Published in final edited form as:
Stat Med. 2013 November 10; 32(25): 4413–4425. doi:10.1002/sim.5839.

Sample Size and Power for a Logrank Test and Cox Proportional Hazards Model with Multiple Groups and Strata, or a Quantitative Covariate with Multiple Strata

John M. Lachin
The Biostatistics Center Departments of Epidemiology and Biostatistics, and Statistics The George Washington University 6110 Executive Boulevard, Suite 750 Rockville, Maryland USA 20852

Summary

General expressions are described for the evaluation of sample size and power for the K group Mantel-logrank test or the Cox PH model score test. Under an exponential model, the method of Lachin and Foulkes [1] for the 2 group case is extended to the $K \geq 2$ group case using the non-centrality parameter of the $K - 1$ *df* chi-square test. Similar results are also shown to apply to the K group score test in a Cox PH model. Lachin and Foulkes [1] employed a truncated exponential distribution to provide for a non-linear rate of enrollment. Expressions for the mean time of enrollment and the expected follow-up time in the presence of exponential losses-to-follow-up are presented. When used with the expression for the non-centrality parameter for the test, equations are derived for the evaluation of sample size and power under specific designs with R years of recruitment and T years total duration.

Sample size and power are also described for a stratified-adjusted K group test and for the assessment of a group by stratum interaction. Similarly computations are described for a stratified-adjusted analysis of a quantitative covariate and a test of a stratum by covariate interaction in the Cox PH model.

Keywords

Sample size; power; logrank test; Cox Proportional Hazards Model; multiple groups; exponential survival; stratified analysis; interactions

1 Introduction

In time-to-event studies with multiple ($K \geq 2$) independent groups, we wish to test the general null hypothesis H_0: $\lambda_1(t) = \lambda_2(t) = \ldots = \lambda_K(t)$ against the alternative H_1: $\lambda_j(t) \neq \lambda_k(t)$ for some $1 \leq j < k \leq K$, where $\lambda_j(t)$ is the time-varying hazard rate within the jth group, $1 \leq j \leq K$, $t > 0$. The Mantel-logrank test is commonly used to test H_0 that also specifies equality of the survival functions of the K groups over time. This test can also be obtained as the score test for a binary covariate in the Cox proportional hazards (PH) model and is fully effcient under the proportional hazards assumption. The simplest instance is the exponential model in which the hazards are constant over time, or $\lambda_j(t) = \lambda_j$ for all $t > 0$. Thus, the exponential model is commonly employed to evaluate the sample size or power for this test.

George and Desu [2] showed that the power of a test of equality of hazards under an exponential survival model is a function of the number of subjects with the outcome event, a result also derived by Schoenfeld [3] for the Cox PH model. Following the work of Lachin [4], Rubenstein, Gail and Santner [5], and Schoenfeld and Richter [6], among others, Lachin and Foulkes [1] described the assessment of sample size and power for the test of two groups under an exponential model with possibly non-linear recruitment over *R* years and follow-up over *T* years, exponential losses to follow-up, and a stratified design where these factors may differ over strata. They did so using a test described in terms of the difference in hazards that Freedman [7] had shown to derive from the limiting distribution of the logrank test, and also using a test of the log hazard ratio that Schoenfeld [8] had shown to derive under a proportional hazards model.

Makuch and Simon [9] provide a generalization of the George-Desu result to determine the number of subjects with the event required to provide the desired level of power for the comparison of $K \geq 2$ groups using a one-way ANOVA-like chi-square test. They also describe how their results could be applied to the assessment of sample size in cases where the total exposure time is also specified. Liu and Dahlberg [10] describe the power of the Wald test in the PH model with results close to those of Makuch and Simon. Ahnn and Anderson [11] describe the power of the Tarone-Ware [12] family of tests that are a generalization of the Mantel or logrank test and present an expression explicitly for the case of equal sample sizes and equal censoring distributions. Halabi and Singh [13] generalize the Ahnn-Anderson expression for unequal sample fractions and describe the stratified adjusted test power.

Herein these results are further generalized to allow for non-uniform recruitment and losses to follow-up that yield variable exposure times, and also stratification as in Lachin and Foulkes [1]. First, the power of Cochran's ANOVA-like χ^2 test of homogeneity [10] of the hazard rates among the *K* groups is described as a function of the non-centrality parameter of the distribution of the test, that then permits assessment of sample size and/or study duration. Then, the non-centrality parameter for the *K*-group score test in a Cox proportional hazards model is shown to be approximately equal to that of Cochran's test, so that the results apply more generally than to a simple exponential model. An equivalent T^2-like contrast test is then used to evaluate power for a stratified-adjusted analysis. Methods are also presented for the assessment of the power of a stratified-adjusted *K*-group test, and for the assessment of a group by stratum interaction. Similar computations are described for a stratified-adjusted analysis of a quantitative covariate, and the test of a strata by covariate interaction in the Cox PH model.

2 Lachin-Foulkes Exponential Model

For the *j*th group, let $\hat{\theta}_j = ln\left(\hat{\lambda}_j\right)$ denote the log of the maximum likelihood estimate of the exponential hazard rate that is distributed as

$$\hat{\theta}_j \sim N\left[\theta_j, \sigma_j^2\right] \quad (1)$$

where $\theta_j = \ln(\lambda_j)$ and $\sigma_j^2 = [E(D_j)]^{-1}$ based on an expected number of subjects $E(D_j)$ to experience the outcome event (termed events), for $j = 1, \ldots, K \geq 2$. Within the jth group, for a given pattern of enrollment and losses-to-follow-up, and hazard rate for the event, the probability π_j that the event is observed is determined. Then, for a given sample size n_j within the jth group, the expected number of subjects with the event is obtained as $E(D_j) = n_j\pi_j$.

To allow for non-uniform entry over a recruitment period of R years, Lachin and Foulkes [1] employed a truncated exponential distribution for the enrollment time r with shape parameter γ and density $g(r)$ and cumulative distribution $G(r)$

$$g(r) = \frac{\gamma e^{-\gamma r}}{1 - e^{-\gamma R}}, \quad G(r) = \frac{1 - e^{-\gamma r}}{1 - e^{-\gamma R}} \quad (2)$$

for $0 < r \leq R$, and $\gamma \neq 0$, that yields a concave pattern for $\gamma < 0$, a convex pattern for $\gamma < 0$. The mean enrollment time is readily shown to be

$$E(r) = \int_{r=0}^{R} \frac{r\gamma e^{-\gamma r}}{1 - e^{-\gamma R}} dr = \frac{1 - e^{-R\gamma}(1 + R\gamma)}{\gamma[1 - e^{-R\gamma}]}. \quad (3)$$

In the case of exponential losses with loss hazard rate η, density $h(u) = \eta e^{-\eta u}$ and cumulative distribution function $H(u) = 1 - e^{-\eta u}$, then the expected potential exposure time is

$$E(f|\eta, \gamma) = \int_{r=0}^{R} \left(\int_{u=0}^{T-r} u\, h(u)\; g(r)\, du + (T - r)[1 - H(T - r)] \right) dr \quad (4)$$

that is easily evaluated numerically.

