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

J. L. Haybittle**

MRC Cancer Trials Office, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH, Great Britain

The log-rank test (Peto and Peto 1972) 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 that are sometimes used are Gehan's generalisation of the Wilcoxon rank sum test (Gehan 1965) and its subsequent modifications by Peto and Peto (1972) and by Prentice (1978). Of these, the latter is to be preferred with censored data (Prentice and Marek 1979), and, as shown by Lee et al. (1975) in a simulation experiment comparing survival curves modelled on Weibull distributions, it may perform better in a nonproportional hazards situation. Similarly, Fleming et al. (1980) have demonstrated the loss of power of the log-rank 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 (1982) 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 log-rank calculation gives equal weights at all event times (Tarone and Ware 1977). 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 (1979) has postulated that the effect of adjuvant therapy

---

* This article is reprinted with permission from Pergamon Journals Ltd., from Haybittle JL (1986) Eur J Cancer Clin Oncol 22: 1279–1283.

** I am indebted to Mr. Laurence Freedman for his 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.

Recent Results in Cancer Research, Vol. 111

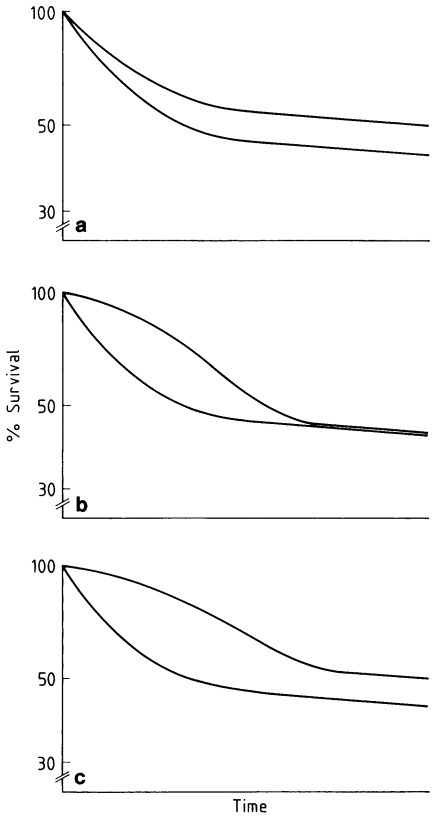


**Fig. 1 a–c.** Possible outcomes of trials of cancer treatment. **a** Increase in proportion of long-term survivors (type A effect); **b** increased survival time in proportion of short-term survivors (type B effect); **c** combination of both A and B effects

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 duration 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 1 a 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. 1 b 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