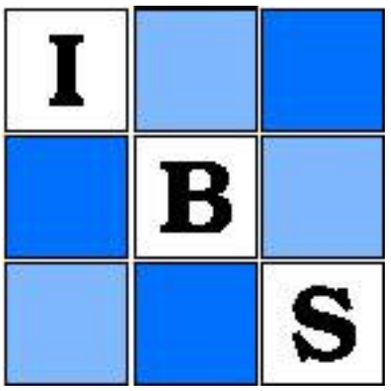

BIOMETRICS 54, 1072–1084
September 1998

A Method for Sequential Analysis of Survival Data with Nonproportional Hazards

M. R. Sooriyarachchi

Department of Statistics and Computer Science, University of Colombo, Sri Lanka

and

John Whitehead*

Medical and Pharmaceutical Statistics Research Unit, The University of Reading, P.O. Box 240, Earley Gate, Reading RG6 6FN, U.K.

SUMMARY

Two tests are proposed for comparing the survival curves of patients randomised between an experimental treatment and a control treatment when it is anticipated that the two survival curves may not satisfy the assumption of proportional hazards. The tests are particularly useful for the situation in which the survival curves are coincident or cross over early in the follow-up period and then diverge. The tests compare the probabilities of survival for longer than some fixed time since randomisation for the two groups of patients. Both methods take account of the right-censored observations, and both are associated with methods for estimating and setting confidence limits for treatment differences. The first method is a mathematically direct approach based on the derivation of the efficient score statistic and Fisher's information. The second method is simpler, being based on Kaplan–Meier estimates and their variances. Conventional methods of sample size determination require the assumption of proportional hazards. Here a sequential approach is used, as it is difficult to set the sample size in advance without strong assumptions about the relationship between the two survival curves. Simulation results giving information on the size and power of the proposed tests are provided and the tests are applied to data from a clinical trial in breast cancer.

1. Introduction

Consider a comparative clinical trial in which the response of interest is the time from randomisation to the occurrence of some well-defined event. The aim is to compare the survival experience of two groups of subjects, one group randomised to an experimental treatment and the other to a control treatment. Most sample size calculations and sequential designs for the comparison of survival curves are based on the assumption of proportional hazards (e.g., Machin and Campbell, 1987; Freedman, 1982; Tsiatis, 1982; Whitehead, 1992; Kim, 1992). This assumption implies that one survival curve dominates the other throughout the trial. Often, at the design stage of the trial, the relationship between the survival curves will be unknown. This is particularly true if patients are to be followed up over a long time period: e.g., 2–5 years of follow-up is common in cancers such as breast or prostate. At the end of a fixed sample trial, it might be possible to identify a suitable model on which to base the final analysis, although this would be data driven. However, at this stage, it would be too late to extend recruitment if it became apparent that a longer trial were needed for definitive comparison of the patterns of survival. In a sequential trial, no suitable model can be reliably identified from early interim analyses. Furthermore, if hazards are nonproportional, then the estimated hazard ratio is dependent on calendar time, resulting in different hazard ratio estimates at different interim analyses. In this case, variation in hazard ratio estimates between

* *Corresponding author's email address:* mps@reading.ac.uk

Key words: Clinical trials; Kaplan–Meier estimates; Proportional hazards; Sequential methods; Survival analysis.

interim analyses is due not only to random variation but also to the fact that the hazard ratio is time dependent (Gregory et al., 1997).

This paper is concerned with two tests for use in sequential trials for the comparison of survival curves when the hazards are or might be nonproportional. The methods proposed allow survival curves to be compared without making any assumptions about the relationship between them. They are especially appropriate in cases in which the survival curves are coincident or cross over early in the follow-up period and then diverge. The tests compare the probabilities of survival for longer than some fixed time (τ) since randomisation for the two groups of patients. The time point τ is selected to reflect long-term survival and therefore to exceed any anticipated point of divergence or crossing and to be of clinical interest. An example of survival curves with early crossing might be a breast cancer clinical trial in which the control treatment is chemotherapy and the experimental treatment is surgery. There may be more deaths on the experimental treatment within the first few months of treatment due to operative mortality, whereas patients on the chemotherapy will have a good chance of surviving this early stage. During the first few months after treatment, the survival curve for the surgery group will be consistently lower than the survival curve of the chemotherapy group. However, when the early phase, with its dangers of operative mortality, has passed, the survivors in the surgery group might have a better chance of subsequent survival than those on chemotherapy due to the long-term effectiveness of the surgical treatment.

The first test considered is a generalisation of the method of Whitehead (1984) for a single-arm study comparing the probability of survival past a fixed point for a group of patients on an experimental treatment with a known value. The method takes account of the right-censored observations and is thus referred to hereafter as the censored binary method. The second test is based on the Kaplan–Meier estimates of survival past time τ for the two treatment groups and their variances as given by Greenwood's formula. Both methods require that some patients in both experimental and control groups pass time τ since recruitment before they can be used. Thus, although sequential, they will not react directly to the extreme situation of a large excess of failures before time τ in one treatment group.

In this paper, the methodologies are applied to sequential designs with straight-line boundaries, although they would fit in with other sequential methods equally well. The approach from which the censored binary method is derived originated in the work of Bartlett (1946), was developed by Whitehead (1978), and is described in detail in Whitehead (1997). In general, the true advantage of one treatment over another is represented by a parameter θ. At any interim analysis, two statistics (Z and V) are calculated, where Z is the score statistic for θ evaluated under the null hypothesis and acts as a cumulative measure of the evidence that θ is positive and V is Fisher's information, which is a measure of the information about θ available in the data. In sequential trials, Z is plotted against V until an appropriate boundary is crossed. Conditional on V, Z has mean θV and variance V. In many special cases, the sample path formed by plotting Z against V over time can be approximated by a Brownian motion with drift θ, as has been proved in the context of specific survival models by Sellke and Siegmund (1983) and Tsiatis, Boucher, and Kim (1995). If such a result were true in the situation considered here, then it would directly justify the properties claimed for the censored binary method. We have not proved this result, and so the validity of both the censored binary and Kaplan–Meier approaches are based only on heuristic considerations and on demonstration through simulation.

