation. As regards the Bayesian analysis of the study, I do not feel persuaded by the methodology used. It seems to me that if you adopt the Bayesian paradigm you must construct a prior that meaningfully reflects your prior opinion. The prior with point mass of 33% at the null does not make any such sense to me, so I am disinclined to take the results seriously. The second prior, Be (3, 12), based on the historical data is a reasonable one only if you believe that the historical observations and the randomized observations are equally reliable and informative. But if you believe this, there is no point in performing a randomized trial in the first place. As I mentioned earlier I believe that there is a qualitative difference in the believability of data from a randomized trial as compared with uncontrolled or historical data. If there is a rational and convincing way of determining the "weights" that one should attach to these different kinds of evidence, then an argument may be made for combining them. But to put them together as if they were identical seems to me to be clearly inappropriate.

Do the additional patients accrued in the nonrandomized phase of the study strengthen the evidence in favor or ECMO? I believe they do to some extent, but not to a great extent. The reason I say this is that the "quality" of these additional patients is much better than the historical data, in that they were accrued using the same eligibility criteria as the randomized patients and treated by the same doctors, so that a more persuasive argument can be made for combining them directly with the randomized patients. However, this is a study with unusual time trends. The CMT results were much more favorable than expected, and it is not unreasonable to suppose that there is a learning curve in the treatment of this difficult condition. Consequently, the use of concurrent controls is especially important. Moreover, by treating more patients with ECMO, you simply reinforce the presupposition that ECMO produces a high success rate in this population. The real deficiency is the absence of sufficient (randomized) controls, and this can never be rectified no matter how many additional patients are treated with ECMO. I might add at this point that I believe that the negative binomial type of design, or indeed any unbalanced scheme, is inefficient. The best plan is to continue balanced randomization until the trial results are conclusive and then stop and publish as soon as possible, rather than tinker around with adaptive schemes.

There is a further issue about the design that I feel is very important. This concerns the fact that the statistical inferences have been based only on the short-term results. The long-term sequelae of ECMO are unknown, though there is a significant risk of brain damage. Clearly, one cannot assess this issue reliably until there are sufficient follow-up data. However, the small sample size severely limits our ability to address this issue, especially as there are only six surviving controls. A larger randomized trial would permit a more comprehensive assessment of the merits of the alternative therapeutic approaches.

Finally, I would like to address the issue of the ethics of randomization when the results are suggestive that one treatment is superior. Although the consequences of the choice of treatment seem especially acute in this trial, since we are dealing with the life and death of infants, the fact is that the ethical dilemma exists for every patient in every randomized trial. Consequently, some commentators believe that randomization is never ethical (Hellman, 1979). That is, the data are never exactly equivocal, and so the doctor invariably will have a preference for a particular treatment. In other words, if you choose to adopt the philosophy of randomization you must sacrifice the interests of patients in the short term in favor of a strategy that will be of the greatest *collective* benefit in the long-term. The question is: when is the appropriate time to stop? The traditional statistical formulation of this question is based on the myopic premise that you only have to convince yourself, since thereafter every patient will be treated with the "better" treatment. Unfortunately, the global practice of medicine does not operate in this simplistic way (Peto, 1985), and the conclusions of individual trials only have an indirect impact on medical practice at best. Therefore, it is important that the conclusions be persuasive to the broad spectrum of medical practitioners, or at least to the influential intellectual leaders among them. For this reason, some commentators, such as Peto (1985), for example, believe that the treatment effects should be significant at the level of three standard deviations rather than the customary two. I am not sure that I would go this far, but I do believe there is an onus on the researcher to consider the global impact of the results on medical practice when deciding on the appropriate moment to terminate a trial.

