Eur J Cancer Clin Oncol, Vol. 22, No. 11, pp. 1279–1283, 1986.
Printed in Great Britain

0277-5379/86$3.00+0.00

Perspectives and Commentaries

Significance Testing in the Comparison of Survival Curves from Clinical Trials of Cancer Treatment

JOHN L. HAYBITTLE
MRC Cancer Trials Office, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH, U.K.

THE logrank rest [1] is now widely used for comparing survival data from randomised clinical trials of cancer treatment that require prolonged follow-up. The test is optimal when the death rate in one group consistently exceeds that in the other group by a given proportion, the so-called "proportional hazards" situation. Alternative tests which are sometimes used are Gehan's generalisation of the Wilcoxon rank sum test [2] and its subsequent modification by Peto and Peto [1] and by Prentice [3]. Of these, the latter is to be preferred with censored data [4] and, as shown by Lee *et al.* [5] in a simulation experiment comparing survival curves modelled on Weibull distributions, may perform better in a non-proportional hazards situation. Similarly Fleming *et al.* [6] have demonstrated the loss of power of the logrank compared with that of the Wilcoxon test in comparing survival curves where the greatest differences occur at early follow-up times, and Harrington and Fleming [7] have shown a similar loss when the hazard ratio is a maximum at time zero and decreases smoothly towards unity as follow up increases. The reason for the difference between the performance of the two tests is that the calculation of the Wilcoxon statistic weights the differences between observed and expected events according to the estimated survival at the time of the event whereas the logrank calculation gives equal weights at all event times [8]. Thus the Wilcoxon test gives more weight to differences which appear early in follow-up.

It may be questioned whether the proportional hazards model is the most appropriate one for many trials of cancer therapy, since the control arm may often contain a subset of long-term survivors or "cured" patients, and the new therapy being tested is unlikely to effect any improvement of survival in this subset. For example, Nissen-Meyer [9] has postulated that the effect of adjuvant therapy following surgery for primary breast cancer might be either to increase the percentage of long-term survivors, or to delay tumour growth in the remainder while not turning them into long-term survivors, or a combination of both these effects. In none of these situations would the outcome be a reduction of the hazard rate by a constant proportion throughout the time of follow-up.

This is illustrated in Fig. 1 where the three possible outcomes are plotted on log-linear graphs and the ratio of the slopes of the curves at any particular time gives the hazard ratio at that time. Figure 1A shows the effect of an increase in the percentage of long-term survivors (Type A effect). As the patients with a poor prognosis become a progressively smaller fraction of those still at risk, the death rates in the two groups become the same i.e. that of the long-term survivors. The hazard ratio (Fig. 2) decreases with time towards unity.

In Fig. 1B the percentage of long-term survivors is the same in both groups but the effect of adjuvant therapy has been to give some increased survival time to the poor prognosis patients (Type B effect). Now the slope of the upper curve is initially less than that of the lower curve, but later the situation reverses as the delayed deaths in the poor prognosis group occur, and eventually the slopes are identical. Thus the hazard ratio (Fig. 2) decreases with time to a minimum value which is less than unity and then increases towards unity.

The combination of both effects is illustrated in Fig. 1C where the resulting hazard ratio plot (Fig. 2) is similar to that for Fig. 1B.

A further consequence of outcomes such as those in Fig. 1 is that the power of a test may be reduced rather than increased with further follow-up even though more events will be recorded. This has been discussed by Peto *et al.* [10] who suggest that the follow-up period may be sub-divided and the logrank test applied separately in each period, so

Accepted 25 April 1986.

