SOME PRACTICAL PROBLEMS IN CLINICAL TRIALS

WILLIAM F. TAYLOR and PETER C. O'BRIEN
MAYO CLINIC, MINNESOTA

1. Introduction

In 1971, the Mayo Clinic received a large grant to carry out clinical research on cancer. Twenty new projects were approved for support. Ordinarily these would be considered to be twenty independent projects and they would each have been provided with statistical services separately by our Medical Statistics Section. Something has been added, however, which provides a new set of problems. There exists now a new organization called a Clinical Cancer Center. It has been our problem (those of us in Medical Statistics) to try to define what a clinical cancer center is and to provide its statistical heart. At this writing we are still trying to develop a workable, unified record system tied in with appropriate computerization. So far the Center helps prepare protocols, designs the research records, handles randomization, edits data promptly, and prepares the data for analysis. Our intent is to examine results frequently, to provide summary reports frequently and, in general, to keep an aggressive watch over the course of the research. (There are a number of our statistical acquaintances around the country who have been through the same thing that we are going through now. They have our respect and admiration. Incidentally, we used to scoff at this sort of work as being pedestrian and dull. We scoff no more.)

In preparation for this grant, a massive effort was made at the Mayo Clinic to come up with suggestions of research projects that could be done with our large clinical practice. (It should be remembered that the Mayo Clinic has an enormous cancer patient load. About 6,000 new ones appear there each year.) As a result of this effort some 100 projects were proposed for support by a grant. These were whittled down by a committee to around 40 projects which were written up in formal NIH style and submitted to the National Cancer Institute. Site visits are never pleasant things if you are the one being visited. In this instance we were confronted by a distinguished panel of cancer experts. The visitors were both highly competent and highly critical. They examined our requests thoroughly. About half of our projects were turned down by the site visitors for one reason or another. Their report said, in part, the following.

"With respect to the scientific merit and design of the proposed studies, practically all suffered from primarily one weakness, namely, an improper experimental design from a biostatistical standpoint. Because of this, protocols were

left open ended and with too many options so that even with all the patient resources each year some of the studies would take years to accumulate enough data for analysis and by this time, of course, the time element would make comparisons impossible. The reviewers judged that those protocols illustrating this weakness were prepared without proper consultation with the biostatistical staff of the Mayo Clinic."

This statement is partly true. We had carefully worked over many of the projects using appropriate sample size considerations and revised protocols. But many were submitted without or in spite of our statistical criticism and some were difficult to love. Even so, a few projects we liked were turned down. On the other hand, some of the projects which were approved by the highly critical group of site visitors were ones that we had disapproved because we could not see that there would be enough patients in a reasonable length of time to come to any conclusions. There is an explanation of sorts. Some of the research topics involving only a few patients were exceedingly interesting to the reviewers and to the world of cancer research. While even the Mayo Clinic's case load might not be very large for some rare diseases, still it was one of the largest case loads in the country and some problems were considered interesting enough to look at in spite of small numbers of patients. We now feel in sympathy with the idea that a reasonable (inexpensive) randomized trial even on very infrequent cases may be better than the alternative we use now. If we had only started this ten years ago on myeloma, we would have some answers now which we need very much.

When we examined the officially accepted research proposals (and the several new ones which are now being prepared), we found, typically, that the research is clinical in nature requiring that each suitable patient be assigned at random to one of several treatment groups. The problem is to make a comparative evaluation of the treatments. How does one determine if a treatment is any good? It is not easy. For example, in the evaluation of patients with cancer of the prostate a "remission" is defined as (i) decrease in the size of the prostatic or paraprostatic mass or the size of the radiographic lesion as measured independently by two urologists, (ii) a decrease in the total acid phosphatase of 50 per cent or the tartrate inhibition fraction by 20 per cent, (iii) improvement in activity or resumption of normal activities since therapy. The first and third of these things are quite subjective and are difficult to measure with assurance. However, once remission occurs, then the usual problem is to determine how long it lasts. We are involved with survival time problems.

