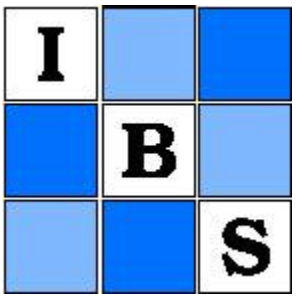

BIOMETRICS 50, 25–38
March 1994

Analysis of Factorial Survival Experiments

Eric V. Slud

Mathematics Department, University of Maryland,
College Park, Maryland 20742, U.S.A.
and
Information Management Services, Inc.,
Rockville, Maryland, U.S.A.

SUMMARY

Several new methodological issues that arise within two-way factorial designs for survival experiments are discussed within the framework of asymptotic theory for the proportional hazards model with two binary treatment covariates. These issues include: the proper formulation of null hypotheses and alternatives, the choice among log-rank and adjusted or stratified log-rank statistics, the asymptotic correlation between test statistics for the separate main effects, the asymptotic power (under the various possible methods of analysis) of tests to detect main effects and interactions, the comparison of power to detect main effects within a 2×2 factorial design with power in a three-group trial where no patients are randomized simultaneously to both treatments, and the problems of analysis arising when accrual or exposure to one of the treatments is terminated early for ethical reasons.

1. Introduction

There are presently a number of ongoing or recently completed clinical trials with survival endpoints and a factorial design. However, as remarked by Byar and Piantadosi (1985), there has been virtually no methodological work directed specifically at factorial designs for survival analysis.

The purpose of this paper is to use existing theoretical tools to analyze clinical trials with survival endpoints and (2×2) factorial designs, within the framework of proportional hazards models with two binary treatment-group indicators as covariates. Some of the new methodological issues that arise in such trials are: (i) choosing among log-rank, adjusted log-rank, and stratified log-rank statistics for detecting the separate main effects, and planning for adequate power to detect desired main effects and interactions; (ii) assessment of the degree of dependence among test statistics for separate main effects, both in ideal settings and in situations likely to be realized in practice, with consequences for the "experimentwise" reporting of significant results; (iii) relative efficiency of factorial trials as compared with trials designed to detect only a single main effect; (iv) justification of the statistical validity of analyses of factorial trials after early termination of accrual or exposure to one of the treatments (with the terminated treatment either eliminated from or applied to all newly accrued patients).

The techniques for large-sample asymptotic analysis that we exploit in the paper rely on counting processes and martingales, as expounded in the books of Gill (1980) or Harrington and Fleming (1990). Comparison of power and efficiency of test statistics is accomplished by consideration of so-called contiguous alternatives, which approach the null hypotheses of interest as sample-size increases. Because of the practical importance of (iv), these techniques are developed also in the context of group sequential or repeated significance tests.

The plan of the paper is as follows: models and statistics are introduced in Section 2; asymptotic large-sample theoretical results are summarized under various headings in Section 3; in Section 4,

Key words: Adjusted score statistic; Counting processes; Early termination of a treatment; Experimentwise probability of stopping; Factorial design; Proportional hazards model; Randomized clinical survival experiment; Repeated significance tests; Stochastic integrals; Stopping boundary; Stratified log-rank statistics.

these results are interpreted and recommendations are based on them. Justification of the theoretical results is deferred to the Appendix.

2. Cox-Model Formulation of Factorial Designs

For convenience of notation, we formulate our models and statistics for the case of a two-way factorial design in which subjects are randomly assigned to treatments. The assumptions and notations are adapted from Cox (1972), Andersen and Gill (1982), and Slud (1984). Suppose that for each of n independent subjects indexed by i, there is a random vector $Z_i = (Z_i^{(1)}, Z_i^{(2)}, Z_i^{(3)})$ whose components respectively denote the ith subject's level j of treatment-factor A, level k of factor B, and the product $j \cdot k$. Suppose also that for each subject there is a pair of latent waiting times X_i, C_i, which are conditionally independent given Z_i, with X_i denoting the ith subject's time from entry until the study endpoint of interest, and C_i the time from entry until loss to follow-up. Denote by ρ_{jk} the fraction $\Pr(Z_i = (j, k, jk)')$ of the study population assigned to A-treatment level j and B-treatment level k. The entry time ε_i for the ith subject is assumed to be independent of all other variables, and is taken to be 0 until we discuss repeated or group sequential tests. The data observed on the ith subject up to study time t consist of $T_i(t) = \min\{X_i, C_i, t - \varepsilon_i\}$ along with the failure-indicator $\Delta_i(t) = I_{[\min(C_i, t-\varepsilon_i) \geq X_i]}$. As usual in survival analysis, denote by $T_i = T_i(\infty)$ the "event time" $\min\{X_i, C_i\}$ for the ith subject, and $\Delta_i = \Delta_i(\infty)$ the corresponding "death indicator" $I_{[X_i \leq C_i]}$.

We assume that the dependence of survival on factorial levels Z is described by a Cox (1972) proportional hazards model. That is, the conditional cumulative hazard function for the latent failure variable X_i given Z_i is assumed to have the form

$$\Lambda(t|Z_i = (j, k, jk)') = \Lambda_0(t)\exp(\beta \cdot Z_i) = \Lambda_0(t)e^{j\beta_1 + k\beta_2 + jk\beta_3}, \tag{1}$$

where $\Lambda_0(t)$ is the baseline cumulative hazard and $\beta = (\beta_1, \beta_2, \beta_3)$ is a parameter vector. The separate effects β_1 for factor A and β_2 for factor B and the interaction β_3 might be parameterized differently in the case of multilevel treatments, and (known) time-dependent functions might be included in the covariates Z to model delays in treatment effect as in Zucker and Lakatos (1990), but we treat only model (1) in the 2×2 case, fixing the possible values of j and k as 0, 1. A further model assumption, which we shall impose in much of what follows and which would hold, for example, in a randomized trial if the only loss to follow-up were due to end-of-study censoring, is:

> The loss-to-follow-up times C_i are independent of the factor-level vectors Z_i. The fractions ρ_{jk} of the study population for which $Z_i = (j, k, jk)'$ factor as products $a_j b_k$. (2)

Ordinarily the factorial study would be conducted to find out whether either treatment has a significant effect on survival. In a testing framework, the null hypothesis H_0 would be formulated as $\beta = (\beta_1, \beta_2, \beta_3) = \mathbf{0}$, and this is the only null hypothesis that makes sense in a discussion of dependence between test statistics to detect the separate A and B main effects. However, if only the single main effect for A were of interest, either for purposes of comparison with a nonfactorial trial or because treatment B has been terminated early for ethical reasons, then the null hypothesis would be $H_{0,A}$: $\beta_1 = \beta_3 = 0$, i.e., that X_i has the same conditional distribution given $Z_i = (j, k, jk)'$ for each level j of A, for all B-exposure levels k. Another scenario is that treatment B might be strongly expected to affect survival whereas the main A-effect and any possible interaction effects are less certain. In this setting also, $H_{0,A}$ is an appropriate null hypothesis.

