25

Group Sequential Designs for Survival Data

CONTENTS

Chris Jennison, University of Bath, UK

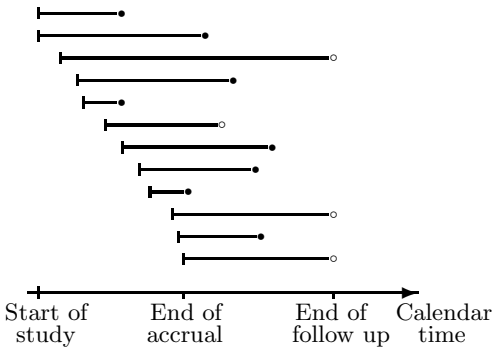
Bruce Turnbull, Cornell University, Ithaca, New York, USA

25.1 Introduction

Consider an experiment or study where entry of subjects is staggered over time. We are interested in a survival or "time to event" response, measured from entry into the trial. The subjects are followed for a certain duration until their event time is observed or censored. The situation is depicted in Figure 25.1 with the horizontal lines in the diagram representing survival times of twelve subjects. A solid circle at the right hand end designates an exact observation (subjects 1, 2, 4, 5, 7, 8, 9 and 11), whereas a hollow circle indicates that the survival time is censored. Note that censoring can occur because of end-of-study (subjects 3, 10 and 12) or for some other reason such as competing risk or loss to follow-up (subject 6). This situation is common in the conduct of clinical trials. Of course, the situation where all subjects start together at the beginning is a special case and this is more common in engineering or product life-testing experiments.

Consider the problem of testing between two hypotheses H_0 and H_1 concerning some parameter θ. The data are analysed not just at the planned end of the study, but also at interim times at calendar time points during

FIGURE 25.1
Accrual and follow up in a survival study



the course of the study, with a maximum of $K > 1$ analyses. At the interim analyses, the decision can be made to stop the study concluding either H_0 or H_1, or to continue on to the next analysis. Figure 25.2 illustrates the case of three analyses. At an interim analysis, subjects are censored if they are still known to be alive at this point. Information on such subjects will continue to accrue at later analyses.

At the first interim analysis, we analyze data on elapsed survival times from randomization. These times have a common starting point of zero and "analysis time" censoring occurs for subjects surviving past the first analysis; see Figure 25.3. Then, at interim analysis 2, we analyze data on survival from randomization time with "analysis time" censoring occurring for subjects surviving past the second analysis; see Figure 25.4. This process continues on through further analyses until the conclusion of the trial.

25.2 Canonical joint distribution of test statistics based on accumulating data

Suppose our main interest is in the parameter θ and let $\widehat{\theta}_k$ denote an estimate of θ based on data available at analysis k. For survival data, θ could be the hazard ratio between two survival distributions, assumed constant over time, or the coefficient for a treatment effect in a Cox (1972) regression model or other type of failure time model.

The information for θ at analysis k is

$$\mathcal{I}_k = \{\mathrm{Var}(\widehat{\theta}_k)\}^{-1}, \quad k = 1, \ldots, K.$$

FIGURE 25.2
Interim analyses

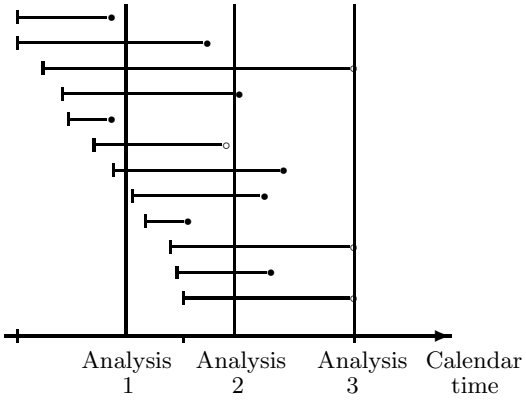


FIGURE 25.3
Interim analysis 1

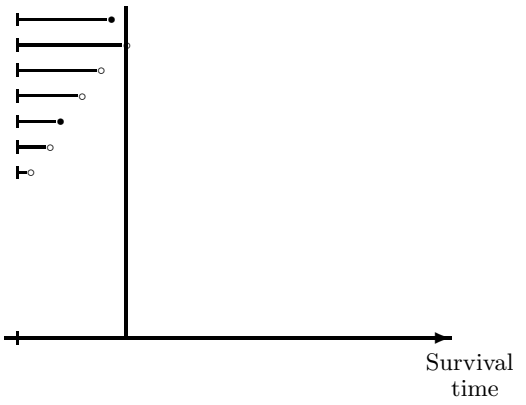
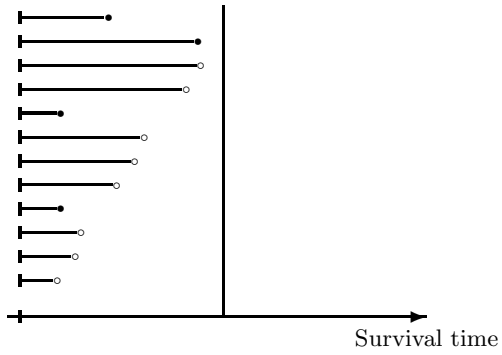


FIGURE 25.4
Interim analysis 2



In many situations, $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ are approximately multivariate normal,

$$\widehat{\theta}_k \sim N(\theta, \{\mathcal{I}_k\}^{-1}), \quad k = 1, \ldots, K,$$

and

$$\mathrm{Cov}(\widehat{\theta}_{k_1}, \widehat{\theta}_{k_2}) = \mathrm{Var}(\widehat{\theta}_{k_2}) = \{\mathcal{I}_{k_2}\}^{-1} \quad \text{for } k_1 < k_2.$$

This is termed the *canonical joint distribution*. It occurs, for example, when $\widehat{\theta}$ is a maximum likelihood estimate or other consistent asymptotically efficient estimator; see Scharfstein, Tsiatis and Robins (1997) and Jennison and Turnbull (1997).