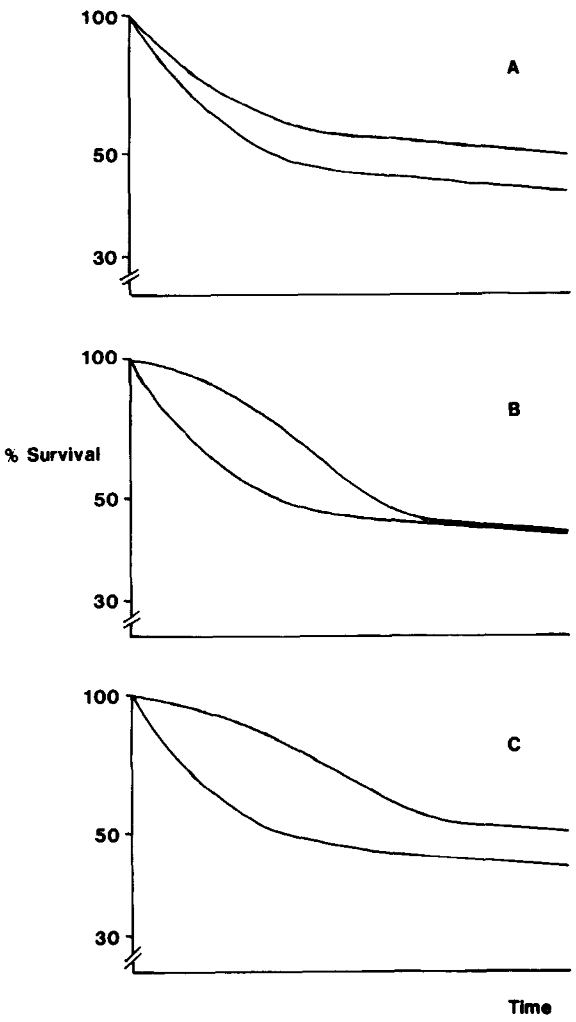

Fig. 1. *Possible outcomes of trial of cancer treatment:*
A. *Increase in proportion of long-term survivors.*
B. *Increased survival time in proportion of short-term survivors.*
C. *Combination of both A and B effects.*

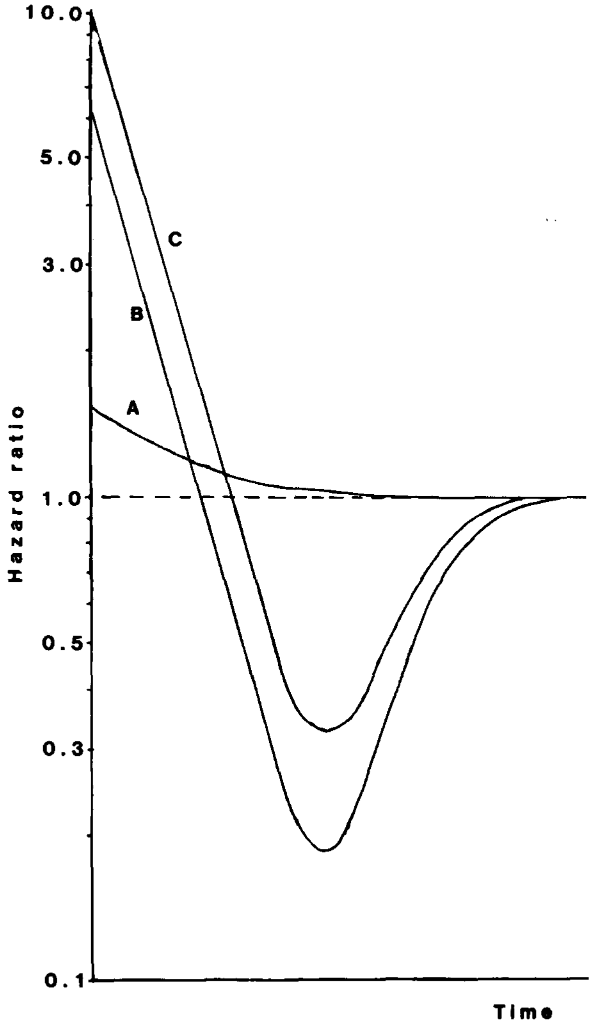

Fig. 2. *Plots of hazard ratio (control arm/new treatment arm) for survival curves of Fig.* 1.

that the significance of early differences can be assessed. They warn, however, that the point of sub-division should not be chosen *after* examination of the two survival curves, but should be based on some consideration of the overall survival curve of all patients together to find when the overall death rate changes.

In spite of this recommendation, very few published clinical trials have been analysed in this way, and a common report following analysis without sub-division has been that an earlier reported significant difference has been reduced or has disappeared after further follow up. For example, the results of the trial of adjuvant polyA-polyU in breast cancer after a maximum of 5 yr follow-up [11] showed a significant difference ($P < 0.03$) in relapse-free survival time of node positive patients. After 8 yr follow-up this difference was non-significant [12]. A report of a UICC meeting on the management of early breast cancer [13] stated that "Reporting data on the results of adjuvant therapy too early can result in premature acceptance of benefits which subsequently may disappear". If the benefit was extension of survival in a poor prognosis fraction of patients, a subsequent non-significant result by the logrank test carried out without sub-division of the time-scale does not necessarily mean that this benefit was illusory, but only that the way the test was used was inappropriate.

SIMULATION EXPERIMENT

The continued uncritical use of the logrank test in spite of its known loss of power when hazards are not proportional may be partly due to lack of appreciation of the magnitude of this loss. In the examples simulated by Fleming *et al.* [6] the power of the logrank was found to be less than 50% of that of the Wilcoxon test in models showing early differences in the survival curves.