Then, for a total study duration of $T \geq R$ years with exponential event hazard rate λ and loss-to-follow-up hazard rate η, Lachin and Foulkes [1] show that

$$\pi = \frac{\lambda}{\lambda + \eta} + \frac{\lambda\gamma e^{-(\lambda+\eta)T}\left[1 - e^{(\lambda+\eta-\gamma)R}\right]}{(\lambda+\eta)(\lambda+\eta-\gamma)(1 - e^{-\gamma R})}. \quad (5)$$

For uniform recruitment with density $g(z) = 1/R$, this simplifies slightly to

$$\pi = \frac{\lambda}{\lambda + \eta}\left[1 - \frac{e^{-(\lambda+\eta)(T-R)} - e^{-(\lambda+\eta)T}}{R(\lambda+\eta)}\right]. \quad (6)$$

Lachin and Foulkes [1] also show that the probability of loss-to-follow-up (non-administrative right censoring) is simply $\pi\eta/\lambda$.

3 An ANOVA-like Test

Under the assumption that there is no difference among groups, or $\theta_1 = \ldots = \theta_K = \theta$, it is well known (cf. [14]) that a consistent estimate of the common log hazard rate θ is provided by the minimum variance linear estimator (MVLE)

$$\hat{\theta}=\frac{\sum_{j=1}^{K}\hat{\sigma}_j^{-2}\hat{\theta}_j}{\sum_{j=1}^{K}\hat{\sigma}_j^{-2}}=\frac{\sum_{j=1}^{K}D_j\hat{\theta}_j}{\sum_{j=1}^{K}D_j} \quad (7)$$

that is obtained from the application of weighted least squares. Then, the hypothesis of no differences among groups, or H_0: $\theta_1 = \ldots = \theta_K$ can be tested using Cochran's [15] test of homogeneity,

$$X^2=\sum_{j=1}^{K}\hat{\sigma}_j^{-2}\left(\hat{\theta}_j-\hat{\theta}\right)^2=\sum_{j=1}^{K}D_j\left(\hat{\theta}_j-\hat{\theta}\right)^2, \quad (8)$$

that is asymptotically distributed as χ^2_{K-1} on $K - 1$ *df*.

The power of this test is provided by the non-centrality parameter $\psi^2 = E[X^2]$ for a given total sample size N under an appropriate model for the other parameters. Makuch and Simon [9] describe the computation of power for given values of $\{E(D_j), \theta_j\}$, $j = 1, \ldots, K$. More generally, the total sample size required, or the total amount of information required, can be obtained from the expressions of Lachin and Foulkes [1] above.

Let ζ_j denote the sample fraction of subjects assigned to the jth group, $j = 1, \ldots, K$, where $\sum_{j=1}^{K}\xi_j=1$; let $\theta_1, \ldots, \theta_K$ denote the specified set of log hazard rates that are of interest to detect under the alternative H_1: $\theta_j \neq \theta_k$ for some $1 \leq j \leq k \leq K$; and let $\eta_1, \ldots, \eta_K$ denote the assumed hazard rates of loss-to-follow-up that may vary among groups. Under an exponential model for each group with either a uniform or non-linear rate of entry yielding event probability π_j, then the resulting non-centrality parameter is

$$\begin{aligned}\psi^2 &= \sum_{j=1}^{K} E(D_j)\, E\left[\left(\hat{\theta}_j-\hat{\theta}\right)^2\right] \\ &= N\sum_{j=1}^{K}\xi_j\pi_j(\theta_j-\theta)^2 \\ &= N\phi^2\end{aligned} \quad (9)$$

where φ^2 is the "non-centrality factor" or the component remaining after factoring N, and where

$$\theta=E\left[\hat{\theta}|H_1\right]=\frac{\sum_{j=1}^{K}\xi_j\pi_j\theta_j}{\sum_{j=1}^{K}\xi_j\pi_j} \quad (10)$$

is the weighted average of the log hazard rate values within the groups under the specified alternative. Thus, the non-centrality factor and power depend on the weighted sum of squares of the deviations of the log hazards within the K groups from the mean hazard.

Values of the non-centrality parameter $\psi^2(\alpha\,\beta\,m)$ providing various levels of power for the non-central χ^2 distribution on *m df* are readily obtained from programs such as the SAS functions PROBCHI for the cumulative probabilities and CINV for quantiles of the χ^2 distribution, both of which provide computations under the non-central distribution. The SAS function CNONCT then computes the value of the non-centrality parameter ψ^2 that provides power $1 - \beta$ for specific levels of α and *m*. For a test at level α, with critical value $\chi^2_{m,1-\alpha}=CINV\,(1-\alpha,m)$, the required non-centrality parameter value is $\psi^2\,(\alpha,\beta,m)=\text{CNONCT}\left(\chi^2_{m,1-\alpha},m,\beta\right)$.

To determine sample size for a study, the value $\psi^2(\alpha, \beta, m)$ of the non-centrality parameter is obtained that will provide power $1 - \beta$ under the non-central χ^2 distribution for a test at level α on *m df*. Then the value of the non-centrality factor φ^2 under the alternative hypothesis in (9) is specified as a function of the parameter sets $\{\zeta_j\}$, $\{\pi_j\}$ and $\{\theta_j\}$. Given the value of φ^2, the *N* required to provide power $1 - \beta$ is that value for which $\psi^2(\alpha, \beta, m) = N\varphi^2$, yielding

$$N=\frac{\psi^2\,(\alpha,\beta,m)}{\phi^2}. \quad (11)$$

Alternately, for a given value of the parameter ψ^2 in (9), the level of power can be computed as

$$1-\beta=1-\text{PROBCHI}\left(\chi^2_{m,1-\alpha},m,\psi^2\right). \quad (12)$$

4 Cox PH Model Score Test

The above expressions are based on the large sample test of the difference among the hazard rates under an exponential model. An appropriate expression can also be obtained from the non-central distribution of the score test for the treatment group coeffcients in a Cox PH model. With the *K*th group as the reference, then the model would employ $K - 1$ binary covariates $(X_1, \ldots, X_{K-1})$ to represent membership in the *j*th group with coeffcient vector $\beta = (\beta_1 \ldots \beta_{K-1})$, where the *j*th coeffcient β_j equals the log hazard ratio for the *j*th group versus the *K*th reference group.