For testing H_0: $\theta = 0$, the *standardized statistic* at analysis k is

$$Z_k \;=\; \frac{\widehat{\theta}_k}{\sqrt{\mathrm{Var}(\widehat{\theta}_k)}} \;=\; \widehat{\theta}_k \sqrt{\mathcal{I}_k}.$$

For this statistic, the canonical joint distribution of $(\widehat{\theta}_1, \ldots, \widehat{\theta}_K)$ implies that

$(Z_1, \ldots, Z_K)$ is multivariate normal,

$Z_k \sim N(\theta\sqrt{\mathcal{I}_k}, 1), \quad k = 1, \ldots, K,$

$\mathrm{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}} \quad \text{for } k_1 < k_2.$

The *score statistics*, $S_k = Z_k\sqrt{\mathcal{I}_k}$, are also approximately multivariate normal with

$$S_k \sim N(\theta\,\mathcal{I}_k,\, \mathcal{I}_k), \quad k = 1, \ldots, K.$$

The score statistics possess the "independent increments" property,

$$\mathrm{Cov}(S_k - S_{k-1},\, S_{k'} - S_{k'-1}) = 0 \quad \text{for } k \neq k'.$$

For computation it is useful to recognize the fact that these score statistics behave as Brownian motion with drift θ observed at times $\mathcal{I}_1, \ldots, \mathcal{I}_K$.

In testing the equality of two survival curves, the successive non-standardized log-rank statistics have, asymptotically, the canonical joint distribution of a sequence of score statistics. Here $\widehat{\theta}$ is an estimate of the log hazard ratio θ in a proportional hazards model and the information $\mathcal{I}$ for θ is roughly equal to a quarter of the number of observed events. The canonical distribution also applies to stratified log-rank statistics; see Jennison and Turnbull (2000, Sec. 13.6.2).

If a Cox (1972) proportional hazards regression model is fitted by maximum partial likelihood, the canonical joint distribution holds approximately for successive estimates of a regression coefficient. Kaplan-Meier (1958) estimates of survival probabilities at a fixed time point or of a specified quantile (e.g., the median) also follow the canonical joint distribution; see Section 25.7.

25.3 Group sequential boundaries and error spending

Suppose we are interested in testing the null hypothesis H_0: $\theta = 0$ versus a one-sided or two-sided alternative hypothesis H_1. At each interim analysis or "stage", we must decide whether to continue the study or to terminate, concluding either H_0 or H_1. At each stage k, $k = 1, \ldots, K$, this decision is based on a statistic Z_k according to the rule

If $Z_k \in \mathcal{C}_k$, continue on to stage $k+1$,

if $Z_k \in \mathcal{A}_k$, stop and conclude H_0,

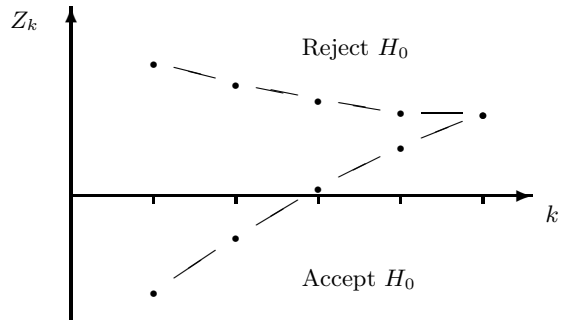
if $Z_k \in \mathcal{B}_k$, stop and conclude H_1,

where $\mathcal{A}_k$, $\mathcal{B}_k$ and $\mathcal{C}_k$ are disjoint and exhaustive subsets of the real line, so $\mathcal{A}_k \cup \mathcal{B}_k \cup \mathcal{C}_k = (-\infty, \infty)$, and we set $\mathcal{C}_K = \emptyset$ in order that the procedure terminates at stage K.

Here, we shall consider the case of one-sided tests for superiority. Results for tests of non-inferiority, two-sided tests and equivalence tests can be developed analogously; see Jennison and Turnbull (2000). In a one-sided test where positive θ values are desirable the hypotheses are H_0: $\theta \leq 0$ and H_1: $\theta > 0$. The type 1 error probability constraint is

$$P_{\theta=0}\{\text{Reject } H_0\} = \alpha \tag{25.1}$$

FIGURE 25.5
A group sequential boundary



and the type 2 error probability is specified through the power requirement at effect size δ,

$$P_{\theta=\delta}\{\text{Reject } H_0\} = 1 - \beta. \tag{25.2}$$

In this case, the continuation and stopping regions are $\mathcal{A}_k = (-\infty, a_k)$, $\mathcal{B}_k = (b_k, \infty)$ and $\mathcal{C}_k = (a_k, b_k)$, where $a_k \leq b_k$ for $k = 1, \ldots, K-1$, and $a_K = b_K$. A typical boundary with critical values $\{(a_k, b_k)\}$ is depicted in Figure 25.5.

The upper boundary, $\{b_k\}$, is often termed the *efficacy* boundary and the lower boundary, $\{a_k\}$, the *futility* boundary. The role of the futility boundary and whether it will be used for guidance or as a binding rule affects the construction of the boundaries. With a ***binding futility boundary***, it is assumed that crossing the lower boundary will definitely lead to stopping and acceptance of H_0, and the type I error probability is calculated as

$$\sum_{k=1}^{K} P_{\theta=0}\{a_1 < Z_1 < b_1, \ldots, a_{k-1} < Z_{k-1} < b_{k-1}, Z_k > b_k\}.$$

A ***non-binding futility boundary*** is appropriate if the study may possibly continue after crossing the lower boundary, so a type I error can still occur. In this case, the type I error probability is calculated as

$$\sum_{k=1}^{K} P_{\theta=0}\{Z_1 < b_1, \ldots, Z_{k-1} < b_{k-1}, Z_k > b_k\}.$$

In either case the type II error probability is calculated as

$$\sum_{k=1}^{K} P_{\theta=\delta}\{a_1 < Z_1 < b_1, \ldots, a_{k-1} < Z_{k-1} < b_{k-1}, Z_k < a_k\}.$$

The Pampallona and Tsiatis (1994) family provides a selection of one-sided group sequential tests. The test with index Δ has critical values of the form

$$\begin{aligned} b_k &= \tilde{C}_1 \, (\mathcal{I}_k/\mathcal{I}_K)^{\,\Delta-0.5}, \\ a_k &= \delta \sqrt{\mathcal{I}_k} - \tilde{C}_2 \, (\mathcal{I}_k/\mathcal{I}_K)^{\,\Delta-0.5}, \quad k = 1, \ldots, K. \end{aligned}$$

Given a specified pattern of information levels, for example, equally spaced values $\mathcal{I}_k = (k/K)\,\mathcal{I}_K$, $k = 1, \ldots, K$, and a choice of binding or non-binding futility boundary, constants $\mathcal{I}_K, \tilde{C}_1$ and $\tilde{C}_2$ can be found such that $a_K = b_K$ and the error probability constraints (25.1) and (25.2) are satisfied.