Sections 2 and 3 of this paper introduce the censored binary and Kaplan–Meier methods, respectively, and Section 4 describes other methods used for comparison. Section 5 presents the simulation results, and Section 6 concerns an illustrative reanalysis of data from a breast cancer trial.

2. The Censored Binary Method

2.1 *Grouping Survival Data*

In order to apply this method to continuous survival data, the data must first be grouped into a small number of intervals $(t_{i-1}, t_i]$, $i = 1, \ldots, h$, $t_0 = 0$, $t_h = \tau$. The selection of the cutpoints of the intervals, t_i, is made for convenience. Mathematically, it is desirable that the interval cutpoints are chosen in such a way as to have equal numbers of events in each interval. From the continuous survival data, information on the occurrence or nonoccurrence of the event during each of the intervals of time since randomisation of the patient $(t_{i-1}, t_i]$ is extracted for each patient. The information on the occurrence of the event in $(t_{i-1}, t_i]$ should only be used after time t_i has elapsed since the patient in question was recruited to the trial, in order to avoid biased overestimation of

absolute event rates. Thus, no data at all should be used on a patient until time t_1 has elapsed after randomisation. Sometimes survival data come naturally in this grouped form and are termed interval-censored data (Whitehead, 1989). As an example, consider data collected in the following way. Patients are examined by a doctor at times $t_1, \ldots, t_h$ after randomisation. At each visit, the doctor determines by way of a diagnostic test whether the event has occurred since the last visit. The datum for a patient consists of the interval $(t_{i-1}, t_i]$, $i = 1, \ldots, h$, in which the event falls.

Continuous survival data can be used in the censored binary method without grouping if each event time is used as a cutpoint in the grid of intervals and has a corresponding parameter. The number of parameters is then very large, and although consistent estimates can be found, the computation becomes complex. In practice, it is better to group them into intervals; it is unlikely that much loss of efficiency will result.

2.2 *The One-Sample Case*

In order to fix ideas, we first suppose that all patients receive the experimental treatment. The null hypothesis H_0: $p_h = p^*$ is to be tested against the alternative H_1: $p_h \neq p^*$, where p_h is the probability of survival past time $\tau = t_h$ and p^* is some known probability. The measure of treatment difference θ between p_h and p^* is taken to be the log odds ratio

$$\theta = \log\left\{\frac{p_h(1-p^*)}{p^*(1-p_h)}\right\}.$$

Further notation is introduced as follows. The quantities involved can be calculated at any time during the trial. The number of survival times since randomisation that have values within $(t_{i-1}, t_i]$ is o_i, and the number of survival times greater than t_i is s_i. The probabilities p_i and q_i are defined as

$$p_i = P(T > t_i) \quad \text{and} \quad q_i = P(T \in (t_{i-1}, t_i] \mid T > t_{i-1}),$$

where T is a typical survival time, $i = 1, \ldots, h$. Let ϕ denote the vector-valued nuisance parameter made up of the parameters $p_1, \ldots, p_{h-1}$ and ϕ^* denote its maximum likelihood estimate given that $\theta = 0$. The log likelihood of θ and ϕ based on the data is denoted by ℓ, and ℓ_θ and $\ell_{\theta\theta}$ denote, respectively, the first and second derivatives of ℓ with respect to θ. Similarly, derivatives with respect to ϕ are denoted by ℓ_ϕ and $\ell_{\phi\phi}$, and the mixed derivative of ℓ with respect to θ and ϕ is denoted by $\ell_{\theta\phi}$.

The sequential method makes use of the statistics Z and V, which are derived from the equations

$$Z = \ell_\theta(0, \phi^*) \quad \text{and} \quad V = -\left\{\ell^{\theta\theta}(0, \phi^*)\right\}^{-1},$$

where $\{\ell^{\theta\theta}\}^{-1} = \ell_{\theta\theta} - \ell'_{\theta\phi}\{\ell_{\phi\phi}\}^{-1}\ell_{\theta\phi}$. Whitehead (1984) gives the derivation of Z and V for the one-sample case. The values of Z and V so derived can be expressed as

$$Z = \eta(1-p^*) \quad \text{and} \quad V = [(1-p^*)^2 s_h + p^{*2}(1-p^*)^2/B_h - (1-p^*)(1-2p^*)\eta],$$

where η satisfies

$$\prod_{i=1}^{h}\left(\frac{s_i - \eta}{o_i + s_i - \eta}\right) = p^*,$$

$$B_j = \frac{B_{j-1}b_j a_{j-1} + b_j + B_{j-1}}{(a_{j-1}B_{j-1} + 1)}, \qquad j = 2, \ldots, h \text{ and } B_1 = b_1,$$

and

$$a_j = \frac{(s_j - s_{j+1} - o_{j+1})}{\hat{p}_j^2}, \qquad j = 1, \ldots, h-1,$$

$$b_j = \frac{(\hat{p}_{j-1} - \hat{p}_j)^2}{o_j}, \qquad j = 1, \ldots, h.$$

2.3 *The Two-Sample Case*

Consider the recruitment of two samples of patients, those randomised to an experimental treatment (E) and those allocated to the control treatment (C). All quantities will be defined as for the one-sample case, with the second subscript E or C added to distinguish between those pertaining to the experimental and control samples, respectively.

