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.