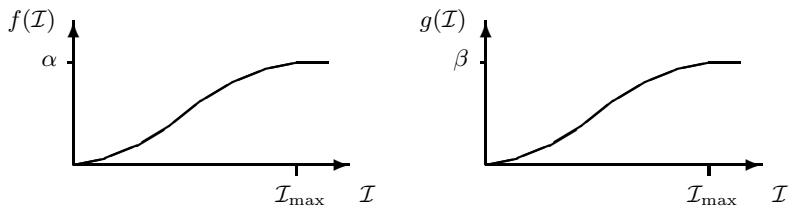
However, for survival data statistics such as those mentioned above, it is impractical to schedule the interim analyses at equal or pre-specified increments of information. Indeed, the increments in information will be both unequal and unpredictable. For example, the information for the log-rank statistic (approximately one quarter of the number of observed events) will only become known at the time of an analysis. Information for a treatment effect in a Cox (1972) regression model or a survival probability or quantile is similarly unpredictable. Thus we shall need to use the error spending approach of Lan and DeMets (1983) in which type I and II error probabilities are "spent" as functions of the observed information.

For a one-sided test of H_0: $\theta \leq 0$ against H_1: $\theta > 0$, we need two functions to spend

Type I error probability α under $\theta = 0$,

Type II error probability β under $\theta = \delta$.

A *maximum information design* works towards a target information level $\mathcal{I}_{\max}$. The type I error probability α spending function $f(\mathcal{I})$ rises from zero to α as $\mathcal{I}$ increases from zero to $\mathcal{I}_{\max}$. Similarly, the type II error spending function $g(\mathcal{I})$ rises from zero at $\mathcal{I} = 0$ to β at $\mathcal{I} = \mathcal{I}_{\max}$.



In implementing this error spending design, boundaries at each interim analysis, k, are constructed so that the cumulative type I error probability thus far is $f(\mathcal{I}_k)$ and the cumulative type II error probability is $g(\mathcal{I}_k)$. This calculation can be carried out treating the futility as binding or non-binding, as required.

At analysis 1:

The observed information is $\mathcal{I}_1$.

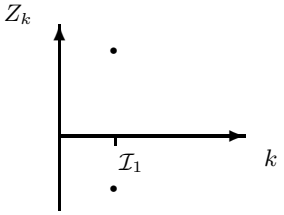
We reject H_0 if $Z_1 > b_1$, where

$$P_{\theta=0}\{Z_1 > b_1\} = f(\mathcal{I}_1)$$

and we accept H_0 if $Z_1 < a_1$, where

$$P_{\theta=\delta}\{Z_1 < a_1\} = g(\mathcal{I}_1).$$

Solving these equations determines the critical values a_1 and b_1.



At analysis 2:

The observed information is $\mathcal{I}_2$.

We reject H_0 if $Z_2 > b_2$ where, for a binding futility boundary,

$$P_{\theta=0}\{a_1 < Z_1 < b_1, Z_2 > b_2\} = f(\mathcal{I}_2) - f(\mathcal{I}_1).$$

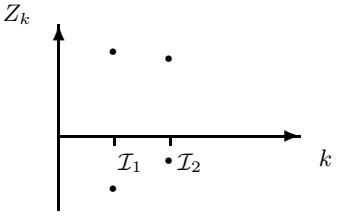
or, for a non-binding futility boundary,

$$P_{\theta=0}\{Z_1 < b_1, Z_2 > b_2\} = f(\mathcal{I}_2) - f(\mathcal{I}_1).$$

We accept H_0 if $Z_2 < a_2$, where

$$P_{\theta=\delta}\{a_1 < Z_1 < b_1, Z_2 < a_2\} = g(\mathcal{I}_2) - g(\mathcal{I}_1).$$

In either case, since a_1 and b_1 have been fixed at the previous analysis, we can solve these equations for a_2 and b_2.



At a general analysis k:

The observed information is $\mathcal{I}_k$.

We reject H_0 if $Z_k > b_k$ where, for a binding futility boundary,

$$P_{\theta=0}\{a_1 < Z_1 < b_1, \ldots, a_{k-1} < Z_{k-1} < b_{k-1}, Z_k > b_k\} = f(\mathcal{I}_k) - f(\mathcal{I}_{k-1}),$$

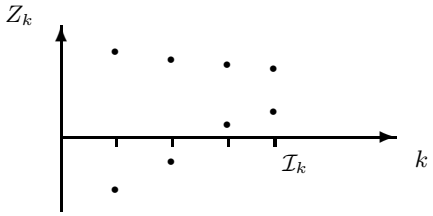
or, for a non-binding futility boundary,

$$P_{\theta=0}\{Z_1 < b_1, \ldots, Z_{k-1} < b_{k-1}, Z_k > b_k\} = f(\mathcal{I}_k) - f(\mathcal{I}_{k-1}).$$

We accept H_0 if $Z_k < a_k$, where

$$P_{\theta=\delta}\{a_1 < Z_1 < b_1, \ldots, a_{k-1} < Z_{k-1} < b_{k-1}, Z_k < a_k\} = g(\mathcal{I}_k) - g(\mathcal{I}_{k-1}).$$

Since $a_1, \ldots, a_{k-1}$ and $b_1, \ldots, b_{k-1}$ were determined at analysis $k-1$, these equations can be solved for a_k and b_k.



