JOHN LACHIN, BIOSTATISTICAL METHODS JOHN WILEY & SONS, NY 2000 pp 409-412



a $c$-fold change in $X(t) = \text{HbA}_{1c}$ ($c > 0$). Therefore, the estimated coefficient $\widehat{\beta} = 3.17$ represent a 35.3% increase in the risk of developing nephropathy per 10% higher value of the current mean $\text{HbA}_{1c}$ at any point in time ($c = 1.1$), or a 28.4% decrease in risk per 10% lower $\text{HbA}_{1c}$ ($c = 0.9$). Using the 95% confidence limits for $\beta$ yields 95% confidence limits for the risk reduction per 10% lower $\text{HbA}_{1c}$ of (16.2, 38.8%).

Among the measures of explained variation described in Section 9.4.8, only the crude approximate measure $R^2_{LR}$ may be readily applied because the model included time-dependent covariates. Based on the likelihood ratio chi-square test, the log of the current mean $\text{HbA}_{1c}$ explains $100(1 - \exp(-17.811/316)) = 5.48\%$ of the variation in risk. In DCCT (1995), such measures were used to describe the relative importance of different covariates, not as an absolute measure of explained variation.

In the analyses of the total conventional group, stratified by primary and secondary cohort, and also adjusting for the baseline level of the log albumin excretion rate, the estimated coefficient $\widehat{\beta} = 2.834$ corresponds to 25.8% risk reduction per 10% lower $\text{HbA}_{1c}$, with similar risk reductions in the primary and secondary intervention cohorts (DCCT, 1995). Nearly equivalent results were obtained in the intensive treatment group, $\widehat{\beta} = 2.639$. Thus virtually all of the difference between the treatment groups in the risk of developing microalbuminuria was attributable to the differences in the level of glycemia as represented by the $\text{HbA}_{1c}$.

## 9.5 EVALUATION OF SAMPLE SIZE AND POWER

### 9.5.1 Exponential Survival

In general, a distribution-free test such as the Mantel-logrank test is used for the analysis of survival (event-time) data from two groups. In principle, the power of any such test can be assessed against any particular alternative hypothesis with hazard functions that differ in some way over time. It is substantially simpler, however, to consider the power of such tests assuming some simple parametric model. The Mantel-logrank test is the most commonly used test in this setting, which is asymptotically fully efficient against a proportional hazards or Lehmann alternative. The simplest parametric form of this model is the exponential model with constant hazard rates $\lambda_1$ and $\lambda_2$ over time in each group.

Asymptotically, the sample estimate of the log hazard rate $\log(\widehat{\lambda})$ is distributed as $\mathcal{N}[\log(\lambda),\ E(D|\lambda)^{-1}]$. Thus the power of the test depends on the expected total number of events $E(D|\lambda)$ to be observed during the study. Here $E(D|\lambda) = NE(\delta|\lambda)$, where $\delta$ is a binary variable representing observation of the event ($\delta = 1$) versus right censoring of the event time ($\delta = 0$); and $E(\delta|\lambda)$ is the probability that the event will be observed as a function of $\lambda$ and the total exposure of the cohort (patient years of follow-up). The test statistic then is $T = \log(\widehat{\lambda}_1/\widehat{\lambda}_2)$. Under $H_0$: $\lambda_1 = \lambda_2 = \lambda$ and the statistic has expectation $\mu_0 = \log(\lambda_1/\lambda_2) = 0$, while under $H_1$, $\mu_1 = [\log(\lambda_1) - \log(\lambda_2)]$. As in Section 3.3.1, let $\zeta_i$ refer to the expected

sample fraction expected in the $ith$ group ($i = 1, 2$) where $E(n_i) = N\zeta_i$. Then the variance of the test statistic under the alternative hypothesis is

$$V(T|H_1) = \sigma_1^2 = \frac{1}{N}\left[\frac{1}{\zeta_1 E(\delta|\lambda_1)} + \frac{1}{\zeta_2 E(\delta|\lambda_2)}\right] \tag{9.112}$$

and under the null hypothesis is

$$V(T|H_0) = \sigma_0^2 = \frac{1}{N}\left[\frac{1}{E(\delta|\lambda)}\left(\frac{1}{\zeta_1\zeta_2}\right)\right]. \tag{9.113}$$

The basic equation relating sample size $N$ and power $Z_{1-\beta}$ is

$$\sqrt{N}\,|\log(\lambda_1) - \log(\lambda_2)| = Z_{1-\alpha}\left[\frac{1}{E(\delta|\lambda)\zeta_1\zeta_2}\right]^{1/2} + Z_{1-\beta}\left[\frac{1}{\zeta_1 E(\delta|\lambda_1)} + \frac{1}{\zeta_2 E(\delta|\lambda_2)}\right]^{1/2}, \tag{9.114}$$

where $\lambda = \zeta_1\lambda_1 + \zeta_2\lambda_2$ analogously to (3.34) in the test for proportions.

