Original Article

Power Calculation for Likelihood Ratio Tests in the Nested Case-control Designs with a Randomly Sampled Control Per Case

Shizue Izumi^{*1} and Yoshinori Fujii^{*2}

^{*1}Faculty of Engineering, Oita University

^{*2}Faculty of Education and Culture, University of Miyazaki

e-mail: shizue@oita-u.ac.jp

This paper presents an asymptotic method of power calculations for likelihood ratio tests in the nested case-control designs with a randomly sampled control per case. It is an extension of the approach described in Self et al. (1992) for proportional hazards models. Our approach here focuses on a simple scenario with 1 : 1 case-control ratio, simple random sampling design, and two independent dichotomous covariates with no interaction effects. The approximation of the noncentrality of the noncentral chi-square distribution for the likelihood ratio statistic is provided. Simulation studies are conducted to examine the accuracy for several parameter values and data configurations. Overall the results suggest that estimates of power using our proposed method are consistent with those of actual power from Monte Carlo simulation. Therefore, the proposed approach can be practically useful in assessing the statistical power for the simple nested case-control design.

Key words: Sample size, Risk set sampling, Proportional hazards models, Failure time data, Censoring, Noncentral chi-square.

1. Introduction

Large-scale cohort studies have been conducted actively to ascertain the molecular epidemiologic disease mechanism with fairly expensive biologic markers. It is crucial what design is chosen under the given conditions (Rundle et al., 2005). A nested case-control design is relatively efficient in estimating the effects of the target variable (Langholz and Goldstein, 1996), which is an attractive feature to many researchers. Controls are sampled from the risk sets at each time a case occurs. The design of sampling controls includes random sampling, matching (stratifying), and counter-matching. When analyzing the nested-case control data, a choice of the design affects the formulation of a partial likelihood function in the proportional hazards model. Because of these unique features, a thorough consideration should be made on the design, sample size, and statistical power to perform a cost effective research and minimize the

number of limited biological resources.

Statistical efficiency, sample size, and statistical power have not been well studied in a nested case-control data analysis. A preliminary simulation study was performed to examine efficiency of the parameter of interests in the regression model in a nested case-control data analysis with counter-matched cases and controls (Cologne et al., 2004). The power calculation of a score test based on a conditional likelihood function was also proposed for a nested case-control data analysis (Sinha and Mukherjee, 2006; Lachin, 2008). Controls are matched to cases with respect to confounders in their methods, and the conditional logistic regression model includes a single exposure variable that can be quantitative or qualitative. On the other hand, an asymptotic method proposed by Self et al. (1992) is frequently used to estimate the power for likelihood ratio tests in the generalized linear models with multiple covariates. However the method is not yet extended to estimate the power for a nested case-control data analysis.

In this paper we present an asymptotic method of power calculations for likelihood ratio tests in the nested case-control designs when a control per case is randomly sampled and two independent dichotomous covariates with no interaction effects are observed. In Section 2.1, an asymptotic method is overviewed in generalized linear models with two covariates. In Section 2.2, a proportional hazards model is described for the simple nested case-control sample. In Section 2.3, the approach described Self et al. (1992) is extended for the proportional hazards models. In Section 3, simulation studies are conducted to examine the accuracy for several parameter values and data configurations. Finally the implementation of the proposed approach and possible future studies are discussed.

2. Asymptotic method of power calculations

2.1 Generalized linear model with two covariates

According to the notation used in McCullagh and Nelder (1989), the class of generalized linear models is specified by assuming independent scalar response variables, $Y_1, \dots, Y_N$ to have probability density functions of the form

$$\exp\{[Y_i\theta_i - b(\theta_i)]/a_i(\phi) + c(Y_i, \phi)\}, \quad (1)$$

for some specific functions $a_i(\cdot)$, $b(\cdot)$, and $c(\cdot)$. When ϕ is known, this is an exponential family model with canonical parameter θ_i . The expected value of $E(Y_i) = \mu_i$ is related to the parameter θ_i by the equation $\mu_i = b'(\theta_i)$, where b' denotes the first derivative of b . The link function g relates the linear predictors η_i to the mean response $\eta_i = g(\mu_i)$. The linear predictors can be written as

$$\eta_i = X_i\psi + Z_i\lambda,$$

where X_i and Z_i are covariates with no interaction effects, and ψ and λ represent the corresponding unknown regression coefficients. The scale parameter ϕ is assumed known. Our interest is to

test the null hypothesis of $H_0: \psi = 0$ against the alternative hypothesis $H_a: \psi \neq 0$, while treating λ as a nuisance parameter.

