Published in final edited form as:

Stat Med. 2008 June 30; 27(14): 2509–2523. doi:10.1002/sim.3057.

Sample Size Evaluation for a Multiply Matched Case-Control Study Using the Score Test From a Conditional Logistic (Discrete Cox PH) Regression Model

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

The conditional logistic regression model (Breslow NE. Covariance adjustment of relative-risk estimates in matched studies. *Biometrics*, 1982; **38**:661–672) provides a convenient method for the assessment of qualitative or quantitative covariate effects on risk in a study with matched sets, each containing a possibly different numbers of cases and controls. The conditional logistic likelihood is identical to the stratified Cox proportional hazards model likelihood with an adjustment for ties (Regression models and life-tables (with discussion). *J. Roy. Statist. Soc., B*, 1972; **34**:187–220). This likelihood also applies to a nested case control study with multiply matched cases and controls selected from those at risk at selected event times. Herein the distribution of the score test for the effect of a covariate in the model is used to derive simple equations to describe the power of the test to detect a coefficient θ (log odds ratio or log hazard ratio), or the number of cases (or matched sets) and controls required to provide a desired level of power. Additional expressions are derived for a quantitative covariate as a function of the difference in the assumed mean covariate values among cases and controls, and for a qualitative covariate in terms of the difference in the probabilities of exposure for cases and controls. Examples are presented for a nested case-control study and a multiply matched case-control study.

Keywords

Sample size; power; conditional logistic model; Cox Proportional Hazards Model; multiple matching; case-control study; nested case-control study

1. INTRODUCTION

In a matched case-control study, one or more disease-free controls are matched with respect one or more characteristics to each case with the disease or outcome of interest. The analysis can then be conducted using the conditional logistic regression model of Breslow [1] to assess the effects of binary and/or quantitative covariates on the risk (log odds) of the outcome. Herein the score test is derived that is applicable to either a binary or quantitative covariate, its distribution under the alternative hypothesis is described, and expressions for the computation of power or sample size are derived.

For a binary exposure variable and a single control ($m = 1$) matched to each case, a simple analysis of matched pairs can be employed consisting of the McNemar [2] test and the computation of a conditional odds ratio (cf. [3], Chapter 5). For $m > 1$ controls per case, Miettinen [4] described methods for analysis of matched sets with $m > 1$ matched controls per case. Walter [5] generalized Miettinen's test to the design with variable numbers of controls matched to each case; i.e. where $m_i \in (1, 2, 3, \dots)$ for the i th matched set. He also

presented a Z or t -test for quantitative covariate assumed to be normally distributed among the cases and controls with differing means and possibly variances. Miettinen [4] and Walter [5] also describe expressions for the power of their tests and sample size calculations. Ury [6], Taylor [7] and Lui [8] present expressions for the asymptotic relative efficiency under a local alternative of Miettinen's test with m controls relative to McNemar's test with a single control. These expressions allow computation of power or sample size, approximately, for Miettinen's test based on a computation of power or sample size for McNemar's test. Ejigou [9] also presents an evaluation of the sample size for a study with multiple controls based on an estimate of the conditional odds ratio and a corresponding estimation-based test statistic. Sinha and Mukerjee [10] derive the power of a score test based on a conditional likelihood as a function of the possible numbers exposed within each set, and allows for polychotomous and ordinal exposure variables. All of these methods only allow for a single case matched to multiple controls.

The conditional logistic regression model also provides a convenient method for the analysis of such data, and provides for further generalizations to studies with variable numbers of cases as well as controls within each matched set. This could apply to a retro-spective case-control study with variable numbers of cases and controls in each matched set, as well as other studies with inherently matched sets, such as a family or sib-ship study of the association of a genotype with the prevalence of a phenotype among family members or siblings within a family. The non-central or non-null distribution of the score test then provides a direct method for the computation of the power or sample size for a study using a quantitative or qualitative covariate.

The cases and controls may also be selected from within a prospective cohort study. In the nested case-control study with no tied event times, a set of N cases is selected at random from among all those that occur during the study. Thus, each of the N cases can be characterized as having an event indicator $\delta_i = 1$ and event time t_i . Then for each case, m subjects are selected as controls who are elements of the risk set at time t_i , and known not to have had the event up to that time. This constructs a set of $(m+1)$ subjects matched at that time. Note that a control for one case could later become a case. The nested case-control design can also be applied to a cohort with tied event times yielding multiple cases at a given time with possibly variable numbers of controls matched to the times of the cases.