Using standard notation, let δ_i be the binary indicator variable to denote whether the *i*th subject had the outcome event ($\delta_i = 1$) or is right censored ($\delta_i = 0$), $R(t_i)$ denote the set of subjects still at risk at the event time t_i, and $n(t_i)$ be the number of subjects in the risk set at that time. For the *i*th subject, $x_i = (x_{i1} \ldots x_{i(K-1)})^T$ where $x_{ij} = 1$ denotes that the event at time t_i occurred in an individual in the *j*th group, $x_{ij} = 0$ otherwise. Then the partial likelihood, assuming no tied event times, is

$$\tilde{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} \left[\frac{e^{\mathrm{x}_i' \boldsymbol{\beta}}}{\sum_{l \in R(t_i)} e^{\mathrm{x}_l' \boldsymbol{\beta}}} \right]^{\delta_i}. \quad (13)$$

Under H_0: $\beta_1 = \ldots = \beta_{K-1} = 0$, the score equation for the jth coeffcient reduces to

$$U_0(\beta_j) = \sum_{i=1}^{N} \delta_i [x_{ij} - p_j(t_i)] \quad (14)$$

where

$$p_j(t_i) = \frac{\sum_{l \in R(t_i)} x_{lj}}{n(t_i)} \quad (15)$$

is the proportion of subjects still at risk at event time t_i that are members of the jth group. Thus,

$$\sum_{i=1}^{N} \delta_i x_{ij} = D_j \quad (16)$$

is the observed number of subjects with the outcome event among subjects in the jth group, and

$$\sum_{i=1}^{N} \delta_i p_j(t_i) = \hat{E}[D_j | H_0] \quad (17)$$

is the estimate of the expected number of events under the null hypothesis. Thus, $U_0(\beta_j) = D_j - \hat{E}[D_j | H_0]$ and the score vector is $\mathrm{U}_0(\boldsymbol{\beta}) = [U_0(\beta_1) \ldots U_0(\beta_{K-1})^T$.

Likewise, the expressions for the elements of the information matrix evaluated under H_0 reduce to

$$\begin{aligned} \mathbf{I}(\boldsymbol{\beta})_{jj} &= \sum_{i=1}^{N} \delta_i p_j(t_i)[1 - p_j(t_i)], \quad j = 1, \ldots, K-1 \\ \mathbf{I}_0(\boldsymbol{\beta})_{jk} &= \sum_{i=1}^{N} -\delta_i p_j(t_i) p_k(t_i), \quad k \neq j. \end{aligned} \quad (18)$$

Under H_0, and the assumption that there is a common censoring distribution among the K groups, then $E[p_j(t_i)] = \zeta_j$ for all event times and $\mathbf{I}_0(\boldsymbol{\beta})$ then has elements

$$\begin{aligned} \mathbf{I}_0(\boldsymbol{\beta})_{jj} &= E(D)\xi_j(1 - \xi_j), \quad j = 1, \ldots, K-1 \\ \mathbf{I}_0(\boldsymbol{\beta})_{jk} &= -E(D)\xi_j\xi_k, \quad k \neq j \end{aligned} \quad (19)$$

where D is the total number of outcome events. Then the score test of H_0 is provided by

$$X_S^2 = \mathbf{U}_0(\boldsymbol{\beta})' \mathbf{I}_0(\boldsymbol{\beta})^{-1} \mathbf{U}_0(\boldsymbol{\beta}). \quad (20)$$

Using the same steps as employed by Schoenfeld [3], under the alternative hypothesis H_1: $\beta \neq 0$, it can then be shown that

$$\mathbf{U}_0(\boldsymbol{\beta}) \sim \mathcal{N}\left[\boldsymbol{\beta}' \mathbf{I}_0(\boldsymbol{\beta}), \mathbf{I}(\boldsymbol{\beta})\right]. \quad (21)$$

Thus, X_S^2 is distributed as non-central chi-square with non-centrality parameter

$$\psi_S^2 = \boldsymbol{\beta}' \mathbf{I}_0(\boldsymbol{\beta})' \mathbf{I}(\boldsymbol{\beta})^{-1} \mathbf{I}_0(\boldsymbol{\beta}) \boldsymbol{\beta}. \quad (22)$$

Since $\mathbf{I}(\beta)$ is approximately equal to $\mathbf{I}_0(\beta)$ under local alternatives, then

$$\psi_S^2 \cong \boldsymbol{\beta}' \mathbf{I}_0(\boldsymbol{\beta})' \boldsymbol{\beta} = \mathbf{E}(\mathbf{D}) \left[\sum_{\mathbf{j}=1}^{\mathbf{K}-1} \beta_{\mathbf{j}}^2 \xi_{\mathbf{j}} (1 - \xi_{\mathbf{j}}) - 2 \sum_{\mathbf{j}=1}^{\mathbf{K}-2} \sum_{\mathbf{k}=\mathbf{j}+1}^{\mathbf{K}-1} \beta_{\mathbf{j}} \beta_{\mathbf{k}} \xi_{\mathbf{j}} \xi_{\mathbf{k}} \right]. \quad (23)$$

Now referring to the prior section, under H_0, and the assumption that there is a common censoring distribution among the K groups, then $E(D_j) \approxeq E(D)\zeta_j$ and the common parameter value is specified as

$$\theta_0 \cong \frac{\sum_{j=1}^{K} E(D) \xi_j \theta_j}{\sum_{j=1}^{K} E(D) \xi_j} = \sum_{j=1}^{K} \xi_j \theta_j \quad (24)$$

The corresponding non-centrality parameter using this simplification is

$$\psi_0^2 = E(D) \sum_{j=1}^{K} \xi_j (\theta_j - \theta_0)^2. \quad (25)$$

Noting that $\beta_j = \theta_j - \theta_K$, then it can be shown that this expression equals that provided in (23) above. Thus, under the assumption of a common censoring distribution among groups, a sample size or power computation using the exponential model also applies to the Cox PH model.

Under the assumption of equal censoring among groups and equal sample sizes, i.e. $\zeta_j = 1/K$ for $\forall j$, the above expression also is equivalent to that described by Ahnn and Anderson [11].

5 A Contrast-Based Test

An equivalent form of the test of homogeneity in (8) is obtained as a T^2-like quadratic form in contrasts among the K log hazards. Let $\boldsymbol{\theta} = (\theta_1 \ldots \theta_K)^T$ designate the vector of log hazard rates within the K groups with sample fractions $\{\zeta_j\}$ as above. For specified hazard rates $\{\lambda_{lj}\}$ and loss hazard rates $\{\eta_{lj}\}$ for the K groups, and recruitment pattern parameter γ, then

we can compute the event probability π_j such that $E(D_j) = N\zeta_j\pi_j$ is the expected number of subjects with the event in the *j*th group.