Table 1 gives the results of a simulation exper-

Table 1. *Results of simulation experiments*

		New treatment arm				% of trials* judged significant† (power of test)			
		(1-*c*) fraction				At 4 yr		At 8 yr	
		Increase in:-							
Row no.	Control arm *c*	*c*	Mean survival (yr)	Half-life (yr)	Type of effect	Logrank	Wilcoxon	Logrank	Wilcoxon
1	0.6	0.8	0	0	A	70	68	87	88
2	0.3	0.5	0	0	A	57	53	86	80
3	0.6	0.6	1.5	0	B	88	95	31	56
4	0.3	0.3	1.0	0	B	98	100	55	94
5	0.6	0.7	1.0	0	A & B	84	89	63	79
6	0.6	0.6	0	1.5	B	51	53	39	50
7	0.3	0.3	0	1.5	B	90	90	92	96

*Out of a total of 400 trials for each row.

†Using $\chi^2 \geqslant 3.841$, $P \leqslant 0.05$ as the criterion for significance.

iment specifically designed to reproduce the situations shown in Fig. 1. Two hundred patients were entered into each arm of a trial at a uniform rate over a period of 4 yr. Analysis of each trial was made at 4 yr after the first patient was entered, i.e. at completion of entry, and also after a further follow-up of 4 yr. Four hundred trials were simulated for each set of conditions.

The model for the survival curve of the control arm was taken to be a two-exponential one, namely:-

$$S(t) = c \cdot \exp(-\lambda_1 t) \ + \ (1 - c) \cdot \exp(-\lambda_2 t)$$
$$(t > 0, 0 < c < 1)$$

where $S(t)$ is the surviving fraction at time t, λ_1 and λ_2 are constants with $\lambda_2 \gg \lambda_1$, and c is the fraction of patients having a smaller mortality rate, i.e. the potentially long-term survivors. With $c = 0.6$ and λ_1 and λ_2 corresponding to half-lives of 20 and 1.5 yr respectively, the model is a reasonably good representation of relapse-free survival up to 8 yr as published by Lacour *et al.* [12] for their control group of node-positive breast cancer patients.

To simulate the effect of the new treatment in the second arm, c could be increased by a specific amount and/or the survival times of patients with the higher mortality rate (λ_2) could be increased. Two methods of achieving this latter change were used. In one the survival times were increased by the simple addition of an amount derived from a normal distribution of increments with a specified mean, μ, and standard deviation, σ . σ was taken to be $\mu/3$ and the distribution was truncated at -3σ so that none of the survival times in the new treatment arm were actually decreased. In the second method, λ_2 was decreased i.e.the half-life for the poor prognosis group was increased.

All simulations were carried out with λ_1 and λ_2 in the control arm corresponding to half-lives of 20 and 1.5 yr respectively. λ_1 was always kept corresponding to a half-life of 20 yr in the new treatment arm. Four simulations were made with $c = 0.6$ in the control arm; three were made with $c = 0.3$.

Rows 1 and 2 of Table 1 show that when the effect of the new treatment was solely to increase the proportion of long-term survivors (Type A effect; Fig. 1A), both tests behaved very similarly, although there is some indication of a small power advantage for the logrank test in row 2 when the proportion of long-term survivors is smaller.

When the effect of the new treatment was solely to increase survival in an additive manner in the poor prognosis patients (Type B effect), then the picture was very different (rows 3 and 4). The tests had similar power at 4 yr, but both lost power with further follow-up due to the curves coming together from about 5 yr onwards. However, the loss of power was much less marked with the Wilcoxon than with the logrank test. At 8 yr in each simulation the power of the logrank was less than 60% of that of the Wilcoxon. Figure 3 is an example of one simulated trial where both tests gave significant results at 4 yr but only the Wilcoxon test still showed a significant difference at 8 yr.

Row 5 shows the results when both increase of c and additively increased survival of poor prognosis patients were simulated. Again the loss of power at 8 yr with the Wilcoxon is less marked than the loss with the logrank test.

Rows 6 and 7 give the results when the Type B effect was simulated by doubling the half-life in the poor prognosis group. In this sub-group the hazards then maintain a constant ratio, but, because of the unaffected good prognosis fraction, the overall picture is still one of non-proportional hazards. The results in row 6 when $c = 0.6$ are qualitatively similar to those in row 3 in that power