We remark that in the above description, the computation of a_k and b_k does ***not*** depend on future information levels, $\mathcal{I}_{k+1}, \mathcal{I}_{k+2}, \ldots$. The error spending design is fully determined once the maximum information, $\mathcal{I}_{\max}$, and the spending functions $f(\mathcal{I})$ and $g(\mathcal{I})$ have been specified, although the critical values will depend on the information levels actually observed. One would like the upper and lower boundaries to meet at a single point at the concluding analysis where $f(\mathcal{I}) = \alpha$ and $g(\mathcal{I}) = \beta$. The maximum information $\mathcal{I}_{\max}$ and functions $f(\mathcal{I})$ and $g(\mathcal{I})$ can be chosen so that this will happen when observed information levels follow a particular pattern, but it is important to be able to handle other observed sequences $\mathcal{I}_1, \mathcal{I}_2, \ldots$.

A convenient choice of error spending functions is provided by the so-called ρ-family, for which

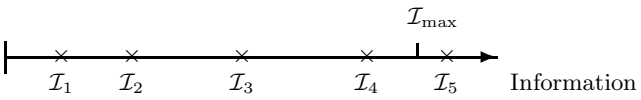
$$f(\mathcal{I}) = \alpha \min\{1, (\mathcal{I}/\mathcal{I}_{\max})^\rho\} \quad \text{and} \quad g(\mathcal{I}) = \beta \min\{1, (\mathcal{I}/\mathcal{I}_{\max})^\rho\}.$$

Values $\rho > 0$ can be used and common choices are $\rho = 1$, 2 or 3. Lower values of ρ correspond to plans with more aggressive early stopping. The value of $\mathcal{I}_{\max}$ should be chosen so that boundaries converge with $a_K = b_K$ at the final analysis under a typical sequence of information levels. So, for design purposes we might plan for a maximum of K analyses at equally spaced information

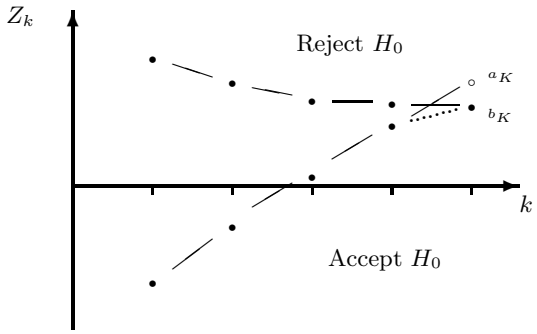
levels, $\mathcal{I}_k = (k/K)\mathcal{I}_{\max}$, $k = 1, \ldots, K$. Then, for each value of ρ there is an associated $\mathcal{I}_{\max}$ that should be used. Barber and Jennison (2002) show that the resulting ρ-family error spending tests have excellent efficiency properties when compared with other designs for the same number of analyses K and maximum information $\mathcal{I}_{\max}$.

Once the trial is running, the occurrences of events are unpredictable. Information levels may not follow the anticipated pattern and it may take more or fewer than K analyses to reach the target information level $\mathcal{I}_{\max}$. Thus, care is needed at the final analysis of a one-sided error spending test.

Over-running: If an analysis is reached with $\mathcal{I}_K > \mathcal{I}_{\max}$, solving the equations for a_K and b_K is liable to give $a_K > b_K$.

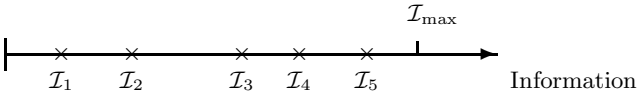


The value calculated for b_K will guarantee that the type I error probability is equal to α. So, in this case, we can reduce a_K to b_K and the power attained under $\theta = \delta$ will be greater than $1 - \beta$.

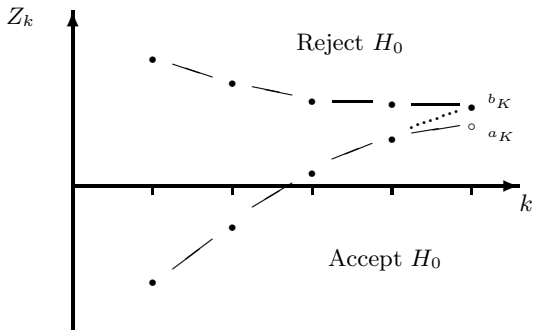


Even when $\mathcal{I}_K = \mathcal{I}_{\max}$, over-running may occur if information deviates from the pattern of, say, equally spaced values used in choosing $\mathcal{I}_{\max}$.

Under-running: A final information level $\mathcal{I}_K < \mathcal{I}_{\max}$ may be imposed as part of the trial design when a final planned analysis is reached, for example, after a maximum length of follow-up of subjects' survival.



In this situation, values $f(\mathcal{I}_K) = \alpha$ and $g(\mathcal{I}_K) = \beta$ are used in the equations for a_K and b_K. Since the information level at this point is lower than $\mathcal{I}_{\max}$, the solutions of these equations are liable to have $a_K < b_K$.



Again, with b_K as calculated, the type I error probability is exactly α. Here, we increase a_K to b_K in order to protect the type I error rate and the attained power at $\theta = \delta$ will be below the planned $1 - \beta$.

There is considerable freedom in implementing error spending group sequential designs. A series of analyses can be stipulated at fixed calendar times and the attained power will vary, depending on the observed information levels. Alternatively, amendments may be made to the original study plan, such as extending follow-up or adding centres to increase patient recruitment, in order to reach the target information $\mathcal{I}_{\max}$. One proviso to protect against any chance of bias in the claimed error probabilities is that such decisions should be made in response to observed information levels and not estimated treatment effects.

25.4 The group sequential log-rank test

We return to the problem of testing the equality of survival distributions $S_A(t)$ and $S_B(t)$ for two treatment arms, A and B, based on accumulating survival data. We denote the hazard rates on treatments A and B by $h_A(t)$ and $h_B(t)$, respectively. At each analysis we observe a failure or censoring time for each subject entered so far, measured from that subject's date of entry or randomization as defined in the study protocol. The way the set of data grows as patients are accrued and follow up on each patient lengthens was shown in Figures 25.1 to 25.4.