Then, the vector of estimated log hazards , $\hat{\boldsymbol{\theta}}=\left(\hat{\theta}_1 \cdots \hat{\theta}_K\right)^T$, is asymptotically distributed as multivariate normal with expectation θ and covariance matrix

$$\Sigma=diag\left[E(D_1)^{-1} \cdots E(D_K)^{-1}\right] \quad (26)$$

that is consistently estimated as $\hat{\Sigma}=diag\left[D_1^{-1} \cdots D_K^{-1}\right]$. A T^2-like contrast test of homogeneity H_0: $\theta_1 = \ldots = \theta_K$ can be constructed as

$$X^2=\left(\mathbf{C}'\hat{\boldsymbol{\theta}}\right)'\left(\mathbf{C}'\hat{\Sigma}\mathbf{C}\right)^{-1}\left(\mathbf{C}'\hat{\boldsymbol{\theta}}\right) \quad (27)$$

where

$$\mathbf{C}'=\begin{bmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \quad (28)$$

is of dimension $(K - 1) \times K$. It is well known that a T^2-like test of homogeneity with a contrast matrix of this form is equivalent to one using the difference between the *j*th group estimate and the overall average as in (8) (Anderson [16], p. 170; cf. Lachin [14], p. 151-2). For the *j*th row of C', the vector product yields the log hazard ratio for the *j*th group relative to the *K*th group, $\hat{\beta}_j=C_j'\hat{\boldsymbol{\theta}}=log\left(\hat{\lambda}_j/\hat{\lambda}_K\right)$, for $j = 1, \ldots, (K–1)$. Thus, the vector of log hazard ratios for the first *K*–1 groups versus the *K*th is $\hat{\boldsymbol{\beta}}=\left(\beta_1 \cdots \beta_{(K-1)}\right)^T$. Since $\hat{V}\left(\hat{\boldsymbol{\theta}}\right)=\hat{\Sigma}$ then $\hat{V}\left(\hat{\boldsymbol{\beta}}\right)=\hat{V}\left(\mathbf{C}'\hat{\boldsymbol{\theta}}\right)=\hat{\Omega}=\mathbf{C}'\hat{\Sigma}\mathbf{C}$,

$$\hat{\Omega}=\begin{bmatrix} D_1^{-1}+D_K^{-1} & D_K^{-1} & \cdots & D_K^{-1} \\ D_K^{-1} & D_2^{-1}+D_K^{-1} & \cdots & D_K^{-1} \\ \vdots & \vdots & \cdots & \vdots \\ D_K^{-1} & D_K^{-1} & \cdots & D_{(K-1)}^{-1}+D_K^{-1} \end{bmatrix} \quad (29)$$

where Ω consists of the like matrix defined in terms of the expected numbers of subjects with the event. Thus the non-centrality parameter of the test is provided by

$$\psi^2=\left(\mathbf{C}'\boldsymbol{\theta}\right)'\left(\mathbf{C}'\Sigma\mathbf{C}\right)^{-1}\left(\mathbf{C}'\boldsymbol{\theta}\right)=\boldsymbol{\beta}'\Omega^{-1}\boldsymbol{\beta} \quad (30)$$

that equals the expression in (9).

Then, evaluating the above covariance matrix under H_0 and the assumption of equal censoring such that $E(D_j) \approx = D\zeta_j$ leads to a non-centrality parameter that equals the expressions obtained under the PH model, ψ_S^2 in (23) and ψ_0^2 in (25).

6 A Statified-Adjusted Analysis

The above can also be generalized to a stratified-adjusted *K df* logrank test over *S* independent strata. While such a test can be conducted using the Cox PH model, an alternate approach is to consider a test based on the multivariate normal distribution of the hazard rate estimates, and its associated power function. This is easily described using the contrast test formulation.

Within the *l*th stratum, let $\boldsymbol{\theta}_l = (\theta_l \ldots \theta_{lk})^T$ designate the vector of log hazard rates within the *K* groups, with sample estimates $\hat{\boldsymbol{\theta}}_l$, for $l = 1, \ldots, S$. Denote the stratum sample fraction as $\omega_l = E(N_l/N)$, N_l being the *l*th stratum sample size, $\sum_{l=1}^{s} \omega_l = 1$; and denote the stratum-group sample fraction as $\zeta_{lj} = E(n_{lj}/N_l)$, n_{lj} being group sample size within the *l*th stratum, $\sum_{j=1}^{K} \xi_{lj} = 1$ for each *l*. For specified hazard rates λ_{lj} and loss hazard rates η_{lj} for the *lj*th cell, and recruitment pattern parameter γ_l for the *l*th stratum, then we can compute the event probability π_{lj} such that $E(D_{lj}) = N\omega_l\zeta_{lj}\pi_{lj}$ is the expected number of subjects with the event in the *lj*th cell. Within the *l*th stratum, asymptotically $\hat{\boldsymbol{\theta}}_l \sim \mathcal{N}(\boldsymbol{\theta}_l, \Sigma_l)$ where $\Sigma l = diag[E(D_{li})^{-1} \ldots E(D_{lK})^{-1}]$ that is consistently estimated as $\hat{\Sigma}_l = diag\left[D_{l1}^{-1} \cdots D_{lK}^{-1}\right]$.

As above, vector the of log hazard ratios in the *l*th stratum is $\hat{\boldsymbol{\beta}}_l = \left(\beta_{l1} \cdots \beta_{l(K-1)}\right)^T$ where $\hat{\beta}_{lj} = C_j'\hat{\boldsymbol{\theta}}_l = log\left(\hat{\lambda}_{lj}/\hat{\lambda}_{lK}\right)$, for $j = 1, \ldots, (K-1)$, with variance $\hat{V}\left(\hat{\boldsymbol{\beta}}_l\right) = \hat{\Omega}_l$ that is of the same form as (29) as a function of the numbers of events $\{D_{lj}\}$, and where Ω_l is the like matrix defined in terms of the expected numbers of subjects with the event $\{E(D_{lj})\}$. Then, the joint minimum variance or weighted least squares estimate of the vector of adjusted log hazard ratios over strata, and the corresponding covariance matrix, are obtained as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left[\sum_{l=1}^{S} \hat{\Omega}_l^{-1}\right]^{-1} \left[\sum_{l=1}^{S} \hat{\Omega}_l^{-1}\hat{\boldsymbol{\beta}}_l\right] \\ \hat{V}\left(\hat{\boldsymbol{\beta}}\right) &= \left[\sum_{l=1}^{S} \hat{\Omega}_l^{-1}\right]^{-1}. \end{aligned} \quad (31)$$

Such equations are described by Lachin [17], among others, in the setting of a stratified multivariate analysis. Then the stratified-adjusted contrast T^2-like test of the hypothesis that the stratified-adjusted hazard ratios are equal among groups, i.e. H_0: $\beta_1 = \ldots = \beta_{(K-1)}$, is constructed as in (27) that equals

$$\begin{aligned} X^2 &= \hat{\boldsymbol{\beta}}' \left[\hat{V}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1} \hat{\boldsymbol{\beta}} \\ &= \left[\sum_{l=1}^{S} \hat{\Omega}_l^{-1} \hat{\boldsymbol{\beta}}_l\right]' \left[\sum_{l=1}^{S} \hat{\Omega}_l^{-1}\right]^{-1} \left[\sum_{l=1}^{S} \hat{\Omega}_l^{-1} \hat{\boldsymbol{\beta}}_l\right]. \end{aligned} \quad (32)$$

Sample size and power are then evaluated using the non-centrality parameter for this test

$$\begin{aligned} \psi^2 &= \boldsymbol{\beta}' \left[V\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1} \boldsymbol{\beta} \\ &= \left[\sum_{l=1}^{S} \Omega_l^{-1} \boldsymbol{\beta}_l\right]' \left[\sum_{l=1}^{S} \Omega_l^{-1}\right]^{-1} \left[\sum_{l=1}^{S} \Omega_l^{-1} \boldsymbol{\beta}_l\right] \end{aligned} \quad (33)$$

where β_l is the vector of assumed log hazard ratios within the lth stratum and Ω_l is of the same form as (29) using $E(D_{lj}) = N\omega_l\zeta_{lj}\pi_{lj}$.