In clinical trials, survival time in some form or another is a very common object of study. The simple survival to death or survival to failure or survival to some sort of change of state is a fairly routine analysis. A first order complication is the case where survival alone is insufficient. Suppose a treatment is tried out in the hope of inducing remission of disease. However, remission may or may not be observed to occur. If a remission occurs, it may last for quite some time and a conditional "duration of remission" study can be carried out. But we really have

the problem of total evaluation, determining both the proportion of patients in whom remission occurs and the duration of such remission. Similarly, when we resort to studying only total survival time, ignoring "quality of survival," we may lose the entire benefit of treatment. Our gynecologists were dismayed, for example, at finding no increase in survival of women with cancer of the ovaries in spite of new treatments which were felt to be helpful. Possibly the quality of life was all that could be improved in this dreadful disease, but as yet we do not know.

In the very large effort expended now in the evaluation of chemotherapy, there is a class of clinical trials called Phase II Research. This is research in which drugs which have been found to have interesting possibilities on the basis of animal research and on very tentative human research are submitted to various investigators to try out on suitable patients. The object is to find those drugs which seem to show some kind of beneficial activity. In spite of years of work, there are still gaps in the common sense of these experiments. One complicating problem is the fact that in most institutions patients are used repeatedly; those failing one drug are tried out on another, and upon failing a second time they are tried out on a third. One wonders how much is missed in a promising drug by testing it out predominantly on patients who have failed one or more previous trials. This occurs frequently in survival analyses.

Suppose a patient starts out with a certain treatment and is followed until he gets either a recurrence of the disease or a new disease. He then is removed from the study and re-entered into a new study to see how he survives under his new condition. The problem is what do you compare him with in his new condition? What kind of comparison survival curves can be constructed? For example, patients with chronic ulcerative colitis are treated without surgery for a certain disease. After a time certain patients are subjected to surgery. Did surgery do any good? Even if selection for surgery were random, the time of such selection in this progressive disease makes for difficulties. We have used a simple variation of a standard survival analysis [2], but we have also seen serious errors made in this situation. Surely our own methods are in need of improvement.

Another problem in the comparison of survival of two or more groups has to do with the differences in the groups due, perhaps, to lack of randomization, or due to "good" randomization which just turned out to be grossly unbalanced. In clinical trials, we think we should attempt to stratify where we think it is important and to randomize within each stratum. There are problems, then, of evaluating overall effects of treatment on the basis of observations from several strata, none of which has sufficient cases to stand alone.

Turn back now to the earlier part of this paper in which we discussed research protocols which we must help develop for the Mayo Clinic Cancer Center. We sometimes have to deal with small samples. The problems are interesting but the patients are slow in appearing at the Clinic. Lengthy observations must be made; research discipline must be maintained for years. It seems obvious that in these projects the investigators will be pushed for decisions as early as possible. Se-

quential analysis appears to be a reasonable approach to take in such cases. In addition, as part of the surveillance of the data collection system, the data must be examined. Waiting three or four years without analyzing the data, even though we edit it and query it and file it carefully away, simply leads to poor data and boredom. We cannot afford to let our data pile up unanalyzed. We must look at them frequently (much more frequently, in fact, than we used to think was necessary). Therefore, there is a natural incentive to evaluate research in progress and to see how we are doing. When our clinicians learn (as they will) of the results of our intermediate statistical analyses, they will say, "Don't we have enough now so we can stop?" It is an inevitable and a natural question. Therefore, it seems to us that in a carefully run data collection protocol, sequential analysis is thrust upon us by the nature of the needs for quality control. Early curiosity is not a scientific sin.

For example, in studying three treatments for chronic hepatitis, a double blind experiment was done. Patients as they came in were assigned at random to one of three groups. Every time a death occurred the investigators "broke the code" and looked at the treatment that the dead person had received. After all too short a time they began seeing deaths occur as follows: placebo, placebo, placebo, treatment 1, placebo, treatment 1, treatment 1. Treatment 2 had no deaths for a long time compared with the other two treatments. The investigators came in and they said, "Look, we think we are killing people with this experiment," but they also said, "We do not want to stop this experiment if we stop it so prematurely that somebody else is going to have to do it again." So faced with this compassionate, yet mature, outlook we made a sequential stopping decision.