To summarize, I feel that the randomized portion of this trial was terminated prematurely. The consequence may be that ECMO will continue to be used by devotees of the therapy, but that its wider acceptability may be impaired, although whether this is true

remains to be seen. In any case, as a precedent for important confirmatory studies in the future, I believe that 19 patients is just too small a sample size to be recommended. One could ask the question: what therapy would I choose if I had a child suffering form persistent pulmonary hypertension? Well, I would certainly choose ECMO based on the available evidence. However, I would also choose ECMO even if the data were only 7/9 versus 6/10 in its favor. In other words, when your own neck is on the line you always want to choose the treatment that appears to be best. Unfortunately, if everyone is permitted to do this the resulting anarchy would totally undermine the scientific rationale on which the best modern medical research is based. For this reason, emotive questions like the preceding one tend to cloud our reasoning when we debate the merits of randomization.

ADDITIONAL REFERENCES

BEGG, C. B. and BERLIN, J. A. (1988). Publication bias: A problem in interpreting medical data (with discussion). *J. Roy. Statist. Soc. Ser. A* **151** 419–463.

HELLMAN, S. (1979). Randomized clinical trials and the doctor-patient relationship: An ethical dilemma. *Cancer Clin. Trials* **2** 189–193.

PETO, R. (1985). Discussion of the papers by J. A. Bather and P. Armitage. *Internat. Statist. Rev.* **53** 31–34.

Comment

Peter Armitage and D. Stephen Coad

Dr. Ware has performed a valuable service in two particular respects. He has given us a carefully documented case study, tracing the role of the statistician from the interpretation of past data, to the planning of a new investigation and the analysis and presentation of its results. Editors of statistical journals frequently bemoan the paucity of case studies amongst the papers submitted to them. Here is an excellent example of such a study.

More specifically, Dr. Ware has described one of the very few clinical trials using any form of outcome-dependent allocation. Armitage (1985) has drawn attention to the need for more interchange of ideas and experience between theorists and practitioners concerned with this aspect of clinical trial methodology. Dr. Ware's paper is a welcome contribution to the literature, from both a practical and a theoretical viewpoint.

There are several examples in therapeutic medicine of unresolved questions, for which the evidence relies almost entirely on nonrandomized comparisons, but where investigators have, for ethical reasons, been reluctant to initiate randomized trials. It is hard to resist the view, expressed, for instance, by Chalmers, Block and Lee (1972), that randomized studies ought to be initiated at a very early stage of the introduction of new methods (they would say for the first patient). In the wake of the earlier inconclusive trial of ECMO, and the controversy to which it gave rise, the present investigators naturally had to tread cautiously, and their wish to restrict the use of CMT as far as possible is understandable. In a rather similar, and equally controversial, situation recently, the (British) Medical Research Council gave firm backing to an extensive trial of vitamin supplementation for women becoming pregnant after an earlier pregnancy resulting in a neural tube defect, to see whether supplementation reduces the risk of a further affected infant. Some had argued that evidence from nonrandomized studies was sufficient to justify routine use of supplementation. The MRC took the view that firm and reliable evidence was needed, and that a randomized trial, carefully monitored, was justified (Wald and Polani, 1984). Its results are awaited eagerly.

The evidence for the superiority of ECMO over CMT, patchy as it is, seems to us fairly convincing. Our view, though, is heavily affected by the fact that patients in phase 2, all of whom received ECMO, were apparently at higher risk than those in phase 1. The eligibility criterion was tightened to exclude some less severely affected patients, and a higher proportion than in phase 1 were outborn, a characteristic apparently conferring higher risk. Had this feature not been present we should have been only moderately impressed, on the grounds that the comparability of phases 1 and 2 was in doubt and that the evidence from phase 1 was weak.

As regards phase 1, we are skeptical of any analysis that suggests a difference much more significant than is given by the Fisher exact test. The Bayesian probabilities for $p_1 > p_2$ are small, partly because an arbitrary amount of prior (and therefore posterior)

Peter Armitage is Professor of Applied Statistics and D. Stephen Coad is a research student, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, England.