Let d_k, $k = 1, \ldots, K$, denote the total number of uncensored failures observed across both treatment arms when analysis k is conducted. Some of these times may be tied and we suppose that d'_k of the d_k failure times are distinct, where $1 \leq d'_k \leq d_k$. We denote these distinct failure times by $\tau_{1,k} < \tau_{2,k} < \ldots < \tau_{d'_k,k}$ and let $r_{iA,k}$ and $r_{iB,k}$ be the numbers at risk on treatment arms A and B, respectively, just before time $\tau_{i,k}$. Finally, we denote by $\delta_{iA,k}$ and $\delta_{iB,k}$ the numbers on treatment arms A and B that fail at time

$\tau_{i,k}$ and define $\delta_{i,k} = \delta_{iA,k} + \delta_{iB,k}$ for $i = 1, \ldots, d'_k$. If there are no ties, then $\delta_{i,k} = 1$ and either $\delta_{iA,k} = 1$ and $\delta_{iB,k} = 0$ or $\delta_{iA,k} = 0$ and $\delta_{iB,k} = 1$ for each pair i and k.

If the survival distributions $S_A(t)$ and $S_B(t)$ are equal, the conditional distribution of $\delta_{iB,k}$ given $r_{iA,k}$, $r_{iB,k}$ and $\delta_{i,k}$ is hypergeometric with expectation

$$e_{i,k} = \frac{r_{iB,k}\ \delta_{i,k}}{r_{iA,k} + r_{iB,k}}$$

and variance

$$v_{i,k} = \frac{r_{iA,k}\ r_{iB,k}\ \delta_{i,k}\,(r_{iA,k} + r_{iB,k} - \delta_{i,k})}{(r_{iA,k} + r_{iB,k} - 1)\,(r_{iA,k} + r_{iB,k})^2}. \tag{25.3}$$

The unstandardized log-rank statistic at analysis k is

$$S_k = \sum_{i=1}^{d'_k} (\delta_{iB,k} - e_{i,k})$$

and the standardized log-rank statistic is

$$Z_k = \frac{\sum_{i=1}^{d'_k} (\delta_{iB,k} - e_{i,k})}{\left(\sum_{i=1}^{d'_k}\ v_{i,k}\right)^{1/2}}. \tag{25.4}$$

The information $\mathcal{I}_k$ for the log hazard ratio is

$$\mathcal{I}_k = \sum_{i=1}^{d'_k} v_{i,k}. \tag{25.5}$$

The log-rank test has optimal power properties to detect alternatives when hazard rates in the two treatment arms are proportional, so $h_A(t) = \lambda h_B(t)$. The sequence of log-rank statistics defined by (25.4) then has, approximately, the canonical joint distribution for a sequence of Z-statistics, given $\mathcal{I}_1, \ldots, \mathcal{I}_K$, with $\theta = \log(\lambda)$, the log hazard ratio.

Since the canonical joint distribution holds, the methods described in Section 25.3 can be used to construct group sequential error spending tests from the sequence of statistics Z_k and information levels $\mathcal{I}_k$. In designing a maximum information trial to meet a given power requirement, it is necessary to predict the information levels that will arise, especially that at the final possible analysis. Here, it is helpful to note from (25.3) that each $v_{i,k}$ is approximately $\delta_{i,k}/4$ if $r_{iA,k} \approx r_{iB,k}$ and either $\delta_{i,k} = 1$ or $\delta_{i,k}$ is small relative to $r_{iA,k} + r_{iB,k}$. Hence, $\mathcal{I}_k$ will be approximately equal to $d_k/4$ and the final information level will be close to one quarter of the total number of observed failures. The illustrative example in the next section will show the usefulness of this approximation in planning the sample size and length of follow-up that may be necessary in a survival study.

25.5 Example: A clinical trial for carcinoma of the oropharynx

We illustrate the methods we have described by applying them to a clinical trial conducted by the Radiation Therapy Oncology Group in the U.S. to investigate treatments of carcinoma of the oropharynx. We use the data from six of the larger institutions participating in this trial as recorded by Kalbfleisch and Prentice (2002, Appendix II). Subjects were recruited to the study between 1968 and 1972 and randomized to a standard radiotherapy treatment or an experimental treatment in which the radiotherapy was supplemented by chemotherapy. The major endpoint was patient survival and patients were followed until around the end of 1973. Several baseline covariates, thought to have strong prognostic value, were also recorded.

TABLE 25.1
Summary data for oropharynx cancer clinical trial

Analysis		Number of subjects entered		Number of deaths	
k	Date	Treatment A	Treatment B	Treatment A	Treatment B
1	12/69	38	45	13	14
2	12/70	56	70	30	28
3	12/71	81	93	44	47
4	12/72	95	100	63	66
5	12/73	95	100	69	73

The conduct of the study did not follow a group sequential plan but, for purposes of illustration, we have reconstructed patients' survival times and their status, dead or censored, at times 720, 1080, 1440, 1800 and 2160 days from the beginning of 1968. This "reconstructed" data set was used by Jennison and Turnbull (2000, Ch. 13). A summary of the reconstructed data is given in Table 25.1: we used the precise death or censoring times in the reconstructed data to compute the statistics and information values in applying retrospectively a group sequential error spending design. As the central survival records would not have been updated continuously, our constructed data sets most likely resemble the information that would have been available at interim analyses conducted a month or two after these times, and so they are an approximation to the data that could have been studied by a monitoring committee meeting at dates a little after 2, 3, 4, 5 and 6 years from

TABLE 25.2
Design parameters for a group sequential procedure assuming equally spaced information levels, $\mathcal{I}_k = (k/5)\,\mathcal{I}_{\max}$, $k = 1, \ldots, 5$