For computation of sample size, denote the cell event probability as $\nu_{lj} = \omega_l\zeta_{lj}\pi_{lj}$ such that $E(D_{lj}) = N\nu_{lj}$. Then $\Omega_l = \Upsilon_l/N$ where Υ_l is patterned as in (29) with jth diagonal element $\nu_{lj}^{-1}+\nu_{lK}^{-1}$ and off diagonal elements ν_{lK}^{-1}. Then it follows that the non-centrality factor is

$$\phi^2 = \left[\sum_{l=1}^{S} \Upsilon_l^{-1} \boldsymbol{\beta}_l\right]' \left[\sum_{l=1l}^{S} \Upsilon^{-1}\right]^{-1} \left[\sum_{l=1}^{S} \Upsilon_l^{-1} \boldsymbol{\beta}_l\right]. \quad (34)$$

For a given set of parameters and total sample size N, power is readily evaluated from (33), whereas the required sample size N is obtained using (34).

7 Test of Group by Stratum Interaction

Alternately, for K groups and S strata it may be desired to conduct a test of homogeneity of the treatment group differences among strata, or a test of a group by stratum interaction, on $(K-1)(S-1)$ df. While such a test is conveniently conducted using the Cox PH model, a large sample test can readily be obtained from the above construction based on the multivariate distribution of the hazard rate estimates.

Within the lth stratum, we have a vector of log hazard ratios $\hat{\boldsymbol{\beta}}_l$ for each group versus the reference (Kth) with covariance matrix provided by $\hat{\Omega}_l$ as in (29). These also yield average estimates $\hat{\boldsymbol{\beta}}$ of the log hazard ratios over the S strata with estimated covariance matrix $\hat{V}\left(\hat{\boldsymbol{\beta}}\right)$ as shown in (31). Then the test of homogeneity, or no group by stratum interaction, is provided by

$$X^2 = \sum_{l=1}^{S} \left(\hat{\boldsymbol{\beta}}_l - \hat{\boldsymbol{\beta}}\right)' \hat{\Omega}_l^{-1} \left(\hat{\boldsymbol{\beta}}_l - \hat{\boldsymbol{\beta}}\right)' \quad (35)$$

with non-centrality factor

$$\phi^2 = \sum_{l=1}^{S} (\boldsymbol{\beta}_l - \boldsymbol{\beta})' \Upsilon_l^{-1} (\boldsymbol{\beta}_l - \boldsymbol{\beta})' \quad (36)$$

where $\Omega_l = N\Upsilon_l$.

An equivalent test can be obtained from a contrast-based test among the log hazard ratios. Let $\underline{\hat{\boldsymbol{\beta}}} = \left(\hat{\boldsymbol{\beta}}_1^T \cdots \hat{\boldsymbol{\beta}}_S^T\right)^T$ denote the column vector of $S(K-1)$ log hazard ratios with estimated covariance matrix $\underline{\hat{\Omega}} = \text{blockdiag}\left(\hat{\Omega}_1 \cdots \hat{\Omega}_S\right)$. Then construct the contrast matrix **C**' such that the *j*th row of the *l*th block consists of a contrast among the *j*th log hazard ratio in that block versus the *j*th log hazard ratio in the last (reference) block. For example, if $(K-1) = 2$ and $S = 3$ then

$$\mathbf{C}' = \begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix} \quad (37)$$

Then the test of homogeneity among the strata, or of no interaction, is provided by

$$X^2 = \left(\mathbf{C}'\underline{\hat{\boldsymbol{\beta}}}\right)' \left(\mathbf{C}'\underline{\hat{\Omega}}\mathbf{C}\right)^{-1} \left(\mathbf{C}'\underline{\hat{\boldsymbol{\beta}}}\right) \quad (38)$$

that is algebraically equivalent to (35). Again partitioning $\Omega_l = \Upsilon_l/N$, such that $\underline{\Upsilon} = \text{blockdiag}\left(\Upsilon_1 \cdots \Upsilon_S\right)$, then the corresponding non-centrality factor is

$$\phi^2 = \left(\mathbf{C}'\underline{\boldsymbol{\beta}}\right)' \left(\mathbf{C}'\underline{\Upsilon}\mathbf{C}\right)^{-1} \left(\mathbf{C}'\underline{\boldsymbol{\beta}}\right) \quad (39)$$

that equals (36) for a specified vector $\underline{\beta}$.

8 Quantitative Covariate Effects

It may also be of interest to describe the effect of a quantitative covariate on the risk of the outcome event after adjusting for differences among groups or strata, and/or to assess the homogeneity of the covariate effect among groups or the interaction of group with the quantitative covariate effect. Hsieh and Lavori [18] describe the assessment of sample size for the effect of a quantitative covariate in a univariate Cox PH model. From their results, it follows that the estimated coefficient $\hat{\beta}$, the log hazard ratio per unit increase in the covariate (X), is asymptotically distributed as

$$\hat{\beta} \sim \mathcal{N}\left[\beta, \left[E(D)\sigma^2\right]^{-1}\right] \quad (40)$$

where σ^2 now denotes the variance of X and D the number of events observed in the cohort. Note that the information in the data that determines power, or the inverse of the variance, is equal to $E(D)\sigma^2$ for a given value of β. For illustration, consider two covariates with the

same value for β but with different variances. Then, for a study with D events, this means that the covariate with the *larger* variance will have greater power because it has a greater range of risk over the range of covariate values, as reflected by the value of σ^2.

Likewise, within the jth group (stratum), $\hat{\beta}_j \sim \mathcal{N}\left[\beta_j, \left[E(D_j)\,\sigma_j^2\right]^{-1}\right]$ as above with covariate variance $\left(\sigma_j^2\right)$ and D_j events in the jth group. As in the prior sections, $E(D_j) = N\zeta_j\pi_j$ is a function of the group sample fraction (ζ_j) and the event probability within that group (π_j) that in turn is a function of the event hazard rate within that group (and other quantities).

Then the minimum variance linear estimator of the common coefficient among groups is provided by

$$\hat{\beta} = \frac{\sum_{j=1}^{K} D_j \sigma_j^2 \hat{\beta}_j}{\sum_{j=1}^{K} D_j \sigma_j^2} \quad (41)$$

with variance

$$V\left(\hat{\beta}\right) = \frac{1}{\sum_{j=1}^{K} E(D_j)\,\sigma_j^2} \quad (42)$$

that is consistently estimated from the observed D_j and the estimated covariate variance $\left(\hat{\sigma}_j^2\right)$ within groups. This then provides a group- or stratified-adjusted test of the covariate effect as

$$X^2 = \frac{\hat{\beta}^2}{\hat{V}\left(\hat{\beta}\right)} \quad (43)$$

that is distributed chi-square on 1 df under H_0: $\beta = 0$. This test is valid when the true coefficients $\{\beta_j\}$ may vary among groups, although there will be loss of power as the degree of heterogeneity increases.