The null hypothesis H_0: $p_E = p_C$ is to be tested against the alternative H_1: $p_E \neq p_C$, where $p_E \equiv p_{Eh}$ and $p_C \equiv p_{Ch}$ are the probabilities of survival beyond time $\tau = t_h$ for patients on E and C, respectively. The parameter of interest measuring the difference between treatments is taken to be the log odds ratio $\theta = \log[p_E(1-p_C)/\{p_C(1-p_E)\}]$. The nuisance parameter ϕ is a vector consisting of $\psi = \log[p_E p_C/\{(1-p_E)(1-p_C)\}]$, $(q_{1,E}, \ldots, q_{h-1,E})$, and $(q_{1,C}, \ldots, q_{h-1,C})$, and ϕ^* is its maximum likelihood estimator given that $\theta = 0$. The likelihood function is given by

$$L = \prod_{i=1}^{h} q_{iE}^{o_{iE}}(1-q_{iE})^{s_{iE}} q_{iC}^{o_{iC}}(1-q_{iC})^{s_{iC}}.$$

The log likelihood is maximised subject to the constraint that $\theta = 0$ using Lagrange's method. This gives

$$\hat{q}_{iE} = \frac{o_{iE}}{(o_{iE} + s_{iE} - \eta)}; \qquad \hat{q}_{iC} = \frac{o_{iC}}{(o_{iC} + s_{iC} + \eta)},$$

where η satisfies

$$\prod_{i=1}^{h}\left(\frac{s_{iE} - \eta}{o_{iE} + s_{iE} - \eta}\right) = \prod_{i=1}^{h}\left(\frac{s_{iC} + \eta}{o_{iC} + s_{iC} + \eta}\right) = p^* \tag{2.1}$$

and p^* is the common value of the two products. It follows after manipulation similar to that carried out in the one-sample case that the statistics Z and V used for sequential testing are given by

$$Z = \eta(1-p^*)$$

and

$$V = -\left[p^{*2}(1-p^*)^2 \ell^{(E)}\ell^{(C)}\right] \Big/ (\ell^{EE} + \ell^{CC}),$$

where $\ell^{(E)} = \ell^{EE} + [(1-2p^*)\eta/\{p^{*2}(1-p^*)\}]$, $\ell^{EE} = -(s_{h,E}/p^{*2} + 1/B_{h,E})$, $\ell^{(C)}$, and ℓ^{CC} are defined similarly for the control treatment, and the B terms are as defined for the one-sample case.

The value of η can be found by solving the polynomial (2.1), which has $2h-1$ roots. As $o_{1,E} + s_{1,E} > s_{1,E} > o_{2,E} + s_{2,E} > s_{2,E} > \cdots > s_{h,E}$ and similarly for the control treatment, it can be shown that only one root lies between $-s_{h,C}$ and $s_{h,E}$. This is the required value of η to make all of the $q_{i,C}$ and $q_{i,E}$ positive.

For this method, the Fisher's information V can be determined only when there are deaths in each of the intervals. If any interval has no deaths, this interval should be combined with a neighbouring interval in such a way as to make all intervals contain deaths.

When there is no censoring, these expressions for Z and V reduce to

$$Z = \frac{r_{1,E}d_C - r_{1,C}d_E}{r_{1,E} + r_{1,C}} \quad \text{and} \quad V = \frac{r_{1,E}r_{1,C}sd}{(r_{1,E} + r_{1,C})^3},$$

where d_E and d_C denote the total numbers of events within time τ observed on the experimental and control treatments respectively, $d = d_E + d_C$, and $s = s_{h,E} + s_{h,C}$ is the total number of survivors. Further, $r_{1,E} = o_{1,E} + s_{1,E}$ and $r_{1,C} = o_{1,C} + s_{1,C}$, the total number of patients in the trial on experimental and control treatments, respectively. The statistic Z^2/V is the familiar statistic of the χ^2 test for a 2×2 contingency table.

Estimates of the magnitude of the treatment effect θ and corresponding confidence intervals can be calculated using any method that allows for interim analyses.

3. The Kaplan–Meier Method

In the censored binary method, the test statistics Z and V satisfy the approximate relationships $E(Z) = \theta V$ and $\text{var}(Z) = V$ mentioned in Section 1, where θ is the log odds ratio. A natural alternative is to use the Kaplan–Meier estimates $\hat{p}_E$ and $\hat{p}_C$ and to equate Z/V to the resulting estimate of θ, i.e.,

$$\frac{Z}{V} = \log\left\{\frac{\hat{p}_E(1-\hat{p}_C)}{\hat{p}_C(1-\hat{p}_E)}\right\}. \tag{3.1}$$

Here the data are used in the grouped form introduced in Section 2, although the ungrouped version now becomes simple to apply. From Greenwood's formula,

$$\text{var}(\hat{p}_\text{E} - \hat{p}_\text{C}) = \left(\hat{p}_\text{C}^2 W_\text{C} + \hat{p}_\text{E}^2 W_\text{E}\right), \tag{3.2}$$

where

$$W_\text{C} = \sum_{i=1}^{h} \left\{ \frac{o_{i\text{C}}}{s_{i\text{C}}(o_{i\text{C}} + s_{i\text{C}})} \right\} \quad \text{and} \quad W_\text{E} = \sum_{i=1}^{h} \left\{ \frac{o_{i\text{E}}}{s_{i\text{E}}(o_{i\text{E}} + s_{i\text{E}})} \right\}.$$

The variance of Z/V can be related to $\text{var}(\hat{p}_\text{E} - \hat{p}_\text{C})$ using a Taylor series approximation. A first-order Taylor series approximation of the logit of $\hat{p}_\text{E}$ is

$$\log\left(\frac{\hat{p}_\text{E}}{1-\hat{p}_\text{E}}\right) = \log\left(\frac{p_\text{E}}{1-p_\text{E}}\right) + (\hat{p}_\text{E} - p_\text{E})\left(\frac{1}{p_\text{E}} + \frac{1}{1-p_\text{E}}\right).$$

Using this result, a first-order Taylor series approximation of Z/V about θ is

$$\frac{Z}{V} = \theta + (\hat{p}_\text{E} - p_\text{E})\left(\frac{1}{p_\text{E}} + \frac{1}{1-p_\text{E}}\right) - (\hat{p}_\text{C} - p_\text{C})\left(\frac{1}{p_\text{C}} + \frac{1}{1-p_\text{C}}\right).$$

Under H_0: $\theta = 0, p_\text{E} = p_\text{C} = \bar{p}$ (say), and so Z/V can be expressed as

$$\frac{Z}{V} = (\hat{p}_\text{E} - \hat{p}_\text{C})\frac{1}{\bar{p}(1-\bar{p})}.$$