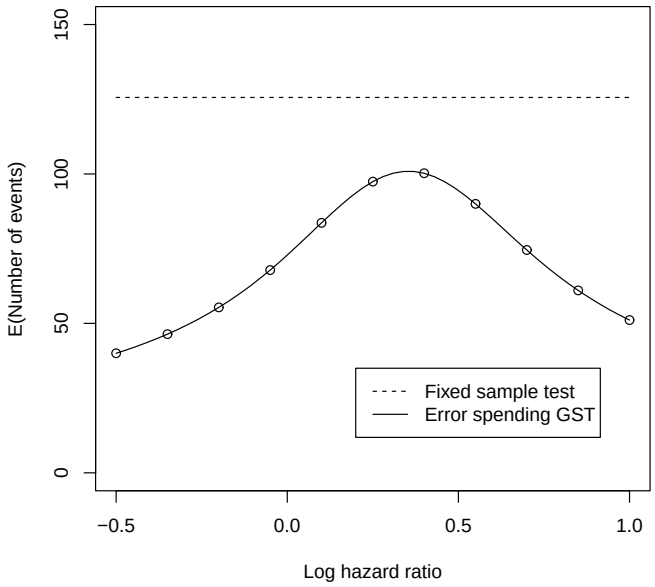
Design parameter	Binding futility boundary		Non-binding futility boundary	
$\mathcal{I}_{\max}$	34.48		35.58	
Maximum number of deaths	138		143	
a_1, b_1	−1.096	3.090	−1.075	3.090
a_2, b_2	−0.053	2.714	−0.023	2.714
a_3, b_3	0.722	2.473	0.758	2.473
a_4, b_4	1.387	2.276	1.429	2.280
a_5, b_5	2.055	2.055	2.114	2.114

the start of the study. The longer waiting period to the first interim analysis is intended to compensate for the slow initial accrual of survival information while only a few patients had been entered to the trial.

Since the experimental treatment involved chemotherapy as well as radiotherapy, the researchers would have been looking for a substantive improvement in survival on this treatment in return for the additional discomfort and short term health risks. A one-sided testing formulation is, therefore, appropriate and we shall conduct our retrospective interim analyses as a group sequential test of the null hypothesis of no treatment difference against the one-sided alternative that the new combination therapy is superior to the standard treatment of radiotherapy alone. For the sake of illustration, we suppose the experiment was designed to achieve a type I error probability of $\alpha = 0.025$ and power $1-\beta = 0.8$ when the log hazard ratio for the experimental treatment versus the standard is equal to 0.5.

We have supposed that at the design stage a maximum of $K = 5$ interim analyses were planned with equally spaced information levels. We use ρ-family error spending functions with index $\rho = 2$ to create efficacy and futility boundaries. Thus, type I error probability α and type II error probability β are "spent" in proportion to $(\mathcal{I}/\mathcal{I}_{\max})^2$. With these specifications we compute the target maximum information, $\mathcal{I}_{\max}$, which gives $a_K = b_K$ when the $\{a_k, b_k\}$ are calculated as described in Section 25.3. The computations depend on whether a binding or a non-binding futility (lower) boundary is to be employed. Table 25.2 displays the design parameters for both situations under

FIGURE 25.6
Expected number of events on termination of the group sequential log-rank test with a binding futility boundary and equally spaced information levels



the planning assumptions of equally spaced information levels, culminating in $\mathcal{I}_{\max}$. Note the familiar "curved triangular" shape of the boundaries as seen in Figure 25.5.

The maximum information $\mathcal{I}_{\max}$ needed in the group sequential trial design is 34.48 if a binding futility boundary is used or 35.58 if the futility boundary is non-binding. Under the approximation $\mathcal{I} \approx d/4$, the maximum numbers of failures that may need to be observed, $d_f = 4\,\mathcal{I}_f$, are 138 and 143, respectively.

A fixed sample study with no interim monitoring but the same type I error rate $\alpha = 0.025$ and power $1 - \beta = 0.8$ at $\theta = 0.5$ requires information

$$\mathcal{I}_f \;=\; \frac{\{\Phi^{-1}(0.975) + \Phi^{-1}(0.8)\}^2}{0.5^2} \;=\; 31.40.$$

Under the approximation $\mathcal{I} \approx d/4$, the total number of failures to be observed is $d_f = 4\,\mathcal{I}_f \approx 126$. Clearly this is smaller than the maximum event numbers of 138 or 143 for the five stage design. However the group sequential procedure benefits from the opportunity to stop before the last stage. Figure 25.6 shows

TABLE 25.3
Summary data for the oropharynx trial and critical values for the error spending design with binding futility boundary

Analysis k	Number entered	Number of deaths	$\mathcal{I}_k$	a_k	b_k	Z_k
1	83	27	5.43	−1.41	3.23	−1.04
2	126	58	12.58	−0.21	2.76	−1.00
3	174	91	21.11	0.78	2.44	−1.21
4	195	129	30.55	1.68	2.16	−0.73
5	195	142	33.28	2.14	2.14	−0.87

the expected number of events for the group sequential design with a binding futility boundary under different values of the hazard ratio. Plotted values for hazard ratios away from one (and log-hazard ratios away from zero) are less accurate since the approximation $\mathcal{I} \approx d/4$ is less reliable in these cases, particularly in later stages of the trial when numbers at risk on the two treatment arms become unequal. Figure 25.6 shows that the group sequential design with a binding futility boundary has an expected number of events under H_0 of 72.9 and under H_1 this becomes 94.5; the maximum expected number of events, which occurs for log hazard ratio $\theta = 0.35$, is 100.9, still considerably less than the 126 events for a fixed sample design. With a non-binding boundary the corresponding numbers are 74.3, 96.4 and 103.3, assuming for purposes of this calculation that the futility boundary is in fact obeyed.

We now turn to the task of applying the monitoring boundaries to the reconstructed data set summarized in Table 25.1. The boundary values $(a_1, b_1), \ldots, (a_5, b_5)$ are calculated using the observed information levels $\mathcal{I}_1, \ldots, \mathcal{I}_5$ rather than the equally spaced ones of the initial design. In doing this, we apply the formulae of Section 25.3 at each interim analysis in succession. From here on, we shall apply designs with binding futility boundaries, noting that the exposition would be very similar if we were to use non-binding futility boundaries instead.

The sequence of standardized log-rank statistics, $Z_1, \ldots, Z_5$, and the corresponding critical values $(a_1, b_1), \ldots, (a_5, b_5)$ are displayed in Table 25.3. We can see from this table that, had this design been used, the trial would have stopped for futility at analysis 2, about three years earlier than the original trial, reaching the same conclusion with only 126 subjects accrued instead of 195. Of course, those last three years may have produced further valuable information about other aspects of the treatments such as toxicity or

quality of life. With this in mind, investigators might have opted to continue the trial despite unpromising interim results. It is in anticipation of such eventualities that a non-binding futility boundary could be chosen since it allows a subsequent positive result for efficacy to be reported without concern that the type 1 error rate is inflated above the specified α.