For a given set of coefficients $\{\beta_j\}$, expected (or realized) numbers of events $\{E(D_j)\}$ and covariate variances $\left\{\sigma_j^2\right\}$ among strata, the non-centrality parameter of the test is

$$\psi^2 = \frac{\beta^2}{V\left(\hat{\beta}\right)} = \frac{\left[\sum_{j=1}^{K} E(D_j)\,\sigma_j^2 \beta_j\right]^2}{\sum_{j=1}^{K} E(D_j)\,\sigma_j^2} \quad (44)$$

from which the power of the test can be obtained. For given $E(D_j) = N\zeta_j\pi_j$, the non-centrality factor is

$$\phi^2=\frac{\left[\sum_{j=1}^{K}\xi_j\pi_j\sigma_j^2\beta_j\right]^2}{\sum_{j=1}^{K}\xi_j\pi_j\sigma_j^2} \quad (45)$$

from which the total required sample size N can be obtained.

Alternately, it may be desired to conduct a test of the hypothesis of homogeneity of the covariate effects among groups, or H_0: $\beta_j = \beta$ for all groups ($j = 1, \ldots, K$). Cochran's test of homogeneity (no interaction) is provided by

$$X^2=\sum_{j=1}^{K}D_j\hat{\sigma}_j^2\left(\hat{\beta}_j-\hat{\beta}\right)^2 \quad (46)$$

on $K - 1$ *df*. For given sets $\{\beta_j\}$, $\{E(D_j)\}$ and $\left\{\sigma_j^2\right\}$ among strata, the non-centrality parameter of the test is

$$\psi^2=\sum_{j=1}^{K-1}E\left(D_j\right)\sigma_j^2(\beta_j-\beta)^2 \quad (47)$$

and the non-centrality factor is

$$\phi^2=\sum_{j=1}^{K-1}\xi_j\pi_j\sigma_j^2(\beta_j-\beta)^2 \quad (48)$$

where

$$\beta=\frac{\sum_{j=1}^{K}\xi_j\pi_j\sigma_j^2\beta_j}{\sum_{j=1}^{K}\xi_j\pi_j\sigma_j^2}. \quad (49)$$

The above simplifies when the variance of the covariate is the same in all groups, $\sigma_j^2=\sigma^2$ for all j.

9 Example

The Glycemia Reduction Approaches in Diabetes Effectiveness, A Cost-Effectiveness Study (GRADE) is designed to compare the effectiveness of four classes of drugs commonly used for treatment of type 2 diabetes. The primary outcome is the time to confirmed inability to maintain adequate glycemic control, which from prior studies is estimated to have a reference hazard rate of $\lambda = 0.0875$ per year in the drug group(s) with the least durable effect on glycemic control. Herein we describe sample size and power computations assuming an $R = 3$ year recruitment interval with a total duration of $T = 7$ years. To allow for a lag in recruitment it is assumed that 40% of subjects are recruited in the first half of the

recruitment period, 60% in the second, that corresponds to a recruitment shape parameter of $\gamma = -0.27$. This yields a mean recruitment time of 1.7 years and a mean potential exposure of 4.8 years assuming no losses to follow-up. Allowing for 4% losses per year ($\eta = 0.04$), from (4) the mean exposure time is reduced to 3.84 years with an event probability of 0.335 and a loss-to-follow-up probability of 0.153.

A single overall global test of the hypothesis of equality among the 4 groups will be conducted on 3 df. The simplest alternative hypothesis is that three treatments all have the reference hazard rate of 0.0875 and one treatment (say the first) is superior to the other four with a hazard ratio of 0.75 versus the others, i.e. with a hazard of 0.0656 for the first and 0.0875 for the other four groups, and the vector of hazard ratios $\beta = (0.75\ 1\ 1)^T$ with group 4 as the reference. With equal sized treatment groups ($\zeta_j = 1/4$), then the expected probabilities of the event are 0.265 in the first and 0.335 in each of the other 3 groups. These yield a weighted mean log hazard $\theta = -2.496$ corresponding to a geometric mean hazard of 0.0824 and a non-centrality factor of $\varphi^2 = 0.004338$. The non-centrality parameter value that provides 90% power for a 3 df test at the 0.05 level is $\psi^2(0.05, 0.10, 3) = 14.1715$. Substituting into (11), $N = \psi^2/\varphi^2 = 3268$ (rounded up from 3267) would be required to provide 90% power to detect the hazard ratio of 0.75 for one therapy versus the others. This yields 216 subjects expected to have the event in the first group and 274 in each of the other 3 groups.

For the case where two therapies are equally superior to the other two with a hazard ratio of 0.75, then an N of 2316 (rounded up from 2315) would provide 90% power. Thus, it is conservative to power the study to detect a single isolated superior drug with $HR = 0.75$, in which case the total sample size selected might be $N = 3300$ to provide 90% power.

However, it is also desired to conduct 6 pairwise comparisons among the 4 drug groups. Although the Hochberg closed test procedure will be employed, for the smallest nominal p-value, the adjustment is equivalent to the Bonferroni correction, i.e. a two-sided significance level of 0.05/6 being required for adjusted significance at the 0.05 level. Two group calculations with $n = 825$ per group shows that the total $N = 3300$ provides only 71% power to detect a $HR = 0.75$ between any two groups with this design. Rather, a sample size of $n = 1242$ per group is required to provide 90% power to detect a $HR = 0.75$ in a two-group comparison at the 0.05/6 level under the above assumptions, thus requiring a total sample size of 4964, rounded up to $N = 5000$ as the target enrollment. In this case, the $K - 1$ df test of homogeneity would provide 98.3% power to detect a single superior drug group with $HR = 0.75$, and 90% power to detect a single group with $HR = 0.796$.

It should be noted that another option might be to conduct 4 pairwise comparisons of each drug group versus the other 3 groups combined. With the smaller total N of 3300, such a test at the 0.05/4 level would provide 93% power to detect $HR = 0.75$. However, the 6 pairwise comparisons are preferred and thus the larger sample size of $N = 5000$ will be employed.

The study will evaluate various stratification or subgroup factors in which case a stratified-adjusted test may be conducted. To assess the effect of heterogeneity among strata, consider the case where one stratum consists of approximately 2000 subjects with a 20% lower

hazard rate of 0.0875 * (0.8) = 0.07/year and a smaller difference between groups with a hazard ratio of 0.85, and the other stratum consists of approximately 3000 subjects with the same risks assumed above. With the same parameters as above, and assuming that a single drug is superior to the others, then the stratum of 2000 subjects would provide an expected 122 events in the first group and 140 events in the other three groups, and the stratum of 3000 subjects would provide 199 and 251 events, respectively. The vector of stratified adjusted log hazard ratios is $\beta = (-0.240713\ 0\ 0)^T$ with the first element corresponding to a hazard ratio of 0.786 for the first group versus the reference (i.e. all others). The corresponding covariance matrix $V\left(\hat{\boldsymbol{\beta}}\right)$ has diagonal elements 0.005678 for the first log odds ratio, 0.005114 for the next 2 diagonal elements, and off-diagonal elements 0.002557. The resulting non-centrality parameter is 14.58 that yields power of 90.1%. Thus, the presence of a mild group by stratum interaction (or heterogeneity) leads to some dilution of power for the 4 group test, but at an acceptable level. However, a test of no interaction or homogeneity would have a low power of only 10% to detect a difference in hazard ratios of 0.75 versus 0.85 between these two strata.