Thus, the null variance of Z/V is

$$\text{var}\left(\frac{Z}{V}\right) = \text{var}(\hat{p}_\text{E} - \hat{p}_\text{C})\frac{1}{\bar{p}^2(1-\bar{p})^2},$$

where $\bar{p}$ can be estimated by $(\hat{p}_\text{E} + \hat{p}_\text{C})/2$ or from a common Kaplan–Meier curve. Using the first of these options, taking $\text{var}(Z/V) = 1/V$ and using expression (3.2), the following expressions can be found for Z and V:

$$Z = \log\left\{\frac{\hat{p}_\text{E}(1-\hat{p}_\text{C})}{\hat{p}_\text{C}(1-\hat{p}_\text{E})}\right\} V$$

and

$$V = \frac{\hat{p}^2(1-\hat{p})^2}{\hat{p}_\text{C}^2 W_\text{C} + \hat{p}_\text{E}^2 W_\text{E}},$$

where

$$\hat{p} = (\hat{p}_\text{E} + \hat{p}_\text{C})/2.$$

When there is no censoring, these expressions for Z and V reduce to

$$Z = \log\left(\frac{s_{h,\text{E}} d_\text{C}}{s_{h,\text{C}} d_\text{E}}\right) V; \qquad V = \frac{(r_{1,\text{E}} s_{h,\text{C}} + r_{1,\text{C}} s_{h,\text{E}})^2 (r_{1,\text{E}} d_\text{C} + r_{1,\text{C}} d_\text{E})^2}{16 r_{1,\text{E}} r_{1,\text{C}} (r_{1,\text{E}}^3 s_{h,\text{C}} d_\text{C} + r_{1,\text{C}}^3 s_{h,\text{E}} d_\text{E})}$$

in the notation of Section 2.3.

The method presented here is a natural extension of the ideas of Cox (1963) for the comparison of two binary parameters. The value of V for no censoring given above does not coincide with that given by Cox because it is evaluated under the null hypothesis. The approach is also related to the more recent paper by Lan and Zucker (1993).

4. Three Alternative Methods

In this paper, the performances of the censored binary method (method 1) and the Kaplan–Meier method (method 2) are compared with three alternative approaches. Method 3 is a version of the censored binary method based on first-order Taylor series approximations to determine roots of the $(2h-1)$-order polynomial (2.1). Considerable manipulation leads to approximate versions of Z and V as

$$Z = \frac{(\hat{p}_\text{E} - \hat{p}_\text{C})\left\{\hat{p}_\text{C}(1-\hat{p}_\text{E}) W_\text{C} + \hat{p}_\text{E}(1-\hat{p}_\text{C}) W_\text{E}\right\}}{(\hat{p}_\text{C} W_\text{C} + \hat{p}_\text{E} W_\text{E})^2}$$

and

$$V = \frac{\{\hat{p}_C (1 - \hat{p}_E) W_C + \hat{p}_E (1 - \hat{p}_C) W_E\}^2 \left(\hat{p}_C^2 W_C + \hat{p}_E^2 W_E\right)}{\left(\hat{p}_C W_C + \hat{p}_E W_E\right)^4}.$$

In Section 5, it is shown that this method achieves the power specification accurately, but the example of Section 6 revealed that V is not always monotonically increasing during the study. The latter property would cause problems in application.

Method 4 is a binary approach that uses only completed binary observations of survival to time τ and ignores the censored observations. In this method failures cannot be used until time τ has elapsed since recruitment of the patient. This is to avoid inflating hazard estimates because, even though information about failures may be known early, information on survivors will be known only once time τ elapses after recruitment. Method 5 is the modified log rank test for interval-censored data explained in Section 3.5 of Whitehead (1997). For five intervals or more, this is approximately equivalent to the log rank test itself. It is used here to represent the log rank test in a form that lends itself more easily to an investigation by simulation.

5. Sequential Comparisons of Survival

5.1 *Stopping Boundaries of the Triangular Test*

The triangular test is used here to illustrate sequential methodology based on the statistics derived in the previous section. It is derived from the boundaries approach and is a descendant of the sequential probability ratio test of Wald (1947). In this approach, test statistics Z_i and V_i are calculated at the ith interim inspection ($i = 1, 2, \ldots$) and are plotted against each other until certain stopping boundaries are crossed or until V_i exceeds some maximum value V_{max}. The form of the stopping boundaries is asymmetric and can be determined from the following specifications:

(a) a specific value of the measure of treatment difference θ (θ_R) > 0, known as the reference improvement;
(b) the two-sided significance level (α) at which the null hypothesis is to be rejected;
(c) the power ($1 - \beta$) to reject the null hypothesis of no treatment difference at level α and to conclude that the experimental treatment is superior when $\theta = \theta_R$.

No restriction is placed on the probability, under $\theta = -\theta_R$, that H_0 is rejected and inferiority concluded. Figure 1 shows the continuous stopping boundaries of the triangular test. These boundaries will satisfy the power requirement only if monitoring is continuous, which is not feasible in practice. As inspections occur at discrete time points, it is possible for excursions outside boundaries to occur between inspections and not be observed. Therefore, a correction referred to as the Christmas tree correction is made. This is described in Whitehead (1997) and consists of the use of less stringent inner boundaries that make stopping easier when inspections do occur. The correction brings the boundaries in by the amount $0.583(V_i - V_{i-1})^{1/2}$ at the ith inspection, $i = 1, 2, \ldots$; $V_0 = 0$. The correction is illustrated in Figure 3, and it has been found to be accurate for the triangular test, in comparison with repeated numerical integration calculations, by Stallard

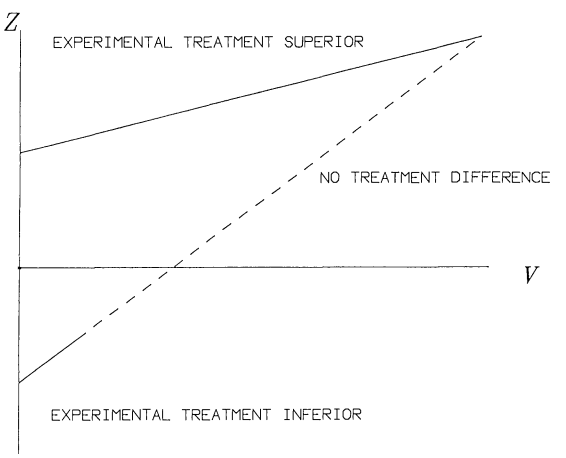

Figure 1. Continuous boundaries of the triangular test.

and Facey (1996). Given α, $1-\beta$, and θ_R, the package PEST3 (Brunier and Whitehead, 1993) can be used to determine the stopping boundaries. A property of the triangular test is that, asymptotically as α tends to zero, it minimises the maximum expected sample size among tests having the same power requirement.