25.6 Monitoring a hazard ratio with adjustment for strata and covariates

The Oropharynx Cancer data set contained information on a number of baseline covariates for each subject. These included gender, initial condition, T-staging, N-staging and two indicator variables describing the tumor site. Each patient was treated at one of six participating institutions and we shall treat institution as a stratifying variable. We model the data by means of a stratified proportional hazards regression model (Cox, 1972) in which the hazard rate for patient i is modeled as

$$h_{il}(t) \;=\; h_{0l}(t)\,\exp\{\beta_1 I(\text{Patient } i \text{ on Treatment B}) \,+\, \Sigma_{j=2}^{7}\, x_{ij}\beta_j\}.$$

The parameter β_1 represents the log hazard ratio between treatments after adjustment for the other covariates and stratification. We take the objective to be to test H_0: $\beta_1 \leq 0$ against the one-sided alternative $\beta_1 > 0$.

Standard software for Cox regression will provide the maximum partial likelihood estimate of the parameter vector, β, and its estimated variance matrix. We are interested in the treatment effect represented by the first component, β_1. At analysis k we have

$$\widehat{\beta}_1^{(k)}, \quad v_k = \widehat{\mathrm{Var}}\,(\widehat{\beta}_1^{(k)}), \quad \mathcal{I}_k = v_k^{-1} \quad \text{and} \quad Z_k = \widehat{\beta}_1^{(k)}/\sqrt{v_k}.$$

The standardized statistics $Z_1, \ldots, Z_5$ have, approximately, the canonical joint distribution of Section 25.2. Thus we may apply the group sequential designs and error spending method of Section 25.3 to monitor the adjusted log hazard ratio at successive interim analyses. In fact we can take exactly the same method that we described in Section 25.5 and simply use the above statistics Z_k and information values $\mathcal{I}_k$, $k = 1, \ldots, 5$, in place of those for the log-rank statistic.

Calculation gives the values $(a_1, b_1), \ldots, (a_5, b_5)$ shown in Table 25.4 for the error spending group sequential design, again with a binding futility boundary, to be applied to $Z_1, \ldots, Z_5$. Under this model and stopping rule, the study would — just — have stopped for futility at the second analysis.

TABLE 25.4
Covariate-adjusted group sequential analysis of the oropharynx data

Analysis k	$\mathcal{I}_k$	a_k	b_k	$\widehat{\beta}_1^{(k)}$	Z_k
1	4.11	−1.75	3.39	−0.79	−1.60
2	10.89	−0.44	2.85	−0.14	−0.45
3	19.23	0.59	2.50	−0.08	−0.33
4	28.10	1.45	2.24	0.04	0.20
5	30.96	2.23	2.23	0.01	0.04

25.7 Further work

In this chapter, we have concentrated on the use of an error spending group sequential design for monitoring a log-rank statistic or a regression coefficient in a Cox regression model. The methods we have presented form a good introduction to other group sequential methods for survival data. The ideas have been extended in two directions:

A. To other features of group sequential designs;

B. To other features of survival analysis.

A. *Further group sequential methods that can be applied to the collection and analysis of survival data.* We have considered the "curved triangular" testing boundaries that arise in one-sided hypothesis tests. These are commonly used in superiority trials where it is hoped to show that a new treatment improves on the current standard; the same forms of boundary also arise in non-inferiority trials where hypotheses H_0: $\theta \leq 0$ and H_1: $\theta > 0$ are replaced by H_0: $\theta \leq -\delta$ and H_1: $\theta > \delta$, where δ represents an acceptable "margin of inferiority". Other boundary shapes are applicable for testing a null hypothesis against a two-sided alternative or in tests of equivalence, where it is hoped to demonstrate that the effect of a new treatment is within a specified tolerance of that of an existing treatment.

In addition to the positive or negative outcome of a hypothesis test, it is usually required to give point or interval estimates of the treatment effect at the termination of a trial or to provide a P-value summarizing the strength of evidence against a null hypothesis. Special methods are needed to construct such quantities, taking into account the sequential nature of the design; see Jennison and Turnbull (2000, Ch. 8).

Repeated confidence intervals permit an interval estimate of a treatment effect to be stated at any stage of the trial (not just the last), with the property that the coverage probability of all the intervals is simultaneously controlled at a given confidence level, $1-\gamma$ say. Such confidence intervals are wider than naive, fixed sample size intervals computed at each stage, but they are free from the "multiple looks" bias of sequential testing. This obviates the problem of "over-interpretation of interim results"; see Jennison and Turnbull (1989).

B. *Further techniques for survival data to which group sequential methods can be applied.* First consider a one-sample problem, where we are interested in the time to an event such as death or the disease recurrence in a homogeneous population. Sometimes a binary outcome is defined to indicate whether failure has occurred after an elapsed time, τ say. If not all subjects are followed for time τ, the simple proportion of those surviving to time τ will be a biased estimate of the survival rate, while omitting subjects with potential censoring times less than τ is inefficient. These difficulties are overcome by use of the Kaplan-Meier estimate (Kaplan and Meier, 1958) of the survival function $S(t)$. Let $\widehat{S}_k(t)$ denote the Kaplan-Meier estimate of the survival probability $S(t)$ at time t based on data available at analysis k. For a given value of τ, suppose $0 < S(\tau) < 1$ and there is a positive probability for each observation to be uncensored and greater than τ, then Jennison and Turnbull (1985) show that the sequence

$$Z_k \;=\; \frac{\{\widehat{S}_k(\tau) - S(\tau)\}}{\sqrt{\mathrm{Var}\{\widehat{S}_k(\tau)\}}}, \quad k = 1, \ldots, K, \tag{25.6}$$

has, asymptotically, the canonical joint distribution of Section 25.2 with $\theta = S(\tau)$ and information levels $\mathcal{I}_k = [\mathrm{Var}\{\widehat{S}_k(\tau)\}]^{-1}$. A consistent estimate of the variance of $\widehat{S}_k(\tau)$ is provided by Greenwood's formula — see, for example, Jennison and Turnbull (1985). Hence, a group sequential test of the hypothesis H_0: $S(\tau) = p_0$, where τ and p_0 are specified, can be based on the standardized statistics