Assume (y_i, x_i, z_i) is a random sample from the joint distribution of (Y, X, Z) with p.d.f. $f(Y, X, Z) = f(Y | X, Z) \cdot f(X, Z)$, where $f(Y | X, Z)$ has the form defined in (1) and $f(X, Z)$ is the joint p.d.f. for X and Z . The form of $f(X, Z)$ is assumed to depend on none of the unknown parameters ψ and λ . The joint likelihood function of these N subjects is

$$L(\psi, \lambda) = \prod_{i=1}^N f(y_i, x_i, z_i) = \prod_{i=1}^N f(y_i | x_i, z_i) \cdot f(x_i, z_i).$$

It follows that the likelihood ratio statistic is $2\{l(\hat{\psi}, \hat{\lambda}) - l(0, \hat{\lambda}_0)\}$, where $l(\psi, \lambda)$ is the log-likelihood function based on $L(\psi, \lambda)$ and $(\hat{\psi}, \hat{\lambda})$ and $(0, \hat{\lambda}_0)$ are the maximum likelihood estimators of (ψ, λ) under the alternative and the null models, respectively. The actual likelihood ratio test statistic is referred to its asymptotic distribution under the null hypothesis, which is a central chi-square distribution with one ($= \text{rank}(X)$) d.f.

In order to perform power analysis, we need to derive distribution of the likelihood ratio statistic under the alternative hypothesis. As in Self et al. (1992), the likelihood ratio statistic is decomposed into three terms,

$$2\{l(\hat{\psi}, \hat{\lambda}) - l(0, \hat{\lambda}_0)\} = 2\{l(\hat{\psi}, \hat{\lambda}) - l(\psi, \lambda)\} - 2\{l(0, \hat{\lambda}_0) - l(0, \lambda_0)\} + 2\{l(\psi, \lambda) - l(0, \lambda_0)\}, \quad (2)$$

where λ_0 is the limiting value of $\hat{\lambda}_0$ as described in Self and Mauritsen (1988). The first term in (2) approximately follows a chi-square distribution with two d.f. ($= \text{rank}(X) + \text{rank}(Z) = 1 + 1$), and the expected value is approximately two. On the contrary the expected value of the second term approximately equals $\text{tr}(M)$, where

$$M = E \left[-\frac{\partial^2}{\partial \lambda^2} l(\psi, \lambda) |_{(0, \lambda_0)} \right]^{-1} E \left[\frac{\partial}{\partial \lambda} l(\psi, \lambda) |_{(0, \lambda_0)}^2 \right].$$

The third term does not involve any maximum likelihood estimators of (ψ, λ) . $E_{(X, Z)}[\cdot]$ denotes the expectation taken with respect to the joint distribution of (X, Z) , which is not dependent on the value of (ψ, λ) . The expected value of the third term can be written as $N\Delta^*$, where

$$\Delta^* = E_{(X, Z)} \left[2 \left\{ \frac{b'(\theta_i)(\theta_i - \theta_i^*) - (b(\theta_i) - b(\theta_i^*))}{a_i(\phi)} \right\} \right].$$

The respective expected value approximations of (2) derived above is $(1 + 1) - \text{tr}(M) + N\Delta^*$. Under the null hypothesis $H_0: \psi = 0$, $\text{tr}(M) = \text{rank}(Z) = 1$ and $\Delta^* = 0$ so that the overall expectation is one, which conforms with the expectation of chi-square distribution with one d.f. Hence, for given parameter values (ψ, λ) and covariate distribution $f(X, Z)$, the actual likelihood ratio statistic of sample size N under the alternative hypothesis approximately follows a noncentral chi-square distribution with one d.f. and noncentrality $1 - \text{tr}(M) + N\Delta^*$.

2.2 Proportional hazards model for the simple nested case-control sample analysis

A nested case-control data set consists of case-control sets in which all cases are extracted and their controls are sampled from the full cohort data (Langholtz, 2005; Fujii et al., 2007). A unique feature of the nested case-control sample analysis is the risk set sampling designs of controls. The sampling designs are related to semiparametric estimation methods for parameters in a proportional hazards model used in the analysis of full cohort data. At each failure time when a case is observed, a risk set is defined to be all subjects under observation. So, the cohort data is organized into the risk sets.