2. Sequential methods

We have recently been studying the work of Robbins and others on sequential experimentation and have been attempting to apply some of these results to our survival time problems. So far our work is only a preliminary effort. We have also worked with Armitage's sequential pairing method [1], but wish to supplement this with sequential interval estimation methods such as those in Robbins' recent article in the *Annals* [3]. Robbins' work makes use of a theorem due to J. Ville [4]; also see A. Wald [5]. The probability that the likelihood ratio based on a sample of size n should exceed a quantity ϵ, for some n greater than 1, is $\leqq 1/\epsilon$, $\epsilon > 1$. This theorem leads to many things; for example, to certain types of confidence interval estimates and to tests of hypotheses with power 1. One problem, discounted by Robbins, has to do with the fact that the sequential plans are open ended and the tests will rarely end under the null hypothesis. We give some examples of approaches that we have been examining.

The first approach is a sequential confidence interval estimate for an assumed constant risk of death. Consider a group of individuals who come under observation at intervals, one at a time. A treatment is applied and each individual is

observed for the rest of his life. At the time of the nth death we observe the number of patients seen so far A_n, the time survived by each patient t_i and the state of each patient ($x_i = 0$ if alive, $x_i = 1$ if dead). We wish to estimate the value λ_0 of λ, the assumed constant risk of death. We carry out a paraphrase of Robbins' sequential confidence intervals. Let hypothesis H_0 specify $\lambda = \lambda_0$ and let alternative H_1 merely state that λ has been obtained at random from a distribution $F(\lambda)$. Let $g_n(\{t, x\})$ be the joint density function of the observations under H_0 and $g'_n(\{t, x\})$ that under H_1 at the time of the nth death. We use Ville's theorem

$$P\left\{\frac{g'_n}{g_n} \geqq \varepsilon \text{ for some } n \geqq 1\right\} \leqq \frac{1}{\varepsilon}, \qquad \varepsilon > 1. \tag{1}$$

In this case,

$$\frac{g'_n}{g_n} = \frac{\int_0^\infty \exp\{-\lambda \sum t_i\}\, \lambda^n\, dF(\lambda)}{\exp\{-\lambda_0 \sum t_i\}\, \lambda_0^n}, \tag{2}$$

where summation is over the number of patients A_n. If we choose $F(\lambda)$ as an exponential distribution with mean λ_0, this reduces to

(3)

$$P\left\{\frac{g'_n}{g_n} = \frac{e^{-1}/(n+1)}{[\exp\{-(\lambda_0 T_n + 1)\}(\lambda_0 T_n + 1)^{n+1}]/(n+1)!} \geqq \varepsilon \text{ for some } n \geqq 1\right\} \leqq \frac{1}{\varepsilon},$$

where $T_n = \sum_{i=1}^{A_n} t_i$. Hence, the sequential confidence interval for λ_0 of size $1 - 1/\varepsilon$ can be defined as the interval $I(n)$ such that $\lambda_0 \in I(n)$, whenever an appropriate Poisson probability satisfies the inequality inside the braces below. Thus,

(4)

$$P\left\{\frac{\exp\{-(\lambda_0 T_n + 1)\}(\lambda_0 T_n + 1)^{n+1}}{(n+1)!} \geqq \frac{e^{-1}}{(n+1)\varepsilon} \text{ for every } n \geqq 1\right\} \geqq 1 - \frac{1}{\varepsilon}.$$

A possible sequential test for two treatments might be to compute these sequential confidence intervals for the λ of each of the treatments and stop as soon as the two confidence intervals failed to overlap. This is quite crude and yet should provide a test of power 1. The expense might be great, for the expected sample size would doubtless be larger than for some other methods.

A second example is really an attempt to utilize more information than Armitage uses in his paired method of sequential experimentation analysis. Suppose the patients are coming in fairly rapidly compared with the rate of death, so that after a while we are pretty sure of having some extra patients alive whenever we observe a death. Consider two groups of patients. Suppose that at the time just before the jth death there are N_{1j} patients in one treatment group still alive and N_{2j} patients in the other. Suppose that the risk in treatment 1 is λ_1 and the risk in treatment 2 is λ_2. Then, given a death occurs, the probability that the death occurs in the first group can be expressed as $N_{1j}\lambda_1/(N_{1j}\lambda_1 + N_{2j}\lambda_2)$. Under the null hypothesis, $\lambda_1 = \lambda_2 = \lambda$. Under the alternative hypothesis $\lambda_1 =$