Subgroup analyses will also be performed to assess the treatment group differences between segments of the population, such as males and females, with a test of a treatment by subgroup interaction. Again assume an overall hazard rate of 0.0875 and losses at 4% per year with the first group being superior to the rest with an overall hazard ratio of 0.75 in the full cohort for one group versus the rest. Within one subgroup, assume that the hazard ratio is 25% less, i.e. a hazard ratio of 0.75 × 0.75 = 0.563, whereas in the other subgroup it is 25% greater, i.e. 0.75 × 1.25 = 0.938. For equally sized subgroups with $n = 2500$ each, the test of homogeneity (no interaction) provides 93.9% power. For a factor with three subgroups, each with sample size 1666, the study would provide 68.9% power to detect hazard ratios of 0.563, 0.75 and 0.938.

Analyses may also be conducted involving a quantitative covariate. As for a qualitative covariate (the *S* strata), one analysis could assess the association of the covariate with the outcome adjusted for treatment group, and another could assess homogeneity of the covariate effect among strata (or a group by stratum interaction).

For $N = 5000$, or 1250 for each of 4 groups, under the above assumptions, approximately 394 events are expected within each group (1576 total). From recent studies, such as of biomarkers in relation to cardiovascular disease, a hazard ratio of 1.4 per standard deviation change (HR_{SD}) in the covariate is desirable to detect. For a standard deviation σ of the covariate, the hazard ratio per unit change in the covariate is $HR = HR_{SD}^{1/\sigma}$. Then the coefficient is $\beta = \log(HR_{SD})/\sigma$. For $\sigma = 10$ the Hsieh-Lavori expression with 1576 events yields virtually 100% power to detect a $HR_{SD} = 1.4$, and 97% power to detect a smaller $HR_{SD} = 1.0$. It is also desired to assess the power of the study to detect heterogeneity of a quantitative covariate effect among groups, such as with coefficient values of 1.25, 1.35, 1.45 and 1.55, the weighted average $\beta = 0.0333$ corresponding to a $HR_{SD} = 1.396$. This yields a non-centrality parameter value of $\psi^2 = 10.3$ on 3 for a test of homogeneity, that yields 76.7% power to detect these small differences in the hazard ratios.

10 Discussion

Makuch and Simon [9] describe the assessment of power of the *K* group logrank test under an exponential model when the numbers of subjects with the event are known (specified), and they provide an equation to compute the average number of such subjects with the event, i.e. assuming that all groups have the same numbers of subjects with the event. Given an assumption about the total exposure time within each group (time to event or censoring), the required sample size can then be obtained. Herein, we take a more general approach by first describing design assumptions including the pattern of recruitment, rate of losses-to-follow-up, and possibly levels of stratification, as in Lachin and Foulkes [1] for the two group case. For a given sample size this then provides the probability in each group that a subject will have the event, which can then be used to determine the required sample size, or to determine power for a given sample size.

The initial approach herein is to describe the non-centrality parameter for a test of homogeneity of the *K* group hazard rates, that is also shown to apply to a T^2 or Wald-type contrast test. These expressions employ the variance of the test statistic evaluated under the alternative, i.e. using the set of specified hazard rates that are desired to be detected. We then derive the non-centrality parameter for the score test in the Cox PH model resulting in an expression identical to that of Ahnn and Anderson [11] when there are equal sized groups and a common pattern of censoring. These expressions employ the variance of the test statistic evaluated under the null hypothesis, i.e. assuming a common probability of the event among the groups. This latter approach based on (23) will provide a smaller number of required events and a smaller total sample size than that using (9). For the above example with 4 groups and a single group superior with a hazard ratio of 0.75, the latter expression yields $N = 3268$ and $D = 1037$ whereas (23) yields an expected total number of events of $D = 913$, for which a smaller sample size $N = 2876$ would be required. For the comparison of 2 groups, Lachin [4] showed that the comparable expression using the alternative hypothesis variance as in (9) was in general more conservative in that it always provides a larger *N* and larger required number of events than the expression based on the null hypothesis variance as in (23). On this basis the exponential model based expression might be preferred.

Generalizations then provide the assessment of sample size or power for a stratified-adjusted *K*-group comparison and for a test of homogeneity or group by stratum interaction, as would be appropriate for a “subgroup” analysis in which the treatment group differences are compared among strata. Likewise, sample size and power are described for a stratified-adjusted analysis of a quantitative covariate effect, and a test of homogeneity of a quantitative covariate effect among strata.

While explicit expressions for a stratified analysis are provided, in many cases a simple approximate computation may suffice. For the above example, with a common recruitment shape parameter in each stratum and a common hazard rate for losses-to-follow-up, the average hazard rate for the event is $(3/5)(0.0875) + (2/5)(0.07) = 0.0805$ and the average log hazard ratio for one group versus the others is $(3/5)\,[\ln(0.75)] + (2/5)\,[\ln(0.85)] = -0.238$ corresponding to an average hazard ratio of 0.788 for one group versus the others. Then a non-stratified computation using (9) with $N = 5000$ yields power of 90.2%, close to the

90.1% provided by the precise stratified computation. Thus, the principal application of the stratified assessment would be to the case where the strata have different patterns of recruitment and/or different periods of enrollment or follow-up duration and different patterns of losses-to-follow-up. An example of this type is described by Lachin and Foulkes [1] to which the above computations would apply for a K group trial.

The Mantel-Logrank test is a member of the family of linear rank tests for survival data described by Anderson, Borgan, Gill and Keiding [19] that includes the Peto-Prentice modified Wilcoxon test that is optimal under a proportional survival odds model. Jung and Hui [20] describe the non-central distribution of this family of tests from which the power of a particular test can be obtained. Their method allows for a period of uniform recruitment and follow-up, and losses-to follow-up, but it requires numerical integration of stochastic integrals to construct the non-centrality parameter. Conversely, the methods herein are quite simple to apply.

In all cases, the sample size is obtained by solving for the total N that yields a desired number of events. In cases where the number of events is known, or pre-specified in advance, power can be assessed by simply substituting the event numbers into the above expressions, as in (9).