Lachin (1981) and Lachin and Foulkes (1986) present a similar expression for the case where the test statistic is the difference in estimated hazard rates, $T = \widehat{\lambda}_1 - \widehat{\lambda}_2$. Freedman (1982) shows that these expressions can also be derived from the null and alternative distributions of the Mantel-logrank statistic. As cited by Lachin and Foulkes (1986), computations of sample size and power using the difference in hazards are conservative relative to those using the log hazard ratio in that the former yields larger required sample sizes, lower computed power for the same values of $\lambda_1$ and $\lambda_2$. As the difference in hazards approaches zero, the difference in the two methods also approaches zero. In some respects the use of the difference in hazards would be preferred because, in fact, the Mantel-Haenszel (logrank) test statistic can be expressed as the weighted sum of the difference in the estimated hazards between groups (see Section 9.6.3). Herein, however, we use the log hazard ratio because generalizations also apply to the Cox PH model.

In the simplest case of a study with no censoring of event times, $E(\delta|\lambda) = 1$ and the event times of all subjects are observed ($N = D$). In this case, the total number of events $D$ and power $Z_{1-\beta}$ are obtained as the solutions to

$$\sqrt{D}\,|\log(\lambda_1) - \log(\lambda_2)| = \frac{Z_{1-\alpha} + Z_{1-\beta}}{\sqrt{\zeta_1\zeta_2}}. \tag{9.115}$$

Thus the total number of events $D$ required to provide power $1 - \beta$ to detect a specified hazard ratio in a test at level $\alpha$ is

$$D = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\zeta_1\zeta_2\,[\log(\lambda_1/\lambda_2)]^2} \tag{9.116}$$

(George and Desu, 1974; Schoenfeld, 1981). Usually, however, there are censored event times because of administrative curtailment of follow-up (administrative censoring) or because of random losses to follow-up.

To allow for administrative censoring, let $T_S$ designate the maximum total length of study. In the simplest case, each subject is followed for $T_S$ years of exposure and $E(\delta|\lambda) = 1 - e^{-\lambda T_S}$. Typically, however, patients enter a study during a period of recruitment of $T_R$ years and are then followed for a maximum total duration of $T_S$ years ($T_S \geq T_R$) so that the first patient entered is followed for $T_S$ years and the last patient for $T_S - T_R$ years. In a study with uniform entry over the recruitment interval of $T_R$ years and with no random losses to follow-up, it is readily shown (see Problem 9.1.5) that

$$E(\delta|\lambda) = \left[1 - \frac{e^{-\lambda(T_S - T_R)} - e^{-\lambda T_S}}{\lambda T_R}\right] \tag{9.117}$$

(Lachin, 1981). Substitution into (9.114) and solving for $N$ yields the sample size needed to provide power $1 - \beta$ in a test at level $\alpha$ to detect a given hazard ratio, say $N(\lambda)$.

Rubenstein, Gail and Santner (1981) and Lachin and Foulkes (1986) present generalizations that allow for randomly censored observations because of loss to follow-up. Let $\gamma_i$ be an indicator variable that denotes loss to follow-up at random during the study prior to the planned end at $T_S$. If we assume that times of loss are exponentially distributed with constant hazard rate $\eta$ over time, then for a study with uniform entry the probability of the event is

$$E(\delta|\lambda, \eta) = \frac{\lambda}{\lambda + \eta}\left[1 - \frac{e^{-(\lambda+\eta)(T_S - T_R)} - e^{-(\lambda+\eta)T_S}}{(\lambda + \eta)T_R}\right] \tag{9.118}$$

and the probability of loss to follow-up is

$$E(\gamma|\lambda, \eta) = \frac{\eta}{\lambda} E(\delta|\lambda, \eta) \tag{9.119}$$

(see Problem 9.1.6). When $\eta_1 = \eta_2$ for the two groups, then the equation relating sample size and power is obtained by substituting $E(\delta|\lambda, \eta)$ in (9.114) evaluated at $\lambda_1$, $\lambda_1$, and $\lambda$. When $\eta_1 \neq \eta_2$, then the probability of the event in each group under $H_0$ will differ. In this case, the general equation in (9.114) is modified to employ the term

$$Z_{1-\alpha}\left[\frac{1}{\zeta_1 E(\delta|\lambda, \eta_1)} + \frac{1}{\zeta_2 E(\delta|\lambda, \eta_2)}\right]^{1/2}. \tag{9.120}$$

In the more general case, where losses to follow-up may not be exponentially distributed, Lachin and Foulkes (1986) show that the sample size with losses that follow any distribution $G(t)$ can be obtained approximately as

$$N[\lambda, G(t)] \doteq N(\lambda)\frac{E(\delta|\lambda)}{E[\delta|\lambda, G(t)]}, \tag{9.121}$$

where $E[\delta|\lambda, G(t)]$ is the probability of an event during the study. Thus random losses to follow-up, whatever their distribution, that result in a 10% reduction in

the probability of the event being observed require that the sample size needed with no losses to follow-up, $N(\lambda)$, be increased by 11.1%.