$c\lambda_2$. We then set up a sequential probability ratio test and arrive at a test of the hypothesis that $c = 1$. This is again sequential over the occurrence of successive deaths in the combined groups. Suppose we stop at the time of the nth death and look at the data. After cancelling λ_2 and λ, we obtain

$$(5)\quad P\left(\frac{g_n'}{g_n} = \prod_{j=1}^{n} \frac{[N_{1j}cx_j + N_{2j}(1 - x_j)]/(N_{1j}c + N_{2j})}{[N_{1j}x_j + N_{2j}(1 - x_j)]/(N_{1j} + N_{2j})} \geqq \varepsilon \text{ for some } n \geqq 1\right) - \frac{1}{\varepsilon},$$

where $x_j = 1$ if jth death occurs in first treatment group, $x_j = 0$ otherwise. Letting $\sum_j^n x_j = n_1$ and $\sum_{j=1}^n (1 - x_j) = n - n_1 = n_2$,

$$(6)\quad P\left(\log \frac{g_n'}{g_n} = n_1 \log c - \sum_{j=1}^{n} \log (N_{1j} + N_{2j}) + \sum_{j=1}^{n} \log (N_{1j}c + N_{2j}) \geqq \log \varepsilon \text{ for some } n \geqq 1\right)$$

$$= P\left(n_1 \geqq \frac{1}{\log c} \{\log \varepsilon - \sum^{n} \log (N_{1j} + N_{2j}) + \sum^{n} \log (N_{1j}c + N_{2j})\} \text{ for some } n \geqq 1\right) \leqq \frac{1}{\varepsilon}$$

if $c > 1$. The inequality reverses (inside) if $c < 1$. If n_1 satisfies this inequality, we stop taking new cases and reject the hypothesis that $\lambda_1 = \lambda_2$. We can also define a confidence interval for c from (6).

A third example is still in the prenatal stage. Suppose we have two treatment groups which, by the nth death, have resulted in: A_{1n} cases started on 1st treatment, n_1 dead, and A_{2n} cases started on 2nd treatment, n_2 dead, where $A_{1n} = A_{2n}$ and $n_1 + n_2 = n$. These two groups have survival times totalling T_{1n} and T_{2n}, respectively. We test the hypothesis $\lambda_1 = \lambda_2 = \lambda$ against the alternative that $\lambda_1 = c\lambda_2$.

The likelihood ratio is

$$(7)\qquad \frac{g_n'}{g_n} = \frac{\exp \{-c\lambda_2 T_{1n}(c\lambda_2)^{n_1}\} \exp \{-\lambda_2 T_{2n}\lambda_2^{n_2}\}}{\exp \{-\lambda(T_{1n} + T_{2n})\}\lambda^n}.$$

If we replace λ_2 and λ by their maximum likelihood estimates, we get an expression of unknown characteristics but with, we feel, reasons for further examination:

$$(8)\quad P\left(\frac{g_n'}{g_n} \geqq \varepsilon \text{ for some } n \geqq 1\right) \approx P\left(\frac{e^{-n}e^{n_1}[n/(cT_{1n} + T_{2n})]^n}{e^{-n}[n/(T_{1n} + T_{2n})]} \geqq \varepsilon \text{ for some } n \geqq 1\right)$$

$$= P\left(n_1 \geqq \frac{1}{\log c} \{\log \varepsilon - n \log (T_{1n} + T_{2n}) + n \log (cT_{1n} + T_{2n})\} \text{ for some } n \geqq 1\right) \leqq \frac{1}{\varepsilon}.$$

If $c < 1$, the inequality for n_1 reverses. Note the similarity to the previous example, (6).

Finally, as a fourth approach, we put the pair preference method of Armitage into the same context. At the nth death let there be M pairs of patients who have arrived in the study and who have been randomized (by pairs) into the two treatment groups. For each such pair, we can establish a preference only if at least one member has died on or before the time of the nth death. We prefer the treatment associated with longer survival. Let the number of preferences be m and let the number of these for which treatment 1 is *not* preferred be m_1. If p is the probability of not preferring treatment 1, the hypothesis to test here is that $p = \frac{1}{2}$. As shown by Armitage $p = \lambda_1/(\lambda_1 + \lambda_2) = c/(1 + c)$, where c, as above, is λ_1/λ_2. The likelihood ratio is now

$$(9) \qquad \frac{g_n'}{g_n} = \frac{p^{m_1}(1-p)^{m-m_1}}{(\frac{1}{2})^m} = \left(\frac{p}{1-p}\right)^{m_1}[2(1-p)]^m = c^{m_1}\left(\frac{2}{1+c}\right)^m.$$

Hence, we write Ville's theorem as

$$(10) \qquad P\left(m_1 \geqq \frac{1}{\log c}\left\{\log \varepsilon - m \log \frac{2}{1+c}\right\} \text{ for some } m \geqq 1\right) \leqq \frac{1}{\varepsilon}.$$

If $c < 1$, the inequality for m_1 reverses. We note once more a similarity in form among this expression and the two previous examples, (6) and (8).