5.2 *Simulation Study*

In the simulations, time was treated as discrete, with months chosen as the time unit. The cut points $t_1, \ldots, t_h$ were all taken to be whole numbers of months. At times $0, 1, \ldots, R$, batches of patients were recruited, and at times $1, 2, \ldots, F$ $(F \geq R)$, numbers of deaths were recorded. No patient was "dead on arrival"—it would be at least 1 month before a death could be notified.

The number of patients recruited to E at month i is denoted by $m_{i\mathrm{E}}$ and was generated as Poisson with mean $\lambda, i = 0, \ldots, R$. The number amongst this batch of $m_{i\mathrm{E}}$ who die after month $i + t_{j-1}$ and at or before month $i+t_j$ is denoted by $d_{ij\mathrm{E}}$ $(i+t_j \leq F)$. In terms of months since recruitment, their survival times lie in $(t_{j-1}, t_j]$, where t_{j-1} is excluded from the interval and t_j is included. The number of these $m_{i\mathrm{E}}$ subjects surviving beyond month $i + t_{j-1}$ is $m_{i\mathrm{E}} - D_{i(j-1)\mathrm{E}}$, where $D_{i(j-1)\mathrm{E}} = d_{i1\mathrm{E}} + \cdots + d_{i(j-1)\mathrm{E}}$. The value of $d_{ij\mathrm{E}}$ was generated from the binomial distribution, parameters $m_{i\mathrm{E}} - D_{i(j-1)\mathrm{E}}$ and $q_{j\mathrm{E}}$. Survival patterns on C were generated in a similar way.

The parameter values were selected as follows. The two-sided significance level was set at $\alpha = 0.05$ and power was set at $1 - \beta = 0.90$. The point of comparison of the two curves was taken to be $\tau = 12$ months. Five intervals $(h = 5)$ were considered. The cutpoints of the intervals were at $t_1 = 1, t_2 = 3, t_3 = 6, t_4 = 9$, and $t_5 = 12$ months, respectively. The value of p_{C} $(= p_{\mathrm{C}5})$ was taken to be 0.30. The reference improvement was taken to be $\theta_R = \log(2) = 0.693$ on the log scale, so that p_{E} was fixed under both the null and alternative hypotheses. The remaining survival probabilities, $p_{\mathrm{C}1}, \ldots, p_{\mathrm{C}4}, p_{\mathrm{E}1}, \ldots, p_{\mathrm{E}4}$, were filled in using the Weibull distribution. Three cases were used: shape parameters (b_{E} and b_{C}) equal (proportional hazards), $b_{\mathrm{E}} < b_{\mathrm{C}}$ (giving curves that cross), $b_{\mathrm{E}} > b_{\mathrm{C}}$ (giving divergent hazards). Figure 2 illustrates the survival curves considered. The number of patients entering per treatment group per month (λ) was taken to be five.

Each case was simulated separately under H_0 and H_1. The number of simulations carried out for each case was 10,000. Five methods of analysis were compared. Interim inspections of the data were conducted after each month.

5.3 *Simulation Results*

Table 1 gives, under H_0 and H_1, the observed proportions of rejections of H_0, with the conclusion that the experimental treatment is superior, the average and 95th percentile of duration, and the average and 95th percentile of sample size for the three shapes of survival curves and five methods of analysis.

The 95% probability interval for the estimate of true power, based on an approximate value of 0.90, is (0.894, 0.906). The 95% probability interval for wrongfully concluding experimental superiority when H_0 is true based on an approximate value of 0.025 is (0.022, 0.028). Methods 1, 3, and 4 are satisfactory with respect to significance level and power for all three cases. Method 2 is satisfactory with respect to power, but the significance level is slightly inflated. Method 5, which is the interval-censored log rank method, gives accurate power and significance level only when the shape parameters are equal, i.e., when the hazards are proportional. When the hazards are not proportional, method 5 gives very poor error rates. The average and 95th percentile of the duration are very similar for methods 1, 2, and 3. The average and the 95th percentile of duration of method 4 are higher than the corresponding values for methods 1, 2, and 3. This illustrates that method 4 is less efficient than methods 1, 2, and 3. When the hazards are proportional, method 5 gives much shorter durations than the other four methods. This illustrates that the log rank test is more powerful than all the other tests when the hazards are proportional. However, as shown by the results of Table 1, it gives completely unpredictable error rates when the hazards are not proportional. Note that, in the case of nonproportional hazards, type I error consists of rejecting H_0 when the survival probabilities past τ on E and C are equal. Other features of the survival patterns may differ, even under H_0. The results for sample size are similar to those for duration.

The main conclusion from the simulation study is that the censored binary method (method 1) is more accurate and efficient than the two conventional methods (methods 4 and 5). The Kaplan–Meier method (method 2) is as efficient as the more complex method 1 but less accurate than method 1. The Taylor series method (method 3) is as successful as method 1.