Central to the application of these methods is the precise specification of the log hazard rates within each of the treatment groups (and possibly strata) worthy of detection. In general, from (9) it is clear that the magnitude of the non-centrality parameter, and thus power, depends explicitly on the weighted sum of squares (SS) among the specified log hazards, weighted by the expected number of events within each group. As employed by Makuch and Simon [9], this parameter depends approximately on the unweighted sum of squares that is easily evaluated for a specified set of hazard rates. In this case, well-known results for a balanced one-way ANOVA F-test of equality of K group means will apply approximately. Consider the set of K ordered means (or log hazard rates) with minimum mean $\theta_{(1)}$ and maximum $\theta_{(K)}$. The maximum power (sum of squares) occurs when the ordered means for half the groups (or for $K/2 \pm 0.5$ if K is odd) equal $\theta_{(1)}$ and the other half (or $K/2 \pm 0.5$) equal $\theta_{(K)}$. Conversely, the least power occurs when $\theta_{(j)} = (\theta_{(1)} + \theta_{(K)})/2$ for $1 < j < K$. The test also has poor power when the hazards in $K - 1$ groups are equal and that in the Kth group is different, the so-called case of a single isolated superiority (or inferiority). For example, for a set of $K = 5$ means with $\theta_{(1)} = 0$ and $\theta_{(K)} = 4$, the maximum SS is 19.2 for ordered mean values of (0, 0, 0, 4, 4) or (0, 0, 4, 4, 4). The SS is 12.8 for the isolated superiority with means (0, 0, 0, 0, 4) or (0, 4, 4, 4, 4); and is a minimum of 8.0 for means of (0, 2, 2, 2, 4). Thus, in the balanced case ($\zeta_j = 1/K$), the total number of events required using the expression under the null (25) for the minimum SS case is 19.2/8 = 2.4-fold higher that for the maximum SS case.

Makuch and Simon also proposed that the global test in (8) could be employed with Fisher's Least Significant Difference (LSD) method to guarantee the experiment-wide type 1 error probability at level α for the set of $K(K - 1)/2$ pairwise tests. However, from the closed testing principle [21], this is true only for $K = 3$ groups. For an illustration, see Chi [24]. For example, with $K = 4$ groups, if the 4 group test is significant at level α, one can then test the

4 separate 3 group differences at level α. A given pairwise comparison, e.g. group 1 versus 2, is a component of the null hypothesis for 2 of the 4 such 3 group tests, specifically H_0: $\theta_1 = \theta_2$ is a component of the "parent" test hypotheses H_0: $\theta_1 = \theta_2 = \theta_3$ and H_0: $\theta_1 = \theta_2 = \theta_4$. If both of these are significant at level α, then the component pairwise test of H_0: $\theta_1 = \theta_2$ can also be tested at level α. If the two parent 3 group null hypotheses are not rejected at level α for a given pairwise hypothesis, then that pairwise test is declared non-significant.

Finally, the methods herein assume that the exponential model or the proportional hazards model apply. If indeed they do not, then the required sample size will be underestimated, and the study power overestimated. For the simple two group design, Lakatos [23] describes a piecewise interval approach to the power of the logrank test when the hazard rate and/or hazard ratio may vary over intervals of time. Ahnn and Anderson [24] describe a generalization of this approach to the *K*-group logrank test for specified hazard rates, numbers of events and numbers within each group at risk within each interval of time, quantities that are not easily obtained in a complex design with staggered non-uniform patient entry, varying hazards and losses to follow-up.

Programs from the author for the computations herein will be available from www.bsc.gwu.edu under the link to programs available for download.

Acknowledgments

This work was partially supported by cooperative agreements from the National Institute of Diabetes, Digestive and Kidney Diseases for the Glycemia Reduction Approaches in Diabetes Effectiveness, A Cost-Effectiveness Study (GRADE) and for the *Diabetes Prevention Program Outcomes Study (DPPOS)*.

REFERENCES

1. Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for non-uniform patient entry, losses to follow-up, non-compliance and stratification. Biometrics. 1986; 42:507–519. [PubMed: 3567285]
2. George SL, Desu MM. Planning the size and duration of a clinical trial studying the time to some critical event. J. Chronic Dis. 1974; 27:15–29. [PubMed: 4592596]
3. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. Biometrics. 1983; 39:499–503. [PubMed: 6354290]
4. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control. Clin. Trials. 1981; 2:93–113. [PubMed: 7273794]
5. Rubenstein LV, Gail MH, Santner TJ. Planning the duration of a comparative clinical trial with losses to follow-up and a period of continued observation. J. Chronic Dis. 1981; 34:469–479. [PubMed: 7276137]
6. Schoenfeld DA, Richter JR. Nomograms for calculating the number of patients needed for a clinical trial with survival as the endpoint. Biometrics. 1982; 38:163–170. [PubMed: 7082758]
7. Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. Stat. Med. 1982; 1:121–129. [PubMed: 7187087]
8. Schoenfeld DA. The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika. 1981; 68:316–319.
9. Makuch RW, Simon RM. Sample size requirements for comparing time-to-failure among k treatment groups. J. Chronic Dis. 1982; 35:861–867. [PubMed: 7142364]
10. Liu PY, Dahlberg S. Design and analysis of multiarm clinical trials with survival endpoints. Control Clin. Trials. 1995; 16:119–130. [PubMed: 7789135]

11. Ahnn S, Anderson SJ. Sample size determination for comparing more than two survival distributions. Stat. Med. 1995; 14:2273–2282. [PubMed: 8552903]
12. Tarone RE, Ware J. On distribution-free tests for equality of survival distributions. Biometrika. 1977; 64:156–160.
13. Halabi S, Singh B. Sample size determination for comparing several survival curves with unequal allocations. Stat. Med. 2004; 23:1793–1815. [PubMed: 15160409]
14. Lachin, JM. Biostatistical Methods: The Assessment of Relative Risks. 2nd Edition. John Wiley & Sons; New York: 2011.
15. Cochran WG. The combination of estimates from different experiments. Biometrics. 1954; 10:101–129.
16. Anderson, TW. An Introduction to Multivariate Analysis. 2nd edition. John Wiley & Sons; New York: 1984.
17. Lachin JM. Some large sample distribution-free estimators and tests for multivariate partially incomplete data from two populations. Stat. Med. 1992; 11:1151–1170. [PubMed: 1509217]
18. Hsieh FY, Lavori PW. Sample-Size Calculations for the Cox Proportional Hazards Regression Model with Nonbinary Covariates. Controlled Clinical Trials. 2000; 21:552–560. [PubMed: 11146149]
19. Andersen PK, Borgan O, Gill RD, Keiding N. Linear nonparametric tests for comparison of counting processes, with applications to censored survival data. Int. Statist. Rev. 1982; 50:219–258.
20. Jung S-H, Hui S. Sample size calculations for rank tests comparing *K* survival distributions. Lifetime Data Analysis. 2002; 8:361–373. [PubMed: 12471945]
21. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. Biometrika. 1976; 63:655–660.
22. Chi GYH. Multiple testings: Multiple comparisons and multiple endpoints. Drug. Inf. J. 1998; 32:1347S–1362S.
23. Lakatos E. Sample sizes based on the log-rank statistic in complex clinical trials. Biometrics. 1988; 44:229–241. [PubMed: 3358991]
24. Ahnn S, Anderson SJ. Sample size determination in complex clinical trials comparing more than two groups for survival endpoints. Stat. Med. 1998; 17:2525–2534. [PubMed: 9819843]