For testing the null hypothesis $\beta = 0$ against alternatives with at most one nonzero coefficient β_i, $i = 1, 2, 3$, the Cox-model score statistics for these three alternatives are the components of the vector score statistic $\mathbf{S} = (S^{(1)}, S^{(2)}, S^{(3)})$ defined by

$$\mathbf{S} = (\mathrm{diag}(\hat{D}_{11}, \hat{D}_{22}, \hat{D}_{33}))^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Delta_i\{Z_i - \bar{Z}_i\},$$
$$\hat{D} = (\hat{D}_{pq}) = \frac{1}{n} \sum_{i=1}^{n} \Delta_i\{Z_i - \bar{Z}_i\}\{Z_i - \bar{Z}_i\}', \tag{3}$$

where

$$\bar{Z}_i = \sum_{m=1}^{n} I_{[T_m \geq T_i]} Z_m \Big/ \sum_{m=1}^{n} I_{[T_m \geq T_i]}.$$

However, it is known that using the log-rank statistic $S^{(1)}$ to test for $\beta_1 \neq 0$ can result in serious loss of power for alternatives with $\beta_2 \neq 0$, as compared with the most powerful test statistic based on the model (1). (See §3 for detailed power comparisons.) The adjusted test statistics $S^{*(i)}$ asymptotically equivalent to the likelihood ratios for the respective alternatives $\beta_i \neq 0$, $i = 1, 2, 3$, are defined through

$$\bar{Z}_i(\beta) = \sum_{m=1}^{n} I_{[T_m \geq T_i]} Z_m e^{\beta Z_m} \Big/ \sum_{m=1}^{n} I_{[T_m \geq T_i]} e^{\beta Z_m},$$

with $\bar{Z}_i$ replaced in the formulas (3) by $\bar{Z}_i(0, \hat{\beta}_2, \hat{\beta}_3)$ for defining $S^{*(1)}$, by $\bar{Z}_i(\hat{\beta}_1, 0, \hat{\beta}_3)$ for defining $S^{*(2)}$, and by $\bar{Z}_i(\hat{\beta}_1, \hat{\beta}_2, 0)$ for defining $S^{*(3)}$, where the maximum partial likelihood estimators (MPLEs) $\hat{\beta}_i$, $i = 1, 2, 3$, are defined by the equation

$$\sum_{i=1}^{n} \Delta_i \{Z_i - \bar{Z}_i(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)\} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{4}$$

If we are testing for the main A-effect but think it likely that B has a non-null effect, then another reasonable choice of test statistic is the stratified log-rank S_A^{str} defined by

$$S_A^{\text{str}} \equiv \frac{\sum_{i=1}^{n} \Delta_i \sum_{k=0}^{1} \left\{ I_{[Z_i = (1, k, k)]} - \dfrac{Y_{1k}(X_i)}{Y_{1k}(X_i) + Y_{0k}(X_i)} I_{[Z_i^{(2)} = k]} \right\}}{\sqrt{\sum_{i=1}^{n} \Delta_i \sum_{k=0}^{1} \dfrac{Y_{1k}(X_i) Y_{0k}(X_i)}{Y_{1k}(X_i) + Y_{0k}(X_i)} I_{[Z_i^{(2)} = k]}}}, \tag{5}$$

where $Y_{jk}(t) \equiv \Sigma_i I_{[T_i \geq t, Z_i = (j, k, jk)]}$. The stratified log-rank statistic S_B^{str} for main B-effect with non-null A-effect is defined analogously.

For purposes of later comparison, we consider also a linear model for survival times in the uncensored case, namely the accelerated failure model, which expresses within factor levels (j, k)

$$\ln(X) = \beta_0 + \beta_1 j + \beta_2 k + \beta_3 jk + V, \tag{6}$$

where β_0 is a constant and the random variable $\exp(\beta_0 + V)$ has cumulative hazard function $\Lambda_0(t)$ with $\mathrm{E}(V|Z = (j, k, jk)') = 0$. This model coincides with (1) in the Weibull cases where $\Lambda_0(t) = \gamma t^{\delta}$ for some $\gamma, \delta > 0$. Within such a model, assuming also (2) for convenience, the main-effect parameter for treatment A is $\mathrm{E}(\ln(X)|Z^{(1)} = 1) - \mathrm{E}(\ln(X)|Z^{(1)} = 0) = \beta_1 + b_1\beta_2$, with b_1 as in (2), and the interaction effect is β_3. The parameters β can be estimated by ordinary least squares with respect to a design matrix with rows $(1, Z_i')$, $i = 1, \ldots, n$, and the least squares estimator for the parameters $(\beta_1, \beta_2, \beta_3)$ has covariance matrix proportional to the inverse of $\mathrm{cov}(Z)$.

Motivated by the linear-model definition of main-effect and interaction parameters, one might also analyze the factorial trial data under (1)–(2) by the MPLEs $\hat{\beta}_1 + b_1\hat{\beta}_3$ and $\hat{\beta}_3$, and these statistics too will be considered below.

3. Summary of Results

3.1 *Asymptotic Distributions of Statistics*

In a moderate-to-large-sample clinical trial satisfying our assumptions including (1), the MPLE $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ is asymptotically normally distributed with mean β and variance–covariance matrix D^{-1}, where

$$D \equiv D(\infty, \beta) \equiv \int_0^{\infty} \left[\frac{u^{(2)}(s, \beta)}{u^{(0)}(s, \beta)} - \left(\frac{u^{(1)}(s, \beta)}{u^{(0)}(s, \beta)} \right)^{\otimes 2} \right] u^{(0)}(s, \beta)\, d\Lambda_0(s) \tag{7}$$

and for $i = 0, 1, 2$,

$$u^{(i)}(s, \beta) \equiv \sum_{j,k} \begin{pmatrix} j \\ k \\ jk \end{pmatrix}^{\otimes i} \rho_{jk} \exp\left\{ \begin{pmatrix} j \\ k \\ jk \end{pmatrix}' \beta - e^{(j,k,jk)\beta} \Lambda_0(s) \right\} \Pr\left(C \geq s | Z = \begin{pmatrix} j \\ k \\ jk \end{pmatrix} \right).$$

Here and in what follows, for a vector v we denote by $v^{\otimes i}$ respectively 1, v, and vv' for $i = 0, 1, 2$. For large samples, the vector $\bar{Z}_i(\beta)$ defined above (4) is asymptotically close to $u^{(1)}(T_i, \beta)/u^{(0)}(T_i, \beta)$. Expression (7) is consistently estimated by the covariance estimator

$$\hat{D}(\infty, \hat{\beta}) \equiv \frac{1}{n} \sum_{i=1}^{n} \Delta_i \{Z_i - \bar{Z}_i(\hat{\beta})\}^{\otimes 2}. \tag{8}$$

Note that formula (8) with $\hat{\beta}$ replaced by $\mathbf{0}$ coincides with the expression $\hat{D}$ in (3), and under H_0 both (3) and (8) are consistent for D in (7). Under $H_{0,A}$, $\hat{D}_{11}$ turns out to be consistent for D_{11}.

It is shown in the Appendix that under H_0 both

$$\mathbf{S} - \sqrt{n}(\text{diag}(D_{11}, D_{22}, D_{33}))^{-1/2} D\hat{\beta}, \quad \mathbf{S}^* - \sqrt{n}\ \text{diag}(D_{11}^{-1}, D_{22}^{-1}, D_{33}^{-1})^{-1/2}\hat{\beta} \tag{9}$$

converge in probability to 0. More generally, under $H_{0,A}$ the first and third components of the second quantity in (9) converge in probability to 0. Thus the "adjusted log-rank statistic" $S^{*(1)}$ for treatment-effect of A is after standardization asymptotically equivalent to the MPLE $\hat{\beta}_1$ under $H_{0,A}$. A further equivalence holds: under H_0 but not under $H_{0,A}$, $S^{(1)} - S_A^{\text{str}} \xrightarrow{P} 0$ as $n \to \infty$.

3.2 *Asymptotic Correlation Among Test Statistics*

When a clinical trial with factorial design is analyzed, inferences for both main treatment effects and for an interaction effect will be performed and simultaneously reported, each test statistic being referred separately to its own nominal distribution. We investigate the dependence among statistics that might be used to test separate treatment effects, in order to assess the "experimentwise" probability of reporting at least one significant effect, under both null and alternative hypotheses.

Formula (7) indicates that under H_0, $D = \text{cov}(Z)\delta$, where $\delta = \Pr(\Delta_1 = 1) = \Pr(X_1 \le C_1)$. When (2) also holds, the first and second components of Z_i are readily seen for each t to be conditionally independent given that subject i is alive and uncensored at study time t (i.e., given that $T_i \ge t$), so that

$$D = \begin{pmatrix} a_1(1-a_1) & 0 & a_1(1-a_1)b_1 \\ 0 & b_1(1-b_1) & b_1(1-b_1)a_1 \\ a_1(1-a_1)b_1 & b_1(1-b_1)a_1 & a_1b_1(1-a_1b_1) \end{pmatrix}\delta.$$

Since D is the asymptotic covariance matrix of $\sqrt{n}D\hat{\beta}$, it follows that the two log-rank statistics $S^{(1)}$ and $S^{(2)}$ are asymptotically independent under H_0, (1), and (2). If (2) fails to hold, for example, because different loss-to-follow-up patterns obtain within different treatment-factor levels, then under H_0 and (1), $S^{(1)}$ and $S^{(2)}$ will asymptotically no longer be precisely independent, but will be very nearly so in practice. For a quantitative assessment of the dependence and its effect on experimentwise rejection probabilities in group sequential testing, see Section 3.4 below.