Our approach here focuses on a simple scenario with a randomly sampled control per case, two independent dichotomous covariates, and no interaction of the covariates in regression models. For example, we may consider that the covariate Z is the genetic variation (e.g., phenotype- A , phenotype- $\bar{A}$), the covariate X is the exposure factor (e.g., treatment, placebo) which is randomly assigned regardless the genetic variation, and the outcome is an occurrence of the disease in an observational study. Let a full cohort data include K individuals of cases and $(N - K)$ individuals of non-cases with their observed time, $\tilde{T}_i$ ($i = 1, \dots, N$) as a minimum of (T_i, C_i) , the event status, τ_i , that equals one if $T_i \leq C_i$ and zero otherwise, and two independent dichotomous covariates X and Z , where T_i ($i = 1, \dots, N$) is continuous failure time, C_i ($i = 1, \dots, N$) is continuous censoring time being assumed to be independent of the failure time T_i , X is a exposure variable of interest, and Z is its potential confounding variable. The failure time of K cases is sorted with an ascending order such as $t_{(1)} < \dots < t_{(k)} < \dots < t_{(K)}$, assuming no ties. The risk set on time $t_{(k)}$, R_k is formed as $R_k = \{i \mid \tilde{T}_i \geq t_{(k)}\}$.

In a proportional hazards model it is assumed that the hazard function for the failure time of the i -th individual is of the form

$$\lim_{\Delta t \rightarrow 0+} \frac{1}{\Delta t} \Pr(T_i < t + \Delta t \mid T_i \geq t, X_i, Z_i) = h_0(t) \exp(X_i \psi + Z_i \lambda),$$

where the baseline hazard, $h_0(t)$ is left unspecified (e.g., parametric or nonparametric). As a consequence of the above model, the density of the distribution of the failure time, T_i is

$$\Pr(t \mid X_i, Z_i) = h_0(t) \exp(X_i \psi + Z_i \lambda) \exp\{-\Lambda_0(t) \exp(X_i \psi + Z_i \lambda)\},$$

where $\Lambda_0(t) = \int_0^t h_0(s) ds$ is the cumulative baseline hazard.

The partial likelihood for nested case-control data is based on the probability that a subject is a case given the case-control set, which leads to a likelihood of the form

$$L(\psi, \lambda) = \prod_{k=1}^K \frac{h_0(t_{(k)}) \exp(X_k \psi + Z_k \lambda) \omega_k}{\sum_{j \in \tilde{R}_k} h_0(t_{(k)}) \exp(X_j \psi + Z_j \lambda) \omega_j} = \prod_{k=1}^K \frac{\exp(X_k \psi + Z_k \lambda) \omega_k}{\sum_{j \in \tilde{R}_k} \exp(X_j \psi + Z_j \lambda) \omega_j},$$

where $\tilde{R}_k$ ($\subset R_k$) is the case-control set and ω_j is the probability of picking a particular case-control set if j is the case. Here ω_j is an inverse of (the number of subjects in the risk set R_k

minus one). As shown in the above formula, the likelihood function for nested case-control data differs from that used in generalized linear models. In the next section, we derive distribution of the likelihood ratio statistic under the alternative hypothesis, $H_a: \psi \neq 0$ for the simple nested case-control sample analysis.

2.3 Distribution of the likelihood ratio statistic under $H_a: \psi \neq 0$ for case-control sets

We propose an asymptotic method of power calculations for likelihood ratio tests in the simple nested case-control sample analysis. Our underlying idea is analogous to that of Self et al. (1992), and the approach extends it to accommodate the failure time models. Using the same decomposition of the likelihood ratio statistic as that of Self et al. (1992), the expected value approximations of (2) is $(1 + 1) - tr(M) + \Delta$, where $\Delta = 2E[l(\psi, \lambda) - l(0, \lambda_0)]$ which does not involve any maximum likelihood estimators of (ψ, λ) . Thus the actual likelihood ratio statistic under the alternative hypothesis $H_a: \psi \neq 0$ approximately follows a noncentral chi-square distribution with one d.f. and noncentrality $1 - tr(M) + \Delta$. The values of λ_0 and $tr(M)$ are obtained in the rest of this section.