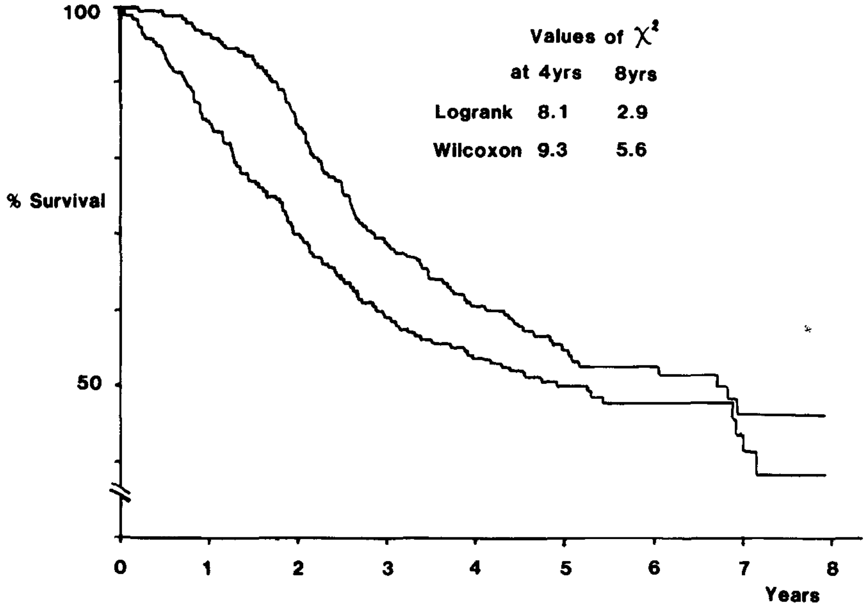

Fig. 3. *Simulated trial result at* 8 yr *after first patient entered.* 200 *patients per arm entered over* 4 yr. *Control group assumed to have* 60% *long-term survivors dying exponentially with a half-life of* 20 yr *and* 40% *short-term survivors dying exponentially with a half-life of* 1.5 yr. *New treatment arm (upper curve) similar to control except that survival times of short-term survivors increased additively by amounts drawn randomly from a normal distribution of increments of mean* 1.5 yr *and standard deviation* 0.5 yr.

decreases with further follow up, though less markedly, and the Wilcoxon test has a higher power than the logrank at the later follow-up point. When *c* is smaller (row 7) the overall situation is nearer to that of proportional hazards and both tests behave similarly with some increase of power at later follow-up.

DISCUSSION

The results demonstrate that the loss of power of the logrank by comparison with that of the Wilcoxon test can be considerable in some situations where the ratio of the hazard rates does not remain constant. The loss is most marked when the new treatment has a Type B effect only, i.e. prolongs survival of a subset of patients with a poor prognosis but does not increase the number of long-term survivors. This could well be the effect to be expected of a number of new cancer treatments and, unless the follow-up period is to be sub-divided for analysis (with the disadvantages that may arise from an arbitrary choice of division points), the Wilcoxon test is more suitable than the logrank for use in such trials.

The choice of test to be used should be governed by the hypothesis to be tested, and clinicians initiating trials should give some thought to the way in which the hoped for benefit of a new treatment is likely to be expressed. Statisticians can then be guided accordingly in the choice of test to be used. In any case, when the result of a trial is analysed, an examination should be made of the way in which the hazard ratio varies with follow-up time. The logrank test should not be used for an overall comparison of curves such as those shown in Fig. 3.

Acknowledgements—I am indebted to Mr. Laurence Freedman for helpful discussions during the course of this work, and to Professor N.M. Bleehen for allowing me to use the facilities of the MRC Cancer Trials Office, Cambridge.

REFERENCES

1. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *J R Stat Soc A* 1972, **135,** 185–198.
2. Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 1965, **52,** 203–217.
3. Prentice RL. Linear rank tests with right censored data. *Biometrika* 1978, **65,** 167–179.
4. Prentice RL, Marek P. A qualitative discrepancy between censored data rank tests. *Biometrics* 1979, **35,** 861–867.
5. Lee ET, Desu MM, Gehan EA. A Monte Carlo study of the power of some two-sample tests. *Biometrika* 1975, **62,** 425–432.
6. Fleming TR, O'Fallon JR, O'Brien PC, Harrington DP. Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics* 1980, **36,** 607–625.

7. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. *Biometrika* 1982, **69,** 553–566.
8. Tarone RE, Ware J. On distribution-free tests for equality of survival distributions. *Biometrika* 1977, **64,** 156–160.
9. Nissen-Meyer R. Adjuvant cytostatic and endocrine therapy: increased cure rate or delayed manifest disease. *Comm Res Breast Dis* 1979, **1,** 95–109.
10. Peto R, Pike MC, Armitage P *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br J Cancer* 1977, **35,** 1–39.
11. Lacour J, Lacour F, Spira A *et al.* Adjuvant treatment with polyadenylic-polyuridylic acid (PolyA.PolyU) in operable breast cancer. *Lancet* 1980, **ii,** 161–164.
12. Lacour J, Lacour F, Spira A. *et al.* Adjuvant treatment with polyadenylic-polyuridylic acid in operable breast cancer: updated results of a randomized trial. *Br Med J* 1984, **288,** 589–592.
13. Canellos GP, Hellman S, Veronesi U. The management of early breast cancer. *N Engl J Med* 1982, **306,** 1430–1432.