The adjusted log-rank statistics $S^{*(1)}$ and $S^{*(2)}$ will never be asymptotically independent: using their equivalence to the standardized MPLEs $\hat{\beta}_1$ and $\hat{\beta}_2$, together with the theory summarized in Section 3.1, we find the asymptotic correlation of $S^{*(1)}$ and $S^{*(2)}$ under H_0 and (2) to be $(D^{-1})_{12}/\sqrt{(D^{-1})_{11}(D^{-1})_{22}} = \sqrt{a_1b_1}$.

Under H_0 and contiguous alternatives, without assuming (2), $S^{(2)}$ will be asymptotically independent of $(S^{*(1)}, S^{*(3)})$. This follows immediately upon taking asymptotic covariances using the representations given in paragraph 3.1:

$$S^{(2)} \approx \sqrt{n/D_{22}} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}' D\hat{\beta}, \quad \begin{pmatrix} S^{*(1)} \\ S^{*(3)} \end{pmatrix} \approx \sqrt{n} \begin{pmatrix} (D^{-1})_{11}^{-1/2} & 0 & 0 \\ 0 & 0 & (D^{-1})_{33}^{-1/2} \end{pmatrix}\hat{\beta}.$$

Thus, suppose that we want to test H_0, but that we believe β_2 is quite likely to be different from 0, i.e., that there may well be a non-null (but not extremely large) main effect from treatment B. Even if we cannot justifiably assume (2), we can first test for the main B-effect using the statistic $S^{(2)}$ at significance level α_B, and then perform an independent test of $H_{0,A}$ using $(S^{*(1)}, S^{*(3)})$. Independence of the two tests makes them relatively easy to interpret, whether from the nominal or experimentwise point of view. In addition, the use of adjusted statistics at the second stage of testing provides a valid test of $H_{0,A}$ regardless of whether $\beta_2 = 0$.

3.3 *Power and Sample Size: Comparison Among Statistics*

All of the foregoing statements about asymptotic equivalence and independence of statistics under null hypotheses extend automatically to contiguous alternatives, under which asymptotic formulas for power will now be developed, following the methods of Gill (1980) and Schoenfeld (1981). The "contiguous" alternatives that we consider are close to H_0 or $H_{0,A}$ in a way that depends on n so as to yield a limiting power strictly between 0 and 1 as $n \to \infty$, viz.

$H_{1,n}$: Z_i and C_i are as under $H_{0,A}$, and conditionally given $Z_i = (j, k, jk)'$, X_i is independent of C_i and has cumulative hazard $\int_0^t \exp(k\beta_2 + jn^{-1/2}(c_1(s) + kc_3(s)))\, d\Lambda_0(s)$,

where c_1 and c_3 are fixed functions that we take to be constants. These constants express the relative degrees of deviation from 0 of the coefficients of $Z^{(1)}$ and $Z^{(3)}$ under the alternative. For convenience, we assume (2). For each of several normalized test statistics, with standard normal asymptotic distribution under $H_{0,A}$, we calculate in the Appendix the limiting expectation μ under $H_{1,n}$ as $n \to \infty$, leading to the expression $\Phi(-z_{\alpha/2} + \mu) + \Phi(-z_{\alpha/2} - \mu)$ for the limiting power of a two-sided size-α test based on that statistic. Formulas for the means μ in terms of the constants c_1, c_3, and $\gamma \equiv \exp(\beta_2)$, and of the matrix D given by (7), are displayed in Table 1. The results are conditional on the observed values of a_1 and b_1.

Table 1

Formulas for asymptotic means of statistics under $H_{1,n}$, in terms of constants c_1, c_3, $a = a_1$, $b = b_1$, D, γ

Statistic	Asymptotic mean
Log-rank $S^{(1)}$	$\dfrac{1}{\sqrt{D_{11}}}(c_1D_{11} + c_3D_{13} - (\gamma - 1)((\gamma - 1)c_1 + \gamma c_3)A)$
Adjusted log-rank $S^{*(1)}$	$\sqrt{D_{11} - D_{13}}\, c_1$
Stratified log-rank S_A^{str}	$(c_1D_{11} + c_3D_{13})/\sqrt{D_{11}}$
Standardized MPLE for main effect $\beta_1 + b\beta_3$	$\dfrac{\sqrt{D_{11} - D_{13}}\,(c_1 + bc_3)}{\sqrt{1 - 2b + b^2D_{11}/D_{13}}}$
Standardized MPLE for β_1 based on data without A & B treatment	$\sqrt{D_{11} - D_{13}}\, c_1$

In the expression for the log-rank in this table, the constant A is defined in terms of $S_C(t) \equiv \Pr(C \geq t)$ by

$$A = a_1(1 - a_1)b_1(1 - b_1) \int S_C(\Lambda_0^{-1}(x)) \frac{xe^{-(1+\gamma)x}\, dx}{be^{-\gamma x} + (1 - b)e^{-x}}.$$

The entries in Table 1, in conjunction with formula (7) for D, can be used generally for the calculation of power or sample size in factorial trials. The asymptotic mean under $H_{1,n}$ for the adjusted log-rank statistic $S^{*(3)}$ for testing interaction is $c_3\sqrt{D_{13} - D_{13}^2/D_{11}}$.

To facilitate comparisons between statistics and the interpretations and recommendations of Section 4, we consider also the simplifications of Table 1 *either* if the alternatives $H_{1,n}$ are contiguous to H_0 (i.e., if $\gamma = 1$) *or* if there is no censoring (i.e., $S_C(t) \equiv 1$). In these cases, if $\delta \equiv \int S_C(\Lambda_0^{-1}(x))e^{-x}\, dx$, then as $n \to \infty$, under $H_{1,n}$

$$\text{Asympt. mean } (S^{*(1)}) = c_1\sqrt{a_1(1 - a_1)(1 - b_1)\delta},$$

$$\text{Asympt. mean (standardized } \hat{\beta}_1 + b_1\hat{\beta}_3) = \text{Asympt. mean } (S_A^{\text{str}}) = (c_1 + b_1c_3)\sqrt{a_1(1 - a_1)\delta}. \tag{10}$$

As in Table 1, the asymptotic mean for the standardized MPLE of β_1 based only on the three levels (0, 0), (1, 0), and (0, 1) for $(Z^{(1)}, Z^{(2)})$ is identical to the asymptotic mean of $S^{*(1)}$. We have remarked already that under alternatives $H_{1,n}$ contiguous to H_0 (when $\gamma = 1$), the log-rank and stratified log-rank are asymptotically the same. On the other hand, for general γ in the absence of censoring,

Table 2 exhibits, for the case $b_1 = .5$, the numerically computed coefficients of c_1 and c_3 in the asymptotic mean for the log-rank statistic.

See Section 4 for discussion and interpretation of the asymptotic means of test statistics presented here and their implications for asymptotic power within factorial survival analyses.

Table 2
Coefficients of c_1 and c_3 under $H_{1,n}$ in Asympt. mean (log-rank)/$(a_1(1 - a_1)^{1/2})$, for $S_C \equiv 1$, $b_1 = .5$

$\gamma = e^{\beta_2}$	Coeff. of c_1	Coeff. of c_3
3	.832	.248
2	.911	.322
1.5	.964	.391
1.2	.992	.452
1	1	.5
.8	.988	.548
.7	.971	.567
.5	.911	.589
.3	.813	.580

3.4 *Extensions to Repeated Significance Tests*

Both to allow monitoring of a factorial clinical trial in a group sequential setting, and to address possible early termination of one of the two treatments, we consider the properties of test statistics calculated repeatedly over chronological time. We introduce additional subscripts τ on variables T_m, Δ_m, $\bar{Z}_m(\beta)$, $\hat{\beta}_i$, $S^{*(i)}$, $S^{(i)}$ to indicate wherever necessary that these quantities are to be calculated using only data that would be observable as of time τ. The effect of the subscript τ on any of the statistics is that observations on subjects for whom T_i was larger than $\tau - \varepsilon_i$ are treated as right-censored at time-on-test $\tau - \varepsilon_i$ in real-time-τ interim analysis. That is, for a time-τ analysis, censorship time C_i is replaced by $\min(\tau - \varepsilon_i, C_i)$.