Given $X_{k,l}$ and $Z_{k,l}$ ($l = 0$ for the case, $l = 1$ for the control) in the risk set R_k , the log-likelihood is written as

$$l(\psi, \lambda) = \sum_{k=1}^K [(X_{k,0}\psi + Z_{k,0}\lambda) - \log\{\exp(X_{k,0}\psi + Z_{k,0}\lambda) + \exp(X_{k,1}\psi + Z_{k,1}\lambda)\}].$$

Then partial derivatives of the log-likelihood function with respect to the parameter λ are written as

$$l_\lambda(\psi, \lambda) = \frac{\partial l(\psi, \lambda)}{\partial \lambda} = \sum_{k=1}^K [Z_{k,0} - \bar{Z}_{k,0}] \text{ and } l_{\lambda\lambda}(\psi, \lambda) = \frac{\partial^2 l(\psi, \lambda)}{\partial \lambda^2} = \sum_{k=1}^K [\bar{Z}_{k,0}^2 - \bar{Z}_{k,0}^2],$$

where

$$\bar{Z}_{k,0} = \frac{Z_{k,0} \exp(X_{k,0}\psi + Z_{k,0}\lambda) + Z_{k,1} \exp(X_{k,1}\psi + Z_{k,1}\lambda)}{\exp(X_{k,0}\psi + Z_{k,0}\lambda) + \exp(X_{k,1}\psi + Z_{k,1}\lambda)} \quad \text{and}$$

$$\bar{Z}_{k,0}^2 = \frac{Z_{k,0}^2 \exp(X_{k,0}\psi + Z_{k,0}\lambda) + Z_{k,1}^2 \exp(X_{k,1}\psi + Z_{k,1}\lambda)}{\exp(X_{k,0}\psi + Z_{k,0}\lambda) + \exp(X_{k,1}\psi + Z_{k,1}\lambda)}.$$

Assuming that the log-likelihood function is conditional independent at each failure time, it is shown that

$$E[l_\lambda(\psi, \lambda)] = KE[Z_{k,0} - \bar{Z}_{k,0}], \quad E[l_{\lambda\lambda}(\psi, \lambda)] = KE[\bar{Z}_{k,0}^2 - \bar{Z}_{k,0}^2].$$

It is assumed that the covariate distribution of (X, Z) at the risk set R_k does not change at each failure time so that it is expressed as $p_{i,j} = f(X, Z) = \Pr(X = i, Z = j) = \Pr(X = i) \Pr(Z = j)$, where $i, j = 0, 1$ when the covariates X and Z are assumed to be independent. The parameters used in the log-likelihood function is called (ψ_*, λ_*) to distinguish them with the parameters used

in the calculation of $E[l_\lambda(\psi, \lambda)]$. Let $(0, \lambda_0)$ be the value of (ψ_*, λ_*) under the null hypothesis, $H_0: \psi = 0$.

First we compute the value of $E[l_\lambda(\psi, \lambda)]$ to a solution of $E[l_\lambda(\psi, \lambda) |_{(0, \lambda_0)}] = 0$. The distribution of $(X_{k,0}, Z_{k,0})$ is written as

$$\Pr(X_{k,0} = i, Z_{k,0} = j) = \frac{p_{i,j} \exp(\psi i + \lambda j)}{\sum_{m,n=0}^1 p_{m,n} \exp(\psi m + \lambda n)}.$$