6. Example

The data used for illustration are drawn from a randomised clinical trial in locally advanced breast cancer conducted by the European Organisation for Research and Treatment of Cancer (EORTC)

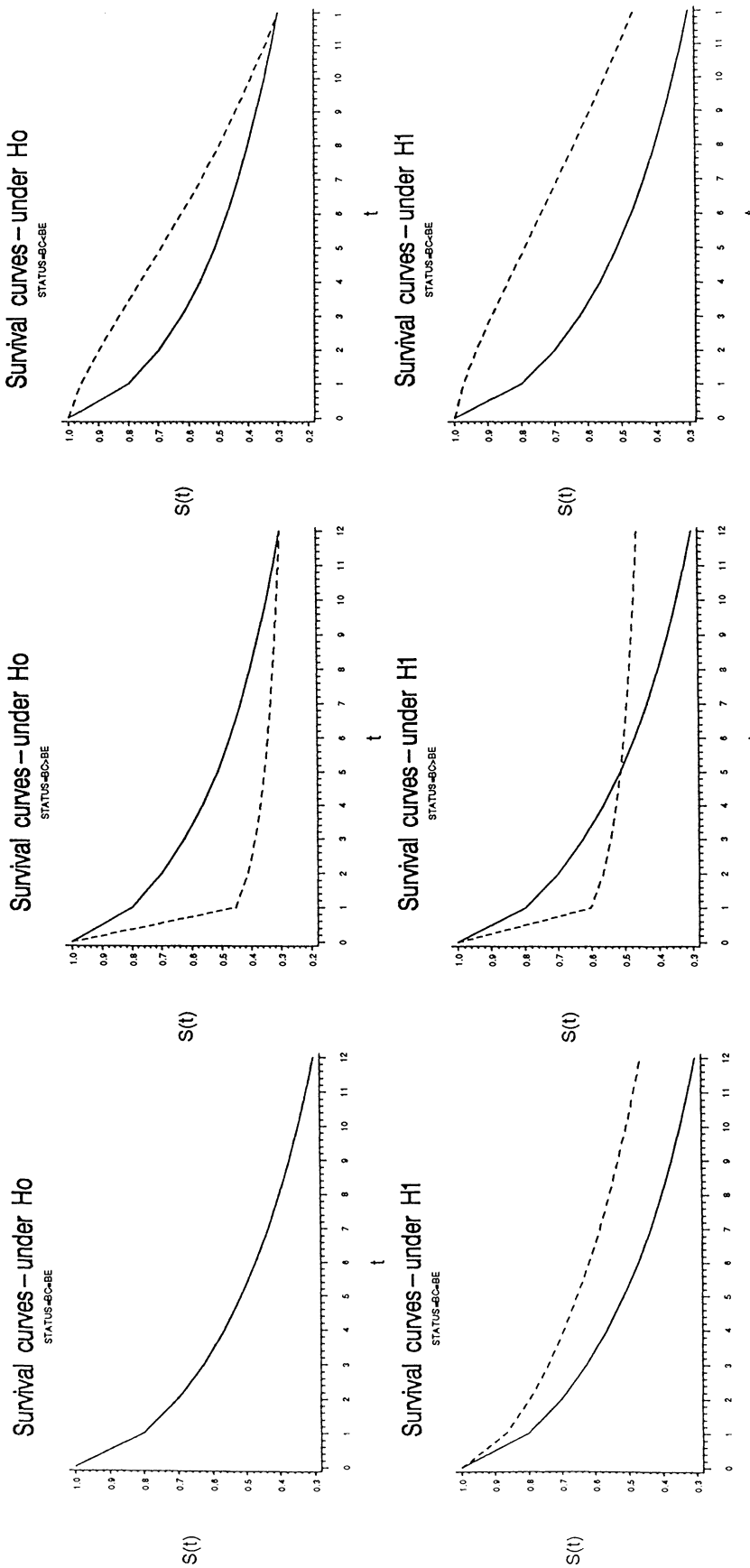

Figure 2. Survival curves of the control and experimental treatment used in the simulation procedure for $p_{Ch} = 0.30$ and $\theta_R = 0.693$. $S(t)$ denotes the survival function and t denotes the time in months. The solid line denotes the control and the dashed line denotes the experimental.

Table 1

Observed proportions of rejections of H_0 concluding that the experimental treatment is superior, the average and 95th percentiles of duration, and the average and 95th percentile of sample sizes for the three shape parameters and five methods of analysis

Shape parameters	Methods of analysis[a]				
	1	2	3	4	5
	Observed Proportion of Rejections of H_0 Under H_0 (Upper Figure) and H_1 (Lower Figure)				
$b_E = b_C$	0.027	0.033	0.027	0.024	0.026
	0.900	0.899	0.902	0.896	0.902
$b_E < b_C$	0.022	0.030	0.025	0.024	0.000
	0.895	0.899	0.896	0.894	0.103
$b_E > b_C$	0.025	0.032	0.028	0.028	0.579
	0.904	0.904	0.900	0.898	0.999
	Average and 95th Percentiles of Duration Under H_0 (Upper Figure) and H_1 (Lower Figure)				
$b_E = b_C$	30, 47	29, 46	30, 48	33, 51	22, 34
	31, 48	29, 47	31, 48	35, 53	27, 43
$b_E < b_C$	28, 45	28, 45	28, 46	34, 51	16, 20
	30, 46	29, 45	31, 47	35, 53	23, 42
$b_E > b_C$	30, 48	30, 47	31, 48	34, 51	28, 43
	32, 49	30, 48	32, 49	35, 52	19, 28
	Average and 95th Percentiles of Sample Size Under and H_0 (Upper Figure) H_1 (Lower Figure)				
$b_E = b_C$	298, 469	292, 463	298, 474	335, 509	215, 339
	311, 479	294, 470	314, 482	350, 529	269, 429
$b_E < b_C$	283, 449	281, 445	284, 462	336, 513	165, 207
	300, 459	287, 452	309, 470	352, 525	233, 414
$b_E > b_C$	304, 476	299, 474	307, 482	335, 512	278, 426
	317, 489	302, 481	321, 491	351, 523	185, 281

[a] Methods are (1) censored binary, (2) Kaplan–Meier, (3) Taylor series approximation, (4) complete binary, and (5) approximate logrank.

and reported in Rubens et al. (1989). A statistical discussion is given by Sylvester, Bartelink, and Rubens (1994). The aim of the trial was to study the efficacy of chemotherapy (CT) and hormone therapy (HT) after radiotherapy (RT). The trial was set up as a 2×2 factorial design with treatment combinations RT, RT+HT, RT+CT, and RT+CT+HT. There were 363 evaluable patients recruited during a period of approximately 6 years. The follow-up period was a further 4 years. The objective of our analysis was to examine the main effect of chemotherapy on mortality, stratifying for hormone therapy, using the censored binary method and the Kaplan–Meier method. The triangular test was adopted as the sequential design. The data available comprised the randomisation date and treatment allocation of all patients and the date of death of those who died. From this information, the status of each patient at each interim analysis could be determined. Of course, if these had been real, prospective interim analyses, then the statistician conducting them would not have had complete up-to-date information on all events that had occurred. No attempt was made in our simulations to model delays in data flow: one can imagine the specified timings as being cut-off dates for data capture, with actual interims being performed about 1 month later.