Using results of Slud (1984)—but see the Appendix of Jennison and Turnbull (1985) for a simpler discussion—it can be shown that the numerators $n^{-1/2} \Sigma_m \Delta_{m,\tau}\{Z_m^{(i)} - \bar{Z}_{m,\tau}^{(i)}(\mathbf{0})\}$ of the log-rank processes $S_\tau^{(i)}$ repeatedly computed at successive times τ have uncorrelated and therefore asymptotically independent increments in τ under H_0 (Slud, 1984, Cor. 2.4). Moreover, $S_\tau^{*(1)}$ and $S_\tau^{(2)}$ are asymptotically independent Gaussian stochastic processes indexed by real time τ, and if (2) holds then $S_\tau^{(1)}$ and $S_\tau^{(2)}$ are, too. Thus, a repeated test based on $S_\tau^{(2)}$ for a number of values of τ of the hypothesis H_0, versus alternatives with $\beta_2 \neq 0$, is approximately independent of a repeated significance test of $H_{0,A}$ based on statistics $S_\tau^{*(1)}$.

To develop a quantitative feeling for the magnitudes of typical correlations among the component main-effect log-rank test statistics when (2) fails to hold, consider the following example. Imagine a clinical experiment with two treatment-factors of two levels each, with staggered patient entry and five times $\tau_1, \ldots, \tau_5$ at which log-rank tests will be performed. Suppose that the increments D_{11,τ_1}, $D_{11,\tau_2} - D_{11,\tau_1}, \ldots, D_{11,\tau_5} - D_{11,\tau_4}$ of asymptotic variance are all anticipated to be equal, as are $D_{22,\tau_1}, D_{22,\tau_2} - D_{22,\tau_1}, \ldots, D_{22,\tau_5} - D_{22,\tau_4}$. Let the repeated log-rank significance tests for each of the two separate treatment-effects on survival be designed as two-sided tests with the Armitage–Pocock stopping boundary (with constant nominal cutoff 2.414) corresponding to overall size .05 for each test. The repeated significance test for effects of treatment A (respectively, B) on survival stops and rejects H_0 at any time τ_i, $i = 1, \ldots, 5$, for which $|S_{\tau_i}^{(1)}| > 2.414$ (respectively, $|S_{\tau_i}^{(2)}| > 2.414$), and the experimentwise significance level is $.05 + .05 - (.05)^2 = .0975$ if (2) holds. If the two log-rank processes were perfectly correlated, then the experimentwise significance level would be .05. Suppose that all survival and loss-to-follow-up times for subjects within each stratum are exponentially distributed random variables, and that

$$\mathrm{E}(X|Z = (j, k, jk)) = 1, \text{ all } j, k, \quad \mathrm{E}(C|Z = (j, k, jk)) = \begin{cases} 3 & \text{if } (j + k) \text{ is even,} \\ 1.5 & \text{if } (j + k) \text{ is odd.} \end{cases}$$

This satisfies the null hypothesis H_0 but not (2). Assume the treatment strata have equal size, i.e., $\rho_{jk} = .25$ for all j, k, and simultaneous entry times $\varepsilon_i = 0$. Then

$$D_{11,\tau} = D_{22,\tau} = \frac{1}{8}\left\{\frac{27}{20} - \frac{3}{4}e^{-4\tau/3} - \frac{3}{5}e^{-5\tau/3}\right\},$$

$$D_{12,\tau} = \frac{1}{8}\left\{\frac{3}{20} - \frac{3}{4}e^{-4\tau/3} + \frac{3}{5}e^{-5\tau/3}\right\}.$$

The times $\tau_1, \tau_2, \tau_3, \tau_4, \tau_5$ at which the values of the asymptotic variance $D_{11,\tau}/D_{11,\infty}$ are equal to .2, .4, .6, .8, and 1 are respectively .1508, .3459, .6220, 1.0973, and ∞. Let (ξ_i, η_i) for $i = 1, \ldots, 5$ denote the increments from times 0 to τ_1, τ_1 to τ_2, etc., for the log-rank process $(S_\tau^{(1)}\sqrt{D_{11,\tau}/D_{11,\infty}}, S_\tau^{(2)}\sqrt{D_{22,\tau}/D_{22,\infty}})$. Then ξ_i, η_i are jointly Gaussian variables, each with variance .2; each ξ_i is uncorrelated with all the variables $\{\xi_j, \eta_j: j \neq i\}$; and

$$(\mathrm{corr}(\xi_i, \eta_1))_{i=1}^5 = (.0121, .0406, .0789, .1377, .2862).$$

Using these correlations, which we observe to be rather small, one calculates the following stopping probabilities for the repeated two-component log-rank significance test under H_0:

k	t_k	*Probability of stopping at* t_k *with 2-component test*	*Probability of stopping at* t_k *with 1-component test*
1	.1508	.033	.01579
2	.3459	.021	.01170
3	.6220	.015	.00901
4	1.097	.015	.00730
5	∞	.012	.00614

The stopping probabilities for the two-component test were obtained by a Monte Carlo simulation with 10,000 replications; the probabilities for a single log-rank process were numerically integrated. Compare the cumulative empirical probabilities .033, .054, .069, .084, .096 of stopping and rejecting using the two-component test, with the corresponding probabilities .031, .054, .071, .086, .0975 calculated from the single-component stopping probabilities as though the two components were independent. Calculations of power yielded similar results for the log-rank tests against alternatives with $n = 144$ and with β_i for $i = 1, 2, 3$ as large as ln(2); i.e., the errors were at most .01 in calculating experimentwise power from componentwise power as though the two component tests were independent.

The primary use of approximate independence of log-rank statistics for separate main effects is as follows. Suppose that a trial is designed so that the experimentwise size and power are to be α_A and $1 - \beta_A$ (respectively α_B and $1 - \beta_B$) for testing H_0 versus the alternative that exposure to A (respectively, to B) multiplies the hazard by ρ_A (resp. by ρ_B), i.e., that $\beta_1 = \ln(\rho_A)$ ($\beta_2 = \ln(\rho_B)$). If the test statistics are $S^{(1)}$ and $S^{(2)}$ and if ρ_A and ρ_B are at most 2, with α_A and α_B equal to .05 and β_A and β_B in the range .80–.90, then simulation studies by Lininger et al. (1979) and Gail, DeMets, and Slud (1982) justify the use of asymptotic theory as it relates to power and approximate independence of increments of log-rank numerators when 100 or more observed deaths are expected. The null hypothesis is rejected whenever either of the two separate log-rank tests reject at nominal sizes α_A and α_B. By approximate independence of the log-rank processes for A and B, the experimentwise significance level is $\alpha_A + \alpha_B - \alpha_A\alpha_B$. Similarly, the rejection probability against the alternative where A multiplies hazards by ρ_A while B does not affect hazards, is approximately $1 - \beta_A + \beta_A\alpha_B$.

3.5 *Early Termination of One Treatment*

One of the novel features of a factorial survival experiment is that one treatment-factor might for ethical reasons be eliminated or applied to all study participants before the study has ended. Assume here that accrual to the level $k = 1$ of treatment factor B is eliminated from some time on, and that we can exclude the need to model crossover effects, which although extremely important are beyond the scope of this paper. Thus we assume either that treatment B has been judged ineffective but toxic—a judgment based on the behavior of a statistic such as $S_\tau^{(2)}$ up to some time τ_1, resulting in switching patients previously treated with B to an untreated status—or that for some similar reason, all patients accrued past time τ_1 receive level 0 of treatment B. In both cases, we are assuming that model (1) continues to hold for all patients, if only because $\beta = 0$. The joint behavior over time of the test statistics must then be invoked to justify significance tests for the continuing treatment A under study.