So

$$\begin{aligned} E[Z_{k,0}] &= \Pr(Z_{k,0} = 1) = \frac{\exp(\lambda) \{p_{0,1} + p_{1,1} \exp(\psi)\}}{\sum_{m,n=0}^1 p_{m,n} \exp(\psi m + \lambda n)}, \text{ and} \\ E[\bar{Z}_{k,0}] &= \sum_{i,j,m,n=0}^1 \Pr(X_{k,0} = i, Z_{k,0} = j) \Pr(X_{k,1} = m, Z_{k,1} = n) \\ &\quad \cdot \frac{j \exp(i\psi_* + j\lambda_*) + n \exp(m\psi_* + n\lambda_*)}{\exp(i\psi_* + j\lambda_*) + \exp(m\psi_* + n\lambda_*)} \\ &= \sum_{i,j,m,n=0}^1 \frac{p_{i,j} \exp(i\psi + j\lambda)}{\sum_{m,n=0}^1 p_{m,n} \exp(m\psi + n\lambda)} p_{m,n} \frac{j \exp(i\psi_* + j\lambda_*) + n \exp(m\psi_* + n\lambda_*)}{\exp(i\psi_* + j\lambda_*) + \exp(m\psi_* + n\lambda_*)} \\ &= \frac{1}{\sum_{m,n=0}^1 p_{m,n} \exp(m\psi + n\lambda)} \left\{ \exp(\lambda) (p_{0,1} + p_{1,1} \exp(\psi)) \left(p_{+,1} + p_{+,0} \frac{\exp(\lambda_0)}{\exp(\lambda_0) + 1} \right) \right. \\ &\quad \left. + \frac{p_{+,1} \exp(\lambda_0)}{\exp(\lambda_0) + 1} (p_{0,0} + p_{1,0} \exp(\psi)) \right\} \quad (\because (\psi_*, \lambda_*) = (0, \lambda_0)) \\ &= E(Z_{k,0}) \left(p_{+,1} + p_{+,0} \frac{\exp(\lambda_0)}{\exp(\lambda_0) + 1} \right) + (1 - E(Z_{k,0})) \frac{p_{+,1} \exp(\lambda_0)}{\exp(\lambda_0) + 1}. \\ \therefore E[(Z_{k,0} - \bar{Z}_{k,0})] &= E(Z_{k,0}) \left\{ 1 - \left(p_{+,1} + p_{+,0} \frac{\exp(\lambda_0)}{\exp(\lambda_0) + 1} \right) \right\} - (1 - E(Z_{k,0})) \frac{p_{+,1} \exp(\lambda_0)}{\exp(\lambda_0) + 1} \\ &= E(Z_{k,0}) \frac{p_{+,0}}{\exp(\lambda_0) + 1} - (1 - E(Z_{k,0})) \frac{p_{+,1} \exp(\lambda_0)}{\exp(\lambda_0) + 1} \\ &= \frac{1}{\exp(\lambda_0) + 1} \{ E(Z_{k,0}) p_{+,0} - (1 - E(Z_{k,0})) p_{+,1} \exp(\lambda_0) \}. \end{aligned}$$

Therefore $E[l_\lambda(\psi, \lambda) |_{(0, \lambda_0)}] = \frac{K}{\exp(\lambda_0) + 1} \{ E(Z_{k,0}) p_{+,0} - (1 - E(Z_{k,0})) p_{+,1} \exp(\lambda_0) \}$. And a solution, λ_0 of $E[l_\lambda(\psi, \lambda) |_{(0, \lambda_0)}] = 0$ is obtained as

$$\lambda_0 = \log [E(Z_{k,0}) p_{+,0} \{ (1 - E(Z_{k,0})) p_{+,1} \}^{-1}]. \quad (3)$$

Next we compute the values of $E[l_\lambda^2(\psi, \lambda) |_{(0, \lambda_0)}]$ and $E[-l_{\lambda\lambda}(\psi, \lambda) |_{(0, \lambda_0)}]$.

$$\begin{aligned} E[l_\lambda^2(\psi, \lambda) |_{(0, \lambda_0)}] &= KE[(Z_{k,0} - \bar{Z}_{k,0})^2] = K \{ E(Z_{k,0}^2) - 2E(Z_{k,0} \bar{Z}_{k,0}) + E(\bar{Z}_{k,0}^2) \} \\ &= K \{ E(Z_{k,0}) - 2E(Z_{k,0} \bar{Z}_{k,0}) + E(\bar{Z}_{k,0}^2) \} \quad (\because Z_{k,0}^2 = Z_{k,0}) \end{aligned}$$

Here

$$\begin{aligned} E(Z_{k,0}\bar{Z}_{k,0}) &= \sum_{i,m,n=0}^1 \Pr(X_{k,0}=i, Z_{k,0}=1) \Pr(X_{k,1}=m, Z_{k,1}=n) \frac{\exp(\lambda_0) + n \exp(n\lambda_0)}{\exp(\lambda_0) + \exp(n\lambda_0)} \\ &= E(Z_{k,0}) \sum_{m,n=0}^1 p_{m,n} \frac{\exp(\lambda_0) + n \exp(n\lambda_0)}{\exp(\lambda_0) + \exp(n\lambda_0)} = E(Z_{k,0}) \left\{ p_{+,1} + \frac{p_{+,0} \exp(\lambda_0)}{\exp(\lambda_0) + 1} \right\}, \end{aligned}$$