3. Monte Carlo simulation

We produced by Monte Carlo methods a simulated situation in which patients "arrive" at the Clinic in a Poisson process at rate 0.2 per day. The first arrival was randomly assigned treatment R_x1 or R_x2; the second received the other treatment. Each odd numbered arrival was thus assigned a treatment at random. Those receiving R_x1 died according to a Poisson process at risk λ_1, those with R_x2 at risk λ_2. Every time a death occurred, we applied the three stopping rules defined by the inequalities in (6), (8), and (10). For each rule, we recorded the sample number defined as the number of deaths required for terminating the experiment.

We obtained 100 sequential experiments. In each experiment, we arbitrarily limited the number of cases to 200 (100 pairs), following each until death. We let $\lambda_1 = 0.003$, $\lambda_2 = 0.009$, and $c = \frac{1}{3}$. For each of the three stopping rules, each of the 100 experiments stopped before the 200 cases had arrived. For the method of (6), the "number at risk" method, the average sample number (ASN) was 23.2 with a range from 6 to 78. The 90th percentile was 44 and the standard deviation 14.6. For method (8), the "time at risk" method, the ASN was 23.0, range 6 to 78, the 90th percentile 41, and the standard deviation 14.8. These are quite close. Out of the 100 experiments the number at risk method led to a smaller n than the time at risk method 20 times, to a larger n 27 times, and to the same n 53 times.

The stopping rule (10), the pair preference method, tended to require a sample number larger than the above two methods. We found the ASN to be 35.7 with

range from 8 to 161, the 90th percentile 61, and the standard deviation 28.2. Out of the 100 experiments this pair preference method was worse (larger n) than both the others 77 times. It was better than both 9 times.

The Monte Carlo procedure was repeated 100 times under the null hypothesis in which both λ_1 and $\lambda_2 = 0.009$ and 100 pairs of patients were followed until all had died. The value of c in (6), (8), and (10), however, was kept at $\frac{1}{3}$. The number at risk method did not stop 97 times, the time at risk method 96 times, and the pair preference method 97 times. In two of the 100 experiments all three methods stopped fairly early (between the 18th and 39th death).

4. Remarks

We have found that two sequential stopping rules (6) and (8) appear to be more sensitive on the average than Armitage's pair preference method (10) in a situation featuring a rather strong difference in the risk of death in two treatment groups. We have also found no apparent difference in the proportion of times the null hypothesis was rejected when really true. The methods used were preliminary. The number of simulated cases might have been too few in the null hypothesis situation. Surely the difference between treatment groups should be varied and simulation retried with smaller differences. Confidence intervals for the value of c in (6), (8), and (10) would also have been valuable.

A word in defense of the pair preference method. Its assumptions are somewhat less restrictive than those in the two other methods given here; the assumption of constant risk can be relaxed somewhat.

◇ ◇ ◇ ◇ ◇

Our thanks go to Mr. Roger Oenning who programmed the Monte Carlo aspects of this study.

REFERENCES

[1] P. Armitage, *Sequential Medical Trials*, Oxford, Blackwell, 1960, Chapter 7.
[2] G. J. Devroede, W. F. Taylor, W. G. Sauer, R. J. Jackman, and G. B. Stickler, "Cancer risk and life expectancy of children with ulcerative colitis," *New England J. Med.*, Vol. 285 (1971), pp. 17–21.
[3] H. Robbins, "Statistical methods related to the law of the iterated logarithm," *Ann. Math. Statist.*, Vol. 41 (1970), pp. 1397–1409.
[4] J. Ville, *Étude Critique de la Notion de Collectif*, Paris, Gauthier-Villars, 1939.
[5] A. Wald, *Sequential Analysis*, New York, Wiley, 1947.