Either by assuming condition (2), or by appealing to Section 3.4, we treat the main-effect log-rank processes $S_\tau^{(1)}$ and $S_\tau^{(2)}$ as approximately independent. Assume also that early termination is based only on double-blind determinations of toxicity and possibly on values of the $S_\tau^{(2)}$ for times τ up to τ_1. The independence of the stochastic process $S_\tau^{(2)}$ from $S_\tau^{*(1)}$, with or without (2), or from $S_\tau^{(1)}$ if (2)

holds, implies that the size and power of any fixed-sample or repeated significance test of $H_{0,A}$, respectively using $S_\tau^{*(1)}$ or $S_\tau^{(1)}$, is approximately unaffected by the early termination of the B-arm of the study. It would be appropriate to base tests on $S_\tau^{*(1)}$ whenever treatment B is not terminated early specifically for lack of effect.

4. Discussion and Recommendations

The reason usually advanced for performing a factorial experiment, such as a 2×2 factorial randomized clinical trial, is the economy of performing two experiments—one trial for each of the two treatments—for the price of one. Apart from the logistical difficulty in securing patient compliance with two different treatment regimens, an objection sometimes raised to designing trials factorially is that when the interaction between treatments is opposite to the individual treatment effects, a factorial design and analysis may lead to nonsignificant results, especially if the trial has been designed to achieve .05 *experimentwise* Type I error rate, whereas a trial for each of the treatment effects might have shown a statistically significant result. Our first task in this section is to shed light on these questions from the large-sample viewpoint of Sections 3.1–3.3. Assume (2).

We first compare, via asymptotic calculations under $H_{1,n}$ as in Section 3.3 for one-sided tests of level α, the power of a 2×2 factorial trial to detect treatment-A effect with the power in an ordinary two-group trial (with population fraction a_1 in the treated group) or in a "three-group design" in which patients are never randomized to simultaneous $A\&B$ treatments. If exposure to B does not occur naturally in the study population, then β_1 but not β_3 can be inferred from the two- or three-group study. The log-rank in the two-group case has power

$$\Phi(-z_\alpha + c_1\sqrt{a_1(1-a_1)\int S_C(t)e^{-\Lambda_0(t)}\,dt}),$$

whereas the power is $\Phi(-z_\alpha + c_1\sqrt{a_1(1-a_1)(1-b_1)\int S_C(t)e^{-\Lambda_0(t)}\,dt})$ in a 2×2 test using $S^{*(1)}$. That is, sample size for inferring β_1 alone is effectively reduced by a multiple $1 - b_1$ under a 2×2 factorial design, just as though one had based inferences on only the subjects with $Z_i^{(2)} = 0$. As Table 1 shows, the power of a test for $\beta_1 \neq 0$ based on the MPLE using only the three factor-levels $(Z^{(1)}, Z^{(2)}) = (0, 0), (0, 1), (1, 0)$ is asymptotically the same as that based on $S^{*(1)}$. Thus in the "three-group design," if none of the sample had been randomized to $A\&B$ treatment, the sample size in the three groups would have been larger by a multiple $(1 - a_1b_1)^{-1}$. The three-group design, with power between two-group and factorial, allows inference for both β_1 and β_2 but not for interactions. Are the losses in effective sample size in the factorial study worth the combination of two trials into one? That depends on whether β_1 is the interesting parameter, whether interactions are interesting for their own sake, and whether interactions will intensify or dilute the separate treatment effects. Note however that Table 1, together with the simplified expressions for the uncensored case, imply that whenever the interaction is in the same direction as the main A-effect, the stratified log-rank statistic for testing A-treatment-effect is at least as powerful as a two-group log-rank test within a two-group trial omitting treatment B.

We examine next the choice of statistics and resulting power in the 2×2 factorial trial. A glance at the results (10) for cases where there is no censoring or alternatives are local to H_0, suggests that one need not consider the MPLE main-effect estimator in place of the stratified log-rank. Combining (10) and Table 2 suggests the following recommendations. If interactions are either weak or in the same direction as both of the separate β coefficients, then the (stratified) log-rank is the statistic of choice for each treatment effect, and can be much more powerful than the adjusted log-rank $S^{*(i)}$. For example, if $a_1 = b_1 = .5$, then for detecting fixed local alternatives to H_0 or $H_{0,A}$ with equal coefficients $c_1 = c_3$, approximately $\frac{2}{9}$ as large a sample would be required to achieve a specified large power with the stratified log-rank as with the log-rank $S^{*(1)}$. However, the situation is much different if the interaction β_3 works strongly in the direction opposite to β_1 or β_2. In such cases, where for example $c_1c_3 < 0$ in contiguous alternatives to H_0, adjusted log-rank or three-group analysis will detect effects of A with greater power than log-rank or stratified log-rank whenever $(1 + b_1c_3/c_1)^2 < 1 - b_1$, in which case the factorial trial will have less power for detecting effects of A than would a two-group trial of the same size omitting treatment B. For alternatives contiguous to $H_{0,A}$ with $\beta_2 > 0$ and $c_1c_3 > 0$, Table 2 indicates that the log-rank $S^{(1)}$ will be less powerful than S_A^{str}. A conservative approach would be always to test for A-effects with either S_A^{str} or $S^{*(1)}$. One reasonable approach, considering also the desirability of independent tests of A and B effects from the standpoint of reporting results, would be to test the treatment less likely to have a strong separate effect by the adjusted log-rank, and to test for the other treatment effect using stratified log-rank. If in advance of the trial there is good reason to suppose one effect stronger than the other, then this information could be used in planning sample size to achieve desired power for the analyses most

likely to be reported. However, the statistician should also plan a fall-back method of analysis in case prior expectations were to prove wrong, and if possible should allow sufficient sample size to achieve adequate power to detect important effects also in such a case.

It is important to understand how much power there is for discriminating interactions as opposed to main effects in a factorial study. In the simplest case of a large trial with simultaneous entry ($\varepsilon_i = 0$ for all i), independent random allocation to each of two treatments, and censoring the same in all treatment groups, formula (9) of Section 3.2 says that for H_0 the standardized MPLE $\sqrt{n}\hat{\beta}$ has asymptotic covariance equal to the inverse of $D = \delta\text{cov}(Z)$, and this is true also for $H_{0,A}$ if there is no censoring. The information to detect $\beta_1 \neq 0$ divided by the information to detect $\beta_3 \neq 0$, which is found as the ratio of the lower-left entry to the upper-left entry of D^{-1}, is $1/b_1$. This result is the same as for the accelerated-failure homoscedastic linear model. Thus, for testing local alternatives to H_0, the importance ascribed to modelling and testing for interaction effects within a factorial survival study are comparable to those for ordinary factorial linear models. However, the analogy breaks down when testing $H_{0,A}$ with $\beta_2 \neq 0$, because the least squares covariance in the homoscedastic linear model is always proportional to $(\text{cov}(Z))^{-1}$, but formula (7) says that the asymptotic covariance matrix D^{-1} for $\sqrt{n}\hat{\beta}$ under (1) is not, when there is some censoring. For example, when $b = .5$, $\beta_2 = \ln(2)$, and censoring occurs in all treatment groups at the fixed time $t = \Lambda_0^{-1}(\ln(2))$, the ratio of information for β_1 to that for β_3 is 5/3 for model (1) but 2 for the linear model.

Factorial clinical trials will usually result in simultaneous reporting of statistical tests for significance of each main effect and possibly for an interaction as well. If the test statistics for separate main effects of treatments A and B are independent, then significance levels and Type II error probabilities can be reported as though the two treatments were tested in separate clinical trials. Such two-in-one reporting is a very desirable feature of factorially designed trials. As we have seen, log-rank statistics *will* be independent in large-sample survival experiments with random treatment allocation whenever loss-to-follow-up patterns are the same for all combinations of treatments, and will be approximately independent even when loss-to-follow-up varies over different treatment groups.