and

$$\begin{aligned} E(\bar{Z}_{k,0}^2) &= \Pr(Z_{k,0}=1, Z_{k,1}=1) \\ &\quad + \{\Pr(Z_{k,0}=1, Z_{k,1}=0) + \Pr(Z_{k,0}=0, Z_{k,1}=1)\} \left(\frac{\exp(\lambda_0)}{\exp(\lambda_0) + 1} \right)^2 \\ &= E(Z_{k,0})p_{+,1} + \{E(Z_{k,0})p_{+,0} + (1 - E(Z_{k,0}))p_{+,1}\} \left(\frac{\exp(\lambda_0)}{\exp(\lambda_0) + 1} \right)^2. \end{aligned}$$

$$\begin{aligned} \therefore E[l_\lambda^2(\psi, \lambda)|_{(0, \lambda_0)}] &= K \left[E(Z_{k,0}) - 2E(Z_{k,0}) \left\{ p_{+,1} + \frac{p_{+,0} \exp(\lambda_0)}{\exp(\lambda_0) + 1} \right\} \right. \\ &\quad \left. + E(Z_{k,0})p_{+,1} + \{E(Z_{k,0})p_{+,0} + (1 - E(Z_{k,0}))p_{+,1}\} \left(\frac{\exp(\lambda_0)}{\exp(\lambda_0) + 1} \right)^2 \right] \\ &= K \left[E(Z_{k,0})p_{+,0} \left\{ 1 - 2 \frac{\exp(\lambda_0)}{\exp(\lambda_0) + 1} + \left(\frac{\exp(\lambda_0)}{\exp(\lambda_0) + 1} \right)^2 \right\} \right. \\ &\quad \left. + (1 - E(Z_{k,0}))p_{+,1} \left(\frac{\exp(\lambda_0)}{\exp(\lambda_0) + 1} \right)^2 \right] \\ &= K \left[E(Z_{k,0})p_{+,0} \left(\frac{1}{\exp(\lambda_0) + 1} \right)^2 + (1 - E(Z_{k,0}))p_{+,1} \left(\frac{\exp(\lambda_0)}{\exp(\lambda_0) + 1} \right)^2 \right] \end{aligned}$$

In contrast,

$$\begin{aligned} E[-l_{\lambda\lambda}(\psi, \lambda)|_{(0, \lambda_0)}] &= K \{E(\bar{Z}_{k,0}^2) - E(\bar{Z}_{k,0})^2\} = K \{E(\bar{Z}_{k,0}) - E(\bar{Z}_{k,0}^2)\} \quad (\because \bar{Z}_{k,0}^2 = \bar{Z}_{k,0}) \\ &= K \left[E(Z_{k,0}) \left(p_{+,1} + p_{+,0} \frac{\exp(\lambda_0)}{\exp(\lambda_0) + 1} \right) + (1 - E(Z_{k,0})) \frac{p_{+,1} \exp(\lambda_0)}{\exp(\lambda_0) + 1} \right. \\ &\quad \left. - E(Z_{k,0})p_{+,1} - \{E(Z_{k,0})p_{+,0} + (1 - E(Z_{k,0}))p_{+,1}\} \left(\frac{\exp(\lambda_0)}{\exp(\lambda_0) + 1} \right)^2 \right] \\ &= K \left(\frac{1}{\exp(\lambda_0) + 1} \right)^2 \{E(Z_{k,0})p_{+,0} \exp(\lambda_0) + (1 - E(Z_{k,0}))p_{+,1} \exp(\lambda_0)\} \\ &= K \left(\frac{1}{\exp(\lambda_0) + 1} \right)^2 \{(1 - E(Z_{k,0}))p_{+,1} \exp(2\lambda_0) + E(Z_{k,0})p_{+,0}\} \quad (\because (3)) \\ &= K \left[E(Z_{k,0})p_{+,0} \left(\frac{1}{\exp(\lambda_0) + 1} \right)^2 + (1 - E(Z_{k,0}))p_{+,1} \left(\frac{\exp(\lambda_0)}{\exp(\lambda_0) + 1} \right)^2 \right] \end{aligned}$$

$$= E[t_{\lambda}^2(\psi, \lambda)|_{(0, \lambda_0)}].$$

Thus $\text{tr}(M) = 1$, which leads noncentrality of $1 - 1 + \Delta = \Delta$. Hence, for given parameter values (ψ, λ) and covariate distribution $f(X, Z)$, the actual likelihood ratio statistic under the alternative hypothesis is performed follows a noncentral chi-square distribution with one d.f. and noncentrality Δ .