In the sequential design that we imposed to reanalyse this completed trial, an overall two-sided significance level of $\alpha = 0.05$ and a power of $1 - \beta = 0.90$ were used. In two separate reanalyses, the survival curves were compared at $\tau = 1$ year and $\tau = 3$ years. The reference improvement in terms of the log odds ratio was set at 0.693. This corresponds, e.g., to survival probabilities at 1 year of 0.82 for experimental and 0.70 for control and at 3 years of 0.57 for experimental and 0.40 for control. The first interim analysis was done 6 months after τ years, and thereafter analyses were conducted at 6 monthly intervals. For $\tau = 1$ year, the interval cutpoints used were 1, 3, 6, 9, and 12 months. For $\tau = 3$, the interval cutpoints used were 6, 12, 18, 24, and 36 months.

Table 2 gives the results of the chemotherapy main effect at $\tau = 1$ year for each interim analysis until termination for the censored binary method and Kaplan–Meier method. The table gives the

Table 2

Results of the chemotherapy main effect at $\tau = 1$ year for each interim analysis until termination for the censored binary method and the Kaplan–Meier method

Inspection	Month	Number of patients	Method: Censored binary Z_i	Method: Censored binary V_i	Method: Kaplan–Meier Z_i	Method: Kaplan–Meier V_i
1	18	79	0.222	0.730	0.220	0.748
2	24	120	0.091	1.251	0.091	1.251
3	30	178	1.917	2.503	2.424	2.593
4	36	222	2.626	2.849	3.149	2.938
5	42	251	3.233	4.002	3.408	4.059
6	48	287	5.086	5.328	5.556	5.456
7	54	319	4.610	5.612	4.877	5.697
8	60	349	4.576	6.111	4.812	6.186
9	66	361	3.580	6.586	3.746	6.635

interim inspection number (i), the corresponding month, the number of patients recruited, and the values of Z_i and V_i. For both methods, the trial continues until the 66th month and ends without a conclusion. The amount of data collected in the study was insufficient for reliable comparison of the 1-year survival rates, and a sequential design would have continued further. At each interim inspection, Z and V were calculated separately for the two strata formed by the hormone effect. In order to test the null hypothesis, the overall values of Z and V were found by summing the values of Z and V over the two strata at each interim inspection. Table 3 is similar to Table 2 and gives the results of the chemotherapy main effect at $\tau = 3$ years for each interim analysis until termination for the censored binary method and Kaplan–Meier method. For both methods, the trial stops at the 66th month due to lack of effect. This leads to the conclusion that, after 3 years, the benefit of chemotherapy is not significant. In this particular example, the pattern of survival curves is such that the conclusion concerning long-term survival (3 years) is clearer than that concerning short-term survival (1 year). Kaplan–Meier curves show a separation of survival curves at 1 and 2 years, with convergence at 3 years.

Tables were also constructed for the Taylor series method (method 3), but they are not presented here. It was observed that, for the Taylor series method, Fisher's information V was not monotonically increasing with the number of deaths but fluctuated, sometimes moving backwards by as much as two to three units. This is an undesirable property and is probably due to the fact that second- and higher-order terms were dropped in the approximations used for determining Z and V.

Figure 3 gives the boundaries of the triangular test with the Christmas tree correction and the sample path for the comparison of chemotherapy versus no chemotherapy at $\tau = 3$ years, stratified for hormone therapy and using the censored binary method. Analysis from PEST gives a p-value of 0.959. The median unbiased estimate of θ is -0.0155 and the 95% confidence interval for θ is $(-0.591, 0.599)$.

Table 3

Results of the chemotherapy main effect at $\tau = 3$ years for each interim analysis until termination for the censored binary method and the Kaplan–Meier method

Inspection	Month	Number of patients	Method: Censored binary Z_i	Method: Censored binary V_i	Method: Kaplan–Meier Z_i	Method: Kaplan–Meier V_i
1	42	251	−0.166	1.351	−0.209	1.363
2	48	287	−0.730	3.220	−0.839	3.421
3	54	319	0.725	5.833	0.726	6.191
4	60	349	2.619	9.122	2.606	8.992
5	66	361	−1.241	12.935	−1.303	13.102

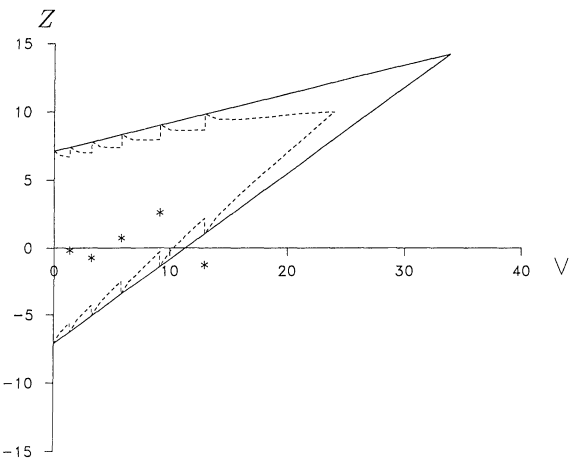

Figure 3. The triangular test with the Christmas tree correction and the sample path for the chemotherapy effect at 3 years using the censored binary method.

7. Discussion

The censored binary method (method 1) is an accurate and valid method for comparing survival curves when the assumption of proportional hazards is invalid or in doubt. It is more efficient than an analysis based on completed binary responses (method 4) as the information on deaths can be used as early as time t_1, long before the time τ at which point the survival curves are compared. Although it is less efficient than the log rank test (method 5) when the assumption of proportional hazards holds, it is far more accurate than the log rank test in cases where this assumption is invalid. An attractive property of the censored binary method is that it becomes the standard approach to 2×2 tables when no censoring is present. The censored binary method requires the solution of a polynomial of order $2h - 1$, where h is the number of intervals. Thus, this method is computationally complex and requires special software for determining the values of Z and V. The Kaplan–Meier method (method 2) is as efficient but somewhat less accurate than the censored binary method. However, the Kaplan–Meier method is computationally very simple and requires no new software for the derivation of Z and V. The approximate method based on the Taylor series approximation (method 3) is perhaps inappropriate for practical use as the test statistic V based on this method is not a monotonically increasing function of the number of deaths.