Acknowledgements

I am indebted to the late Dr David Byar of the National Cancer Institute for numerous fruitful discussions on the subject of this paper and for insightful comments on this and many earlier papers. I gratefully dedicate this paper to his memory.

I thank also Larry Freedman and Ed Lakatos for useful conversations about the topics discussed here. I owe to Larry Freedman the idea in Section 3.3 of comparing power for 2×2 and "three-group" trials.

Résumé

Plusieurs résultats méthodologiques récents obtenus pour des dipositifs factoriels à deux voies, dans des études de survie, sont discutés dans le cadre de la théorie asymptotique du modèle à forces de mortalité proportionnelles en présence de deux covariables binaires. Ces résultats incluent: la formulation correcte des hypothèses nulles et alternatives, le choix entre les statistiques du log-rank et du log-rank ajusté ou stratifié, la corrélation asymptotique entre les statistiques de test pour les effets principaux individualisés, la puissance asymptotique (pour les différentes méthodes d'analyse) des tests pour les effets principaux et les interactions, la comparaison entre la puissance pour le test des effets principaux dans un dispositif factoriel 2×2 et la puissance dans un essai à trois groupes sans randomisation simultanée des patients entre les traitements; enfin les problèmes d'analyse survenant lorsque les traitements sont arrêtés précocemment pour des raisons éthiques.

References

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: A large-sample study. *Annals of Statistics* **10,** 1100–1120.

Byar, D. and Piantadosi, S. (1985). Factorial designs for randomized clinical trials. *Cancer Treatment Reports* **69,** 1055–1063.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B* **34,** 187–219.

Gail, M., DeMets, D., and Slud, E. (1982). Simulation studies of increments of the two-sample logrank test for survival time data, with application to group sequential boundaries. In *Special Topics in Survival Analysis*, I.M.S. Monograph Series, J. Crowley and R. Johnson (eds). Hayward, California: Institute of Mathematical Statistics.

Gill, R. (1980). *Censoring and Stochastic Integrals*. MC Tracts. Amsterdam: Mathematical Centre.
Harrington, D. and Fleming, T. (1990). *Survival Analysis and Counting Processes*. New York: Wiley.
Jennison, C. and Turnbull, B. (1985). Repeated confidence intervals for the median survival time. *Biometrika* **72,** 619–625.
Lininger, L., Gail, M., Green, S., and Byar, D. (1979). Comparison of four tests for equality of survival curves in the presence of stratification and censoring. *Biometrika* **66,** 419–428.
Schoenfeld, D. (1981). The asymptotic properties of rank tests for the censored two-sample problem. *Biometrika* **68,** 316–319.
Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika* **70,** 315–326.
Slud, E. (1984). Sequential linear rank tests for two-sample censored survival data. *Annals of Statistics* **12,** 551–571.
Slud, E. (1992). Relative efficiency of the log rank test within a multiplicative intensity model. *Biometrika* **78,** 621–630.
Tsiatis, A., Rosner, G., and Tritchler, D. (1985). Group sequential trials with censored survival data adjusted for covariates. *Biometrika* **72,** 365–373.
Zucker, D. and Lakatos, E. (1990). Weighted linear rank statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* **77,** 853–862.

Received July 1989; *revised July and December* 1992; *accepted December* 1992.

Appendix

Asymptotic Distributional Results

This Appendix provides sources and theoretical justifications for the assertions of Sections 2–4. The approach is via stochastic integrals and compensated counting processes, as applied by Gill (1980) and Andersen and Gill (1982) to censored-data linear rank statistics, and by Slud (1984) and Sellke and Siegmund (1983) to sequential and repeated significance tests based on such statistics. The notations and assumptions are as in Sections 2–3. Assume also that the conditional laws of (T_i, Δ_i) given $Z_i = (j, k, jk)'$ are the same for all i, j, and k.

A.1 Asymptotic Behavior of Fixed-Sample Statistics

Let the process $N_{jk}(t)$ count individuals with $Z = (j, k, jk)'$ who are observed to die after no more than time t on test. Thus

$$N_{jk}(t) \equiv \sum_i I_{[Z_i = (j,k,jk)']} \Delta_i I_{[T_i \le t]}, \quad N(t) = \sum_j \sum_k N_{jk}(t).$$

Under (1), it is well known (Gill, 1980, pp. 34–37, 44–49) that

$$M_{jk}^{\beta}(t) \equiv N_{jk}(t) - \int_0^t Y_{jk}(s) e^{(j,k,jk)'\beta} \, d\Lambda_0(s)$$

is a martingale in t, as is $M^{\beta}(t) \equiv \Sigma_j \Sigma_k M_{jk}^{\beta}(t)$, where $Y_{jk}(t) \equiv \Sigma_i I_{[Z_i=(j,k,jk),\ T_i \ge t]}$. Here the underlying filtration $\mathcal{F}_t$ is generated by the data $\{Z_i, \Delta_i I_{[T_i \le t]}, T_i I_{[T_i \le t]}: i = 1, \ldots, n\}$ observable up to study time t. As in Andersen and Gill (1982), the partial likelihood score statistic numerator for parameter β, using data up to study time t, is

$$\mathbf{L}(t, \beta) = \sum_{i=1}^{n} \Delta_i I_{[T_i \le t]} (Z_i - \bar{Z}_i(\beta)) = \sum_{j,k} \int_0^t \left[\begin{pmatrix} j \\ k \\ jk \end{pmatrix} - \frac{U^{(1)}(s, \beta)}{U^{(0)}(s, \beta)} \right] dM_{jk}^{\beta}(s), \qquad \text{(A.1)}$$

where for $i = 0, 1, 2$, with $v^{\otimes i}$ defined as in Section 3.1,

$$U^{(i)}(s, \beta) \equiv \sum_{j,k} \begin{pmatrix} j \\ k \\ jk \end{pmatrix}^{\otimes i} Y_{jk}(s) e^{(j,k,jk)\beta}.$$

Then $n^{-1}U^{(i)}(s, \beta)$ converges uniformly in probability as $n \to \infty$, to $u^{(i)}(s, \beta)$ defined following (7). For large n, $\mathbf{L}(t, \beta)$ is asymptotically normally distributed with mean $\mathbf{0}$ and asymptotic variance

$$D(t, \beta) \equiv \int_0^t \left\{ \frac{u^{(2)}(s, \beta)}{u^{(0)}(s, \beta)} - \left(\frac{u^{(1)}(s, \beta)}{u^{(0)}(s, \beta)} \right)^{\otimes 2} \right\} u^{(0)}(s, \beta)\, d\Lambda_0(s). \tag{A.2}$$

The MPLE $\hat{\beta}$ defined by (3) also satisfies under model (1)

$$\sqrt{n}(\hat{\beta} - \beta) - \frac{1}{\sqrt{n}} (D(\infty, \beta))^{-1} \mathbf{L}(\infty, \beta) \xrightarrow{P} 0 \quad \text{as } n \to \infty, \tag{A.3}$$

and $\hat{D}(\infty, \hat{\beta})$ defined in (8) consistently estimates $D \equiv D(\infty, \beta)$. Note that $\hat{D}$ in (3) converges in probability to

$$D^{(0)}(\beta) \equiv \int_0^\infty \left\{ u^{(2)}(s, \beta) - 2u^{(1)}(s, \beta) \left(\frac{u^{(1)}(s, 0)}{u^{(0)}(s, 0)} \right)' + u^{(0)}(s, \beta) \left(\frac{u^{(1)}(s, 0)}{u^{(0)}(s, 0)} \right)^{\otimes 2} \right\} d\Lambda_0(s),$$

so that $S^{(i)} - L^{(i)}(\infty, \mathbf{0})/\sqrt{nD_{ii}^{(0)}(\beta)} \xrightarrow{P} 0$ for $i = 1, 2, 3$. Under model (1) with assumption (2), it turns out that $D_{11}^{(0)}(0, \beta_2, 0) = D_{11}(\infty, 0, \beta_2, 0)$.