3. Simulation studies

Simulation studies are performed for evaluating the accuracy of the proposed approach for several parameter values and data configurations in the simple nested case-control sample analysis. A full cohort data set includes K individuals of cases and $(N - K)$ individuals of non-cases with their observed time, event status, and covariate values. Two independent dichotomous covariates, (X, Z) are used with $\Pr(X = 1) = \pi_1$ for $\pi_1 = (0.25, 0.5)$ and $\Pr(Z = 1) = \pi_2$ for $\pi_2 = 0.5$. These covariates can be considered as sex, exposure to the environmental factors, smoking status, other lifestyle factors, or genetic markers. Proportional hazards model such as $h(t | X, Z) = h_0(t) \exp(X\psi + Z\lambda)$ is used to generate the observed time and event status. The parameter of interest, ψ is taken to be $\ln(1.25)$, $\ln(1.5)$, and $\ln(2)$. The confounding parameter λ is set as $\ln(2)$. A Weibull distribution is used for the baseline hazard function to generate the failure times, T_i ($i = 1, \dots, N$). The exponential distribution is used to generate the censoring time, C_i ($i = 1, \dots, N$). Finally, we obtain the observed time, $\tilde{T}_i$ ($i = 1, \dots, N$) as a minimum of (T_i, C_i) and the event status, τ_i , that equals one if $T_i \leq C_i$ and zero otherwise. Then the failure times of K cases are sorted with an ascending order such as $t_{(1)} < \dots < t_{(k)} < \dots < t_{(K)}$, assuming no ties. The risk set, R_k is defined as $R_k = \{i \mid \tilde{T}_i \geq t_{(k)}\}$. The total number of subjects, N is arbitrary chosen to achieve the number of cases (K) such as $K = (100, 200, 300, 400, 500)$. Given the full cohort data, all cases are extracted and a control per case is randomly sampled from each risk set to form a nested case-control data set.

For a given model, covariate distribution, regression coefficients, and failure time distribution with specified number of cases, estimates of actual power for testing the null hypothesis $H_0: \psi = 0$ against the alternative hypothesis $H_a: \psi \neq 0$ with significance level 0.05 are computed through Monte Carlo simulation based on 10,000 replicate nested case-control samples. A unique nested case-control sample is extracted from each full cohort data set (i.e., 10,000 replicate full cohort data sets are prepared).

Estimates of power associated with the above same parameter values and data configurations are then computed with our proposed method. All calculations are performed using program written with Statistical Software R (R Development Core Team, Version 2.1.1, 2005). The R programs for our proposed approach can be obtained from the authors.

Results of the simulation studies are summarized in Table 1. In general, estimates of power using our proposed method are consistent with those of actual power from Monte Carlo simula-

Table 1. Comparison of the simulated power (RS1) and asymptotic power (RS2) for given $\Pr(X = 1)$, number of cases, and parameter values of $\psi = (\ln(2.0), \ln(1.5), \ln(1.25))$.

$\Pr(X = 1)$	Number of cases	ψ					
		$\ln(2.0)$		$\ln(1.5)$		$\ln(1.25)$	
		RS1	RS2	RS1	RS2	RS1	RS2
0.5	100	0.64	0.64	0.28	0.28	0.12	0.12
	200	0.91	0.91	0.50	0.50	0.19	0.19
	300	0.98	0.98	0.67	0.67	0.26	0.26
	400	1.00	1.00	0.79	0.79	0.33	0.33
	500	1.00	1.00	0.87	0.87	0.40	0.40
0.25	100	0.59	0.59	0.24	0.24	0.10	0.10
	200	0.87	0.87	0.43	0.43	0.16	0.16
	300	0.97	0.97	0.59	0.59	0.22	0.22
	400	0.99	0.99	0.71	0.71	0.28	0.28
	500	1.00	1.00	0.80	0.80	0.33	0.33

tion, which suggests that the proposed method performs well and can be practically useful.