The model, parameter, and statistics introduced here could be used with other forms of sequential design, including the α-spending function methods of Lan and DeMets (1983) and the designs incorporated into the software package EaSt (Cytel, 1992). In fact, the choices "one-sided test" and "early rejection of H_0 or H_1" in EaSt lead to a design very similar to the triangular test. In view of the accuracy demonstrated in the case of the triangular test, it seems likely that the method would be accurate for other designs, too. The Christmas tree correction makes simulation of triangular tests relatively easy. The unequal and unpredictable increments of information inherent in this method mean that, in most other methods, individual stopping rules would have to be calculated separately for each of the 10,000 replicate simulations, rendering the exercise more computationally onerous.

When the censored binary method was compared with the full binary method (method 4) in the simulations, the reduction in duration was approximately 3 months, with a modest decrease in the sample size of 50 patients. In a case where one treatment group is seriously disadvantaged, the reduction in duration will be greater. If a naive binary method that uses deaths as soon as they occur is adopted, problems will arise when there is early transient elevation of hazard in one group due, e.g., to surgery.

Acknowledgements

The authors thank the Association of Commonwealth Universities (ACU) for funding this research and the European Organisation for Research and Treatment of Cancer (EORTC) for providing and allowing for use of the data set on breast cancer.

Résumé

Deux tests sont proposés pour comparer les courbes de survie de patients randomisés entre un groupe expérimental et un groupe contrôle, quand l'hypothèse des risques proportionnels peut ne pas être satisfaite. Les tests sont particulièrement intéressants lorsques les courbes de survie sont proches ou se croisent tôt dans la période de follow-up et divergent ensuite. Les tests comparent les probabilités de survie après un temps fixé depuis la randomisation, pour les 2 groupes de patients. Les 2 méthodes prennent en compte les données censurées à droite et sont toutes deux associées avec des méthodes estimant et déterminant les limites de confiance pour les différences de traitement. La première méthode est une approche directe dérivée de la statistique du score d'efficience et de l'information de Fisher. La deuxième méthode est plus simple et est basée sur les estimations de Kaplan-Meier et de leurs variances. Les méthodes conventionnelles d'estimation de taille d'échantillon nécessitent l'hypothèse de hasards proportionnels. Une approche séquentielle est utilisée à cause de la difficulté de déterminer la taille d'échantillon par avance sans hypothèses fortes sur les relations entre les 2 courbes de survie. Les résultats de simulations donnant de l'information, sur le seuil de signification et la puissance des tests proposés, sont donnés et les tests sont appliqués aux données d'un essai clinique dans le cancer du sein.

References

Bartlett, M. S. (1946). The large sample theory of sequential tests. *Proceedings of the Cambridge Philosophical Society* **42,** 239–244.

Brunier, H. and Whitehead, J. (1993). *PEST 3.0 Operating Manual.* Reading, U.K.: Reading University.

Cox, D. R. (1963). Large sample sequential tests for composite hypotheses. *Sankhyā Series A* **25,** 5–12.

Cytel. (1992). *East: A software Package for the Design and Monitoring of Group Sequential Clinical Trials.* Cambridge, Massachusetts: Cytel Software Corporation.

Freedman, L. S. (1982). Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine* **1,** 121–129.

Gregory, W. M., Bolland, K., Whitehead, J., and Souhami, R. L. (1997). Cautionary tales of survival analysis: Conflicting analysis from a clinical trial in breast cancer. *British Journal of Cancer* **76,** 551–558.

Kim, K. (1992). Study duration for group sequential clinical trials with censored survival data adjusting for stratification. *Statistics in Medicine* **11,** 1477–1488.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70,** 659–663.

Lan, K. K. G. and Zucker, D. M. (1993). Sequential monitoring of clinical trials: The role of information and Brownian motion. *Statistics in Medicine* **12,** 753–765.

Machin, D. and Campbell, M. J. (1987). *Statistical Tables for the Design of Clinical Trials.* Oxford: Blackwell.

Rubens, R. D., Bartelink, H., Engelsman, E., Hayward, J. L., Rotmensz, N., Sylvester, R., Van der Schueren, E., Papadiamantis, J., Vassilaros, S. D., Wilders, J., and Winters, P. J. (1989). Locally advanced breast cancer: The contribution of cytotoxic and endocrine treatment to radiotherapy. *European Journal of Cancer and Clinical Oncology* **25,** 667–678.

Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika* **70,** 315–326.

Stallard, N. and Facey, K. M. (1996). Comparison of the spending function method and the Christmas tree correction for group sequential trials. *Journal of Biopharmaceutical Statistics* **6,** 361–373.

Sylvester, R., Bartelink, H., and Rubens, R. (1994). A reversal of fortune: Practical problems in the monitoring and interpretation of an EORTC breast cancer trial. *Statistics in Medicine* **13,** 1329–1335.

Tsiatis, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association* **77,** 855–861.

Tsiatis, A. A., Boucher, H., and Kim, K. (1995). Sequential methods for parametric survival models. *Biometrika* **82,** 165–173.

Wald, A. (1947). *Sequential Analysis.* New York: Wiley.

Whitehead, J. (1978). Large sample sequential methods with application to the analysis of 2×2 contingency tables. *Biometrika* **65,** 351–356.

Whitehead, J. (1984). The sequential analysis of survival data. *Medizinische Informatik und Statistik* **56,** 115–123. Berlin: Springer-Verlag.

Whitehead, J. (1989). The analysis of relapse clinical trials, with application to a comparison of two ulcer treatments. *Statistics in Medicine* **8,** 1439–1454.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials,* revised 2nd edition. New York: Wiley.

Received February 1996; *revised July* 1997; *accepted September* 1997.