The adjusted log-rank statistics studied by Tsiatis, Rosner, and Tritchler (1985) were defined from numerators $L(t, \beta)$ by substituting restricted MPLEs for the nonzero components of β (under the null hypothesis). Our usage in defining $\mathbf{S}^*$ is to replace the nonzero components of β by corresponding components of the unrestricted MPLE $\hat{\beta}$, i.e., to define $\mathbf{S}^*$ by standardizing the numerators

$$\mathbf{L}^* \equiv (L^{(1)}(\infty, 0, \hat{\beta}_2, \hat{\beta}_3), L^{(2)}(\infty, \hat{\beta}_1, 0, \hat{\beta}_3), L^{(3)}(\infty, \hat{\beta}_1, \hat{\beta}_2, 0))'$$

using variance expression $n\hat{D}_{ii}$ for component i, $i = 1, 2, 3$. By (A.3) and the delta method, the components of $\mathbf{S}^*$ differ asymptotically negligibly from the adjusted log-rank statistics of Tsiatis, Rosner, and Tritchler (1985) when the corresponding components of β are of order $n^{-1/2}$. Moreover, under H_0 (respectively, under $H_{0,A}$), all (respectively, the first and third) components of

$$n^{-1/2}(\mathbf{L}^* - \operatorname{diag}(D_{11}, D_{22}, D_{33}) D^{-1} \mathbf{L}(\infty))$$

converge in probability to 0. It follows immediately that under $H_{0,A}$, $S^{(2)}(\infty)$ is asymptotically uncorrelated (and therefore, by joint asymptotic normal distribution, independent) of $(S^{*(1)}, S^{*(3)})$. By (A.3), $n^{1/2}(\hat{\beta}_1/\sqrt{(D^{-1})_{11}}, \hat{\beta}_3/\sqrt{(D^{-1})_{33}})$ differs asymptotically negligibly from $(S^{*(1)}, S^{*(3)})$ under $H_{0,A}$.

In (2), we assume that the first and second components of Z_i are independent, so that $\rho_{jk} = a_j b_k$ and Y_i is independent of Z_i. If in addition $\beta = 0$, then for all (j, k), $\Pr\{Z = (j, k, jk)' | T_1 \geq t\} = a_j b_k$. Throughout the Appendix, wherever it simplifies notation, we write a and $(1 - a)$ instead of a_1 and a_0 and b, $(1 - b)$ instead of b_1 and b_0. For each t the first two components of Z_i are conditionally independent given $T_i \geq t$. By (A.2), in this case $D(t)$ is equal to

$$\begin{pmatrix} a(1-a) & 0 & ab(1-a) \\ 0 & b(1-b) & ab(1-b) \\ ab(1-a) & ab(1-b) & ab(1-ab) \end{pmatrix} \Pr(\Delta = 1, T \leq t).$$

Thus $S^{(1)}$ and $S^{(2)}$ are asymptotically independent, but $S^{*(1)}$ and $S^{*(2)}$ are not, since the (1, 2) entry of $D(t)^{-1}$ is nonzero. In the setting of Tsiatis, Rosner, and Tritchler (1985), which was not factorial since the interaction term involving β_3 was not present, we could conclude as they did that the first two components of $\mathbf{S}$ would be asymptotically equivalent to those of $\mathbf{S}^*$, but that is definitely *not* the case here, since $\mathbf{S}^*$ is obtained by standardizing

$$\begin{pmatrix} 1/(1-b) & a/(1-b) & -1/(1-b) \\ b/(1-a) & 1/(1-a) & -1/(1-a) \\ \dfrac{-b(1-ab)}{(1-a)(1-b)} & \dfrac{-a(1-ab)}{(1-a)(1-b)} & \dfrac{1-ab}{(1-a)(1-b)} \end{pmatrix} \mathbf{L}(\infty, \mathbf{0}).$$

A further statistic discussed in the paper under $H_{0,A}$ is the stratified log-rank (5) for testing the presence of a nonzero effect for A. Since the numerator of this statistic has the representation

$$\sum_{k=0}^{1} \int \left\{ \frac{Y_{0k}}{Y_{\cdot k}}\, dM_{1k}^{\beta} - \frac{Y_{1k}}{Y_{\cdot k}}\, dM_{0k}^{\beta} + \frac{Y_{1k}Y_{0k}}{Y_{\cdot k}}\, e^{k\beta_2}(e^{\beta_1 + k\beta_3} - 1)\, d\Lambda_0 \right\}, \tag{A.4}$$

the stratified log-rank statistic is asymptotically equivalent to the ordinary log-rank $S^{(1)}$ under H_0.

A.2 Power Formulas and Comparisons

All of the foregoing statements about asymptotic equivalence and independence of statistics under null hypotheses extend automatically to contiguous alternatives, under which asymptotic formulas for power will now be developed, following the methods of Gill (1980) and Schoenfeld (1981). The alternatives are taken to be close to H_0 or $H_{0,A}$ in a way that depends on n so as to yield a limiting power strictly between 0 and 1 as $n \to \infty$, viz.

$H_{1,n}$: Z_i and C_i are as under $H_{0,A}$, and conditionally given $Z_i = (j, k, jk)'$, X_i is independent of C_i and has cumulative hazard $\int_0^t \exp(k\beta_2 + jn^{-1/2}(c_1(s) + kc_3(s)))\, d\Lambda_0(s)$,

where c_1 and c_3 are fixed functions such that $\int e^{\varepsilon(|c_1(s)| + |c_3(s)|) - \exp(k\beta_2)\Lambda_0(s)}\, d\Lambda_0(s) < \infty$ for some $\varepsilon > 0$. Alternatives $H_{1,n}$ are *contiguous* to $H_{0,A}$ in the LeCam–Hajek sense that any convergence in probability under $H_{0,A}$ persists under $H_{1,n}$ (Gill, 1980).

Assume that (2) holds, let $\gamma \equiv e^{\beta_2}$, and take c_1 and c_3 to be constants. Under alternatives $H_{1,n}$, the martingales $M_{jk}^{\beta}(t)/\sqrt{n}$ have asymptotic means $ja_jb_k(c_1 + kc_3)\int_0^t \gamma^k e^{-\gamma^k \Lambda_0(s)} S_C(s)\, d\Lambda_0(s)$ and asymptotic variances $a_jb_k \int_0^t \gamma^k e^{-\gamma^k \Lambda_0(s)} S_C(s)\, d\Lambda_0(s)$, where $S_C(s) = \Pr(C \geq s)$ as in Section 3.3. After some algebra and the change-of-variable $x = \Lambda_0(s)$, the asymptotic covariance matrix D of formula (A.2) takes the form

$$D(\infty, 0, \beta_2, 0) = \int \sum_{j,k} a_jb_k\gamma^k e^{-\gamma^k x} S_C(\Lambda_0^{-1}(x)) \begin{pmatrix} j - a \\ k - \dfrac{b\gamma e^{-\gamma x}}{b\gamma e^{-\gamma x} + (1-b)e^{-x}} \\ jk - \dfrac{ab\gamma e^{-\gamma x}}{b\gamma e^{-\gamma x} + (1-b)e^{-x}} \end{pmatrix}^{\otimes 2} dx.$$

Further manipulations with D show that

$$D = \begin{pmatrix} D_{11} & 0 & D_{13} \\ 0 & D_{22} & aD_{22} \\ D_{13} & aD_{22} & D_{13} + a^2 D_{22} \end{pmatrix},$$

where

$$\begin{pmatrix} D_{13} \\ D_{11} - D_{13} \end{pmatrix} = a(1-a) \int \begin{pmatrix} b\gamma e^{-\gamma x} \\ (1-b)e^{-x} \end{pmatrix} S_C(\Lambda_0^{-1}(x))\, dx,$$

$$D_{22} = b(1-b) \int \frac{\gamma e^{-(1+\gamma)x}}{b\gamma e^{-\gamma x} + (1-b)e^{-x}}\, S_C(\Lambda_0^{-1}(x))\, dx,$$

so that

$$D^{-1} = \frac{1}{D_{11} - D_{13}} \begin{pmatrix} 1 & a & -1 \\ a & \dfrac{D_{11} - D_{13}}{D_{22}} + a^2 \dfrac{D_{11}}{D_{13}} & -a\dfrac{D_{11}}{D_{13}} \\ -1 & -aD_{11}/D_{13} & D_{11}/D_{13} \end{pmatrix}.$$