$$Z_k \;=\; \frac{\{\widehat{S}_k(\tau) - p_0\}}{\sqrt{\{\widehat{V}_k(\tau)\}}}, \quad k = 1, \ldots, K,$$

and associated information levels $\mathcal{I}_k = \{\widehat{V}_k(\tau)\}^{-1}$, where $\widehat{V}_k(\tau)$ denotes a consistent estimate of $\mathrm{Var}\{\widehat{S}_k(\tau)\}$. Since information depends on the number and times of observed failures, the error spending approach of Section 25.3 is needed for the construction of such tests. The Greenwood estimate is straightforward to calculate and is typically available in the output of standard statistical computer software for estimating survival curves. Alternatively, the "constrained" variance estimator introduced by Thomas and Grunkemeier (1975, Sec. 4) can be used in place of the Greenwood formula: simulations reported by Thomas and Grunkemeier and by Barber and Jennison (1999) show this should lead to more accurate attainment of error rates and coverage probabilities for repeated confidence intervals. Barber and Jennison (1999) go

on to propose further methods to achieve error rates and coverage probabilities more accurately in smaller sample sizes.

Sometimes, interest is in a certain *quantile* of the survival distribution. For $0 < p < 1$, we define the pth quantile of the survival distribution $S(t)$ to be $t_p = \inf\{t\colon\ S(t) \geq p\}$. Assuming $S(t)$ to be strictly decreasing in t, a group sequential test of H_0: $t_p = t^*$ for specified t^* and p is equivalent to a test of H_0: $S(t^*) = p$ and the same Kaplan-Meier test statistics can be used with $\tau = t^*$ and $p_0 = p$. Jennison and Turnbull (1985) have investigated repeated confidence intervals for the median survival time.

Analogous methods can also be used in a two-sample comparison. If $S_A(t)$ and $S_B(t)$ denote survival functions on treatments A and B in a randomized trial, a test of H_0: $S_A(\tau) = S_B(\tau)$, for a given choice of τ, can be based on successive statistics

$$Z_k \ = \ \frac{\{\widehat{S}_{Ak}(\tau) - \widehat{S}_{Bk}(\tau)\}}{\sqrt{\{\widetilde{V}_{Ak}(\tau) + \widetilde{V}_{Bk}(\tau)\}}}, \quad k = 1, \ldots, K,$$

where $\widehat{S}_{Ak}(\tau)$ and $\widehat{S}_{Bk}(\tau)$ are Kaplan-Meier estimates of $S_A(\tau)$ and $S_B(\tau)$, respectively, at analysis k and $\widetilde{V}_{Ak}(\tau)$ and $\widetilde{V}_{Bk}(\tau)$ are their estimated variances. The problem of comparing the pth quantiles of two survival distributions has been addressed by Keaney and Wei (1994).

25.8 Concluding Remarks

A variety of software packages is now available to implement the methods we have described. One choice that can compute the error spending boundaries described in Section 25.3 and that has a dedicated module for planning and analyzing survival trials is East (Cytel, 2012). Another choice is the **gsDesign** package in R.

It should be noted that not all sequences of standardized statistics follow the canonical joint distribution of Section 25.2. As an example, Slud and Wei (1982) have shown that this property does *not* hold for some weighted log-rank test statistics when there is staggered entry. These statistics include those arising in Gehan's (1965) procedure for modifying the Wilcoxon test to allow censored data.

This chapter has provided a basic overview of the use of group sequential methods for survival data. There is a large literature on the subject which we have not attempted to summarize here: some more references can be found in Jennison and Turnbull (2000, Ch. 13). In particular, there is an emerging literature on the adaptive clinical trial designs for survival data. The availability at interim analyses of partial information about patients' continuing survival causes particular problems in adaptive designs: for one

example with correlated survival endpoints, and a solution to the adaptive design problem, see Jenkins, Stone and Jennison (2011).

References

Barber, S. and Jennison, C. 1999. Symmetric tests and confidence intervals for survival probabilities and quantiles of censored survival data. *Biometrics*, 55:430–436.

Barber, S. and Jennison, C. 2002. Optimal asymmetric one-sided group sequential tests, *Biometrika*, 89:49–60.

Cox, D.R. 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220.

Cytel Software Corporation, 2012. *EaSt v.6: A software package for the design and interim monitoring of group-sequential clinical trials*, Cytel Software Corporation, Cambridge, Massachusetts.

Gehan, E.A. 1965. A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, 52:203–223.

Jenkins, M., Stone, A. and Jennison, C. 2011. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics*, 10:347–356.

Jennison, C. and Turnbull, B.W. 1985. Repeated confidence intervals for the median survival time. *Biometrika*, 72:619–625.

Jennison, C. and Turnbull, B.W. 1989. Interim analyses: the repeated confidence interval approach (with discussion). *Journal of the Royal Statistical Society, Series B*, 51:305–361.

Jennison, C. and Turnbull, B.W. 1997. Group sequential analysis incorporating covariate information. *Journal of the American Statistical Association*, 92:1330–1341.

Jennison, C. and Turnbull, B.W. 2000. *Group Sequential Methods with Applications to Clinical Trials*, Chapman & Hall/CRC, Boca Raton.

Kalbfleisch, J.D. and Prentice, R.L. 2002. *The Statistical Analysis of Failure Time Data, Second Edition*, Wiley, New York.

Kaplan, E.L. and Meier, P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.

Keaney, K.M. and Wei, L.J. 1994. Interim analyses based on median survival times. *Biometrika*, 81:279–286.

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

Pampallona, S. and Tsiatis, A.A. 1994. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference*, 42:19–35.

Scharfstein, D.O., Tsiatis, A.A. and Robins, J.M. 1997. Semiparametric

efficiency and its implication on the design and analysis of group sequential studies. *Journal of the American Statistical Association*, 92:1342–1350.

Slud, E.V. and Wei, L-J. 1982. Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association*, 77:862–868.

Thomas, D.R. and Grunkemeier, G.L. 1975. Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, 70:865–871.