For cases where the exponential model is known not to apply, Lakatos (1988) describes the assessment sample size based on the power function of a weighted Mantel-Haenszel test against an arbitrarily specified alternative hypothesis. In this procedure one specifies the hazard rates in the control and experimental groups over intervals of time along with other projected features of the study, such as the proportion censored or lost to follow-up within each interval. This allows the assessment of power under any alternative, including cases where the hazards may not be proportional, or may even cross, and where the pattern of random censoring is uneven over time and may differ between groups. Wallenstein and Wittes (1993) also describe the power of the Mantel-logrank test in an analysis where the hazards need not be constant nor proportional over time, although the Lakatos procedure is more general.

### 9.5.2 Cox's Proportional Hazards Model

As with the logistic regression model, the power function of a test for the vector of coefficients in the Cox PH model is a function of the joint distribution of the vector of covariates among those with the event and those not at each event time, which are unknown. However, one can assess the power of a Wald or score test in a PH model that is stratified by other factors. Under an exponential model, Lachin and Foulkes (1986) describe the relationship between sample size and power for a test of the difference in hazards for two groups that is stratified by other factors. Since the logrank test is the score test in a Cox PH model, these methods also apply, approximately, to a test of the difference between groups in a Cox PH model with other binary covariates that define a set of strata.

Schoenfeld (1983) showed that the probability of events in each of two groups under the assumption of constant proportional hazards over strata can be obtained if one has information on the survival function over time in the control group. Let $E(\delta)$ denote the resulting probability of an event in the total sample. Then the sample size required to provide power $1-\beta$ in a test at level $\alpha$ to detect a given hazard ratio ($RR$) is provided as $N = D/E(\delta)$, where $D$ is the total number of events required from (9.116) substituting $RR$ for $\lambda_1/\lambda_2$. Many, such as Palta and Amini (1985), describe generalizations of this approach.

**Example 9.10** *Lupus Nephritis: A Study*

Lewis, Hunsicker, Lan, et al. (1992) describe the results of a clinical trial of plasmapheresis (plasma filtration and exchange) plus standard immunosuppressive therapy versus standard therapy alone in the treatment of severe lupus nephritis. The sample size for this study was based on two previous studies in which the survival function was log-linear with constant hazard approximately $\lambda = 0.3$ yielding median survival of 2.31 years. Since lupus is a rare disease, recruitment was expected to be difficult. Thus initial calculations determined the sample size required to provide 90% power to detect either a 40 or 50% risk reduction (relative hazards of 0.6 and 0.5) with a one-sided test at le
events required are 131.3 for a
assuming no right censoring. In
events, respectively.

The study was planned to re
a total study duration of $T_S =$ (
up, to detect a 40% risk reduct
required. Among those in the c
of the event is $E(\delta|\lambda_2 = 0.3) =$
plasmapheresis group with haza
$0.18) = 0.5027$ with 55.8 expect
sample size required is $N = 12$
unchanged, yielding 43.3 expecte
$\lambda_1 = 0.15$, the event probability
events.

In each case, the total expecte
is approximately equal to the to
censoring. Thus the sample size
the required expected number of
the average period of exposure,
smaller the required sample size
over a period of only six months
of patients required to detect a 4
the 222 required with a four yea
period, the average duration of
expected events among the 179
average duration of follow-up i
among the 222 patients.

Because lupus is a life-threat
tients would be lost to follow-u
ditional calculation allowing for
$\eta_1 = \eta_2 = \eta = 0.05$ in each grou

| $\lambda_1$ | $N$ | $E(\gamma\|\lambda_1, \eta$ |
|---|---|---|
| 0.18 | 241.6 | 0.128 |
| 0.15 | 138.8 | 0.135 |

In the control group, for both va
lost to follow-up ($E(\gamma|\lambda_2, \eta) =$
is reduced to $E(\delta|\lambda_2, \eta) = 0.62$
40% risk reduction ($\lambda_1 = 0.18$)
group, $E(\gamma|\lambda_1, \eta)$, would be lost
event is reduced to $E(\delta|\lambda_1, \eta) =$
the total sample size is increase
no losses to follow-up. The tot