Now under $H_{1,n}$ the standardized MPLEs $n^{1/2}(\hat{\beta}_1/\sqrt{(D^{-1})_{11}},\ \hat{\beta}_3/\sqrt{(D^{-1})_{33}})$, which are equivalent to $(S^{*(1)}, S^{*(3)})$, have asymptotic expectations $(c_1/\sqrt{(D^{-1})_{11}},\ c_3/\sqrt{(D^{-1})_{33}})$. Using the displayed form for D^{-1}, we find the asymptotic means $c_1\sqrt{D_{11} - D_{13}}$ for $S^{*(1)}$ and $c_3\sqrt{D_{13} - D_{13}^2/D_{11}}$ for $S^{*(3)}$. The main-effect MPLE $\hat{\beta}_1 + b\hat{\beta}_3$ would be standardized by the multiple

$$\sqrt{n(D_{11} - D_{13})}/\sqrt{1 - 2b + b^2 D_{11}/D_{13}},$$

and the standardized statistic would have asymptotic mean as given in the fourth line of Table 1.

Using the representation (A.4) for a statistic proportional to S_A^{str}, up to a nonrandom constant multiple, we find the asymptotic mean to equal $\sqrt{n}(c_1 D_{11} + c_3 D_{13})$ and the asymptotic standard deviation $\sqrt{n D_{11}}$. Therefore the standardized statistic S_A^{str} has asymptotic mean given by the third line of Table 1.

Finally, we need the asymptotic mean and variance for the log-rank statistic for treatment A under $H_{1,n}$. Since β_2 is not 0, the two-group model that ignores treatment-B groups does not follow a proportional hazards model, but has hazard intensity

$$\lambda(t|Z^{(1)} = j) = \Lambda_0'(t) e^{j\beta_1} \frac{b\gamma e^{j\beta_3 - \gamma e^{j(\beta_1+\beta_3)}\Lambda_0(t)} + (1-b)e^{-e^{j\beta_1}\Lambda_0(t)}}{be^{-\gamma e^{j(\beta_1+\beta_3)}\Lambda_0(t)} + (1-b)e^{-e^{j\beta_1}\Lambda_0(t)}}. \tag{A.5}$$

Since this conditional intensity does not depend on j under $H_{0,A}$, the asymptotic variance of the log-rank numerator $L^{(1)}(\infty, \mathbf{0})$ is still nD_{11}. To find the asymptotic expectation of $L^{(1)}(\infty, \mathbf{0})$ under $H_{1,n}$ for the model (A.5), one reasons as in Slud (1992, p. 624) to obtain the expression

$$\sqrt{n}\, a(1-a) \int S_C(s)(b\gamma e^{-\gamma\Lambda_0(s)} + (1-b)e^{-\gamma\Lambda_0(s)})$$

$$\left(c_1 \frac{\partial}{\partial\beta_1} \ln\lambda(t|Z^{(1)} = 1) + c_3 \frac{\partial}{\partial\beta_3} \ln\lambda(t|Z^{(3)} = 1)\right)_{\beta_1=\beta_3=0} d\Lambda_0(s).$$

After some algebra and the change-of-variable $x = \Lambda_0(s)$, the resulting asymptotic expectation for the standardized log-rank statistic $S^{(1)}$ reduces to the expression given in the first line of Table 1.

In the "three-group" case where the sample of size n is split only among $(Z^{(1)}, Z^{(2)}) = (0, 0)$, $(1, 0)$, and $(0, 1)$ in the proportions $(1-a)(1-b)$, $a(1-b)$, and $(1-a)b$, the model (1)–(2) is the same as a proportional hazards model with covariate $\zeta = (\zeta_1, \zeta_2)' = (Z^{(1)}, Z^{(2)})'$ with coefficients $\beta = (\beta_1, \beta_2)$, where $\Pr(\zeta = (1,0)') = a(1-b)/(1-ab)$ and $\Pr(\zeta = (0,1)') = b(1-a)/(1-ab)$. Under $H_{0,A}$, the quantities $u^{(i)}(s, \beta)$ in the formula (A.2) for matrix $D \equiv D(\infty, \beta)$ must now be replaced by

$$\bar{u}^{(i)}(s, \zeta) \equiv \frac{1}{1-ab} \sum_{(j,k)\neq(1,1)} \binom{j}{k}^{\otimes i} \rho_{jk} \exp\left\{\binom{j}{k}' \beta - \Lambda_0(s) e^{(j,k)'\beta}\right\} S_C(s).$$

After some algebra, the resulting information matrix $\bar{D}$ reduces to

$$\bar{D} \equiv \begin{pmatrix} \dfrac{D_{11} - D_{13}}{1-ab} + a^2 \bar{D}_{22} & -a\bar{D}_{22} \\ -a\bar{D}_{22} & \bar{D}_{22} \end{pmatrix},$$

where

$$\bar{D}_{22} = \frac{(1-a)b(1-b)}{1-ab} \int \frac{\gamma e^{-(1+\gamma)x}}{(1-a)b\gamma e^{-\gamma x} + (1-b)e^{-x}} S_C(\Lambda_0^{-1}(x))\, dx.$$

Thus $(\bar{D}^{-1})_{11} = (1-ab)/(D_{11} - D_{13})$, and the asymptotic mean divided by standard deviation for the three-group MPLE of β_1 under $H_{1,n}$ is $c_1\sqrt{D_{11} - D_{13}}/\sqrt{1-ab}$. In the setting of Table 1, where the total sample size allocated to the treatment cells (0, 0), (1, 0), and (0, 1) is $(1-ab)n$ instead of n, this asymptotic standardized mean must be multiplied by $\sqrt{1-ab}$, yielding the last line of Table 1.

A.3 Extensions to Repeated Significance Tests

Up to this point, the statistics to be used for testing have been calculated in a fixed-sample setting without regard to chronological times of entry. We follow Slud (1984) in extending now to significance testing based on statistics computed repeatedly at successive chronological times. [See the Appendix of Jennison and Turnbull (1985) for a simplified discussion.] Suppose that the entry times ε_i are independent of each other and all other survival, treatment, and censoring variables and have distribution function $F_\varepsilon(\cdot)$. Using the additional subscript τ as indicated in Section 3.4 for statistics and processes to indicate that they are calculated based only on the data that would be observable up to chronological time τ, we find that $\mathbf{L}_\tau(t, \beta)$ are martingales in t for each τ, given as in (A.1) by

$$\mathbf{L}_\tau(t, \beta) = \sum_{j,k} \int_0^t \{(j, k, jk)' - U_\tau^{(1)}(s, \beta)/U_\tau^{(0)}(s, \beta)\}\, dM_{jk\tau}^{\beta}(t)$$

and have exactly uncorrelated increments with respect to τ (Slud, 1984, Prop. 2.5, Thm 4.1, and Lemma 4.2). The asymptotic variance of $\mathbf{L}_\tau(t, \beta)/\sqrt{n}$ is given now by formula (A.2) after inserting an extra factor $F_\epsilon(\tau - s)$ into the integrand, and again a consistent variance estimator is provided by $\hat{D}_\tau$ defined in (8) with all quantities $\hat{\beta}$, $\bar{Z}_i(\beta)$, Δ_i given subscripts τ, i.e., calculated based only on data available up to real time τ. Moreover, the equivalences proved above between statistics **S** or **S*** and expressions involving $\hat{\beta}$ and D remain valid with subscripts τ for each chronological time τ. Similarly, assertions of approximate independence for log-rank statistics $S^{(1)}$ and $S^{(2)}$ or $S^{(1)}$ and $S^{*(2)}$ continue to hold for processes in τ for these processes because all of them differ only by nonrandom functions of τ from asymptotically Gaussian independent-increments processes.