4. Discussion

We propose in this paper an approach for power calculation for likelihood ratio tests in the simple nested case-control sample analysis. This approach is an extension of the work by Self et al. (1992) to accommodate the risk set sampling of controls and the use of proportional hazards models. Our focused scenario is 1 : 1 case-control ratio, simple random sampling design, and two independent dichotomous covariates with no interaction effects. The approximation of the noncentrality parameter in a noncentral chi-square distribution of the likelihood ratio statistic is also derived. Under this simple scenario our approach maintains great accuracy based on the simulation results.

Effects of the covariate Z on the risk of the outcome are differently adjusted in two methods. As for the method proposed, the effects of the covariate Z are adjusted in a proportional hazards model. On the other hand, the methods proposed by Sinha and Mukherjee (2006) and Lachin (2008) uses the matching design (i.e., controls are matched to cases with respect to the covariate Z). However the model-based adjustment of the covariate effects may be feasible in some situation where appropriate controls are not found to be matched with cases.

Structure of sampling design is incorporated into a partial likelihood function in our proposed method, not in the score test based method. The design of selecting a case-control set includes matching (stratifying) and counter-matching besides the random sampling that was considered in this paper. A similar approach can be used to extend the proposed asymptotic method for other designs. A choice of the design affects the probability of picking a particular case-control set and the partial likelihood function. Therefore the extension may yield an intricate calculation

of $tr(M)$.

Other extensions can be made to permit larger case-control ratios, 3 or more covariates, and an interaction term of two covariates. When a larger value of the case-control ratio is chosen, the term $\exp(X_{k,1}\psi + Z_{k,1}\lambda)$ in the log-likelihood function $l(\psi, \lambda)$ needs to be summed up for the number of controls. And when the model includes 3 or more covariates and an interaction term, the second order derivatives of the log-likelihood function $l(\psi, \lambda)$ with respect to the parameter λ being a vector will be a matrix. In each extension the computation of $tr(M)$ will be fairly complicated. These interesting technical challenges may be undertaken in future researches.

Acknowledgements

Supported in part by Grants-in-Aids for Scientific Research (C) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (No. 21500279) and for Collaborative Research Program (2) from the Institute of Mathematical Statistics of Japan (No. 20-CR-2054, 21-CR-2060). We are grateful to the referees for their constructive suggestions that improved the presentation of this paper.

REFERENCES

- Cologne, J.B., Sharp, G.B., Neriishi, K., Verkasalo, P.K., Land, C.E., Nakachi, K. (2004). Improving the efficiency of nested case-control studies of interaction by selecting controls using counter matching on exposure. *International Journal of Epidemiology* **33**(3), 485-492.
- Fujii, Y., Cologne, J., Izumi, S. (2007). Sampling design: Counter matching. *Japanese Journal of Biometrics* **28**(1), 47-58. (in Japanese).
- Lachin, J.M. (2008). Sample size evaluation for a multiply matched case-control study using the score test from a conditional logistic (discrete Cox PH) regression model. *Statistics in Medicine* **27**(14), 2509-2523.
- Langholz, B. (2005). Case-control study, nested. In *Encyclopedia of Biostatistics*, Second edition. (P. Armitage & T. Colton, eds), 646-655. John Wiley & Sons, Ltd, Chichester.
- Langholz, B. and Goldstein, L. (1996). Risk set sampling in epidemiologic cohort studies. *Statistical Science* **11**, 35-53.
- McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, 2nd edition. London: Chapman and Hall.
- R Development Core Team. (2005). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL <http://www.R-project.org>.
- Rundle, A.G., Vineis, P. and Ahsan, H. (2005). Design options for molecular epidemiology research within cohort studies. *Cancer Epidemiology, Biomarkers, and Prevention* **14**(8), 1899-1907.

- Self, S.G. and Mauritsen, R.H. (1988). Power/sample size calculations for generalized linear models. *Biometrics* **44**, 79-86.
- Self, S.G., Mauritsen, R.H. and Ohara, J. (1992). Power calculation for likelihood ratio tests in generalized linear models. *Biometrics* **48**, 31-39.
- Shieh, G. (2000). On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics* **56**(4), 1192-1196.
- Sinha, S. and Mukherjee, B. (2006). A score test for determining sample size in matched case-control studies with categorical exposure. *Biometrical Journal* **48**(1), 35-53.