Liddell, McDonald, Thomas and Cunliffe [11] describe a matched case-control study embedded within a prospective cohort follow-up study. They conducted an analysis using the Miettinen methods. However, in a supplement to the main paper, Thomas [12] shows that the Cox Proportional Hazards (PH) model [13] could be applied to this analysis. Goldstein and Langholz [14] use the martingale theory for counting processes to prove that the resulting coefficient estimates are asymptotically normally distributed under nested case-control sampling.

The Cox likelihood with Cox's adjustment for tied event times [13] is equivalent to the conditional logistic regression model of Breslow [1], the two yielding an equivalent likelihood. Thus, the conditional logistic regression model can be fit using a stratified Cox PH model with Cox's adjustment for ties.

Schoenfeld [15] describes an equation for the total number of deaths (subjects with the outcome event) required to provide a desired level of power for a single binary covariate in a Cox PH model. Hsieh and Lavori [16] extend the Schoenfeld derivation for the case of a single quantitative covariate. Both are based on the expression for the score test of a covariate in the general PH model.

Herein I derive the form of the score test for a qualitative or quantitative covariate effect on the log odds in a conditional logistic regression model with variable numbers of cases and controls in each matched set. The test can be applied to a qualitative (binary) or quantitative covariate. The test can also be applied to the Cox PH model for a nested case-control study. For a binary covariate, the resulting test is a generalization of the tests of Miettinen [4] and Walter [5]. For a quantitative covariate, the score test is simpler and more general than that of Walter [5] that assumes an underlying normal distribution. I then derive the power function of the score test and describe sample size calculations for the design of a matched case-control study. In doing so I derive a useful expression for the log odds ratio in the regression model as a function of the difference in means of a quantitative covariate between cases and controls, or in probabilities of a qualitative covariate, that may further facilitate the evaluation of sample size for such studies.

2. MULTIPLY MATCHED SCORE TEST

2.1 A Quantitative Covariate

Consider a study with N matched sets. The i th matched set consists of n_i members of whom $d_i \leq D$ are cases and $m_i \leq M$ are controls, where D and M are the maximum numbers of cases and controls among all sets.

In a nested case-control study, these matched sets are constructed as follows. At the i th event time, assume that $G_i \geq 2$ subjects may have the event at time t_i (or during the i th interval) among the risk set of R_i subjects at risk at that time. Then d_i subjects are randomly selected from among the G_i to be cases, and m_i are randomly selected from among the remaining $R_i - G_i$ to be controls.

Consider a single covariate X , possibly quantitative, with values x_{ij} for the j th member of the i th matched set with the subjects ordered such that $x_{i1}, \dots, x_{id_i}$ refers to the covariate values of the d_i cases, and $x_{i(d_i+1)}, \dots, x_{in_i}$ refers to the values for the m_i controls. The resulting conditional logistic regression model likelihood with matched sets containing d_i cases and m_i controls is

$$L = \prod_{i=1}^N \left[\frac{\prod_{j=1}^{d_i} e^{x_{ij}\theta}}{\sum_{l=1}^{n_i} \binom{n_i}{d_i} \prod_{j(l)=1}^{d_i} e^{x_{ij(l)}\theta}} \right] \quad (1)$$

where θ represents the log odds ratio [1]. This is identical to the Cox [13] discrete proportional hazards model for tied event times that would apply to the nested case-control sampling design in which θ is the log hazard ratio. Let $s_i = \sum_{j=1}^{d_i} x_{ij}$ for the i th set and

$s_{i(l)} = \sum_{j(l)=1}^{d_i} x_{ij(l)}$ for the l th of $\binom{n_i}{d_i}$ combinations of d_i subjects within that set. The resulting score equation is

$$U(\theta) = \sum_{i=1}^N \left[s_i - \frac{\sum_l s_{i(l)} e^{s_{i(l)}\theta}}{\sum_l e^{s_{i(l)}\theta}} \right] = \sum_{i=1}^N [s_i - E(s_i | H_1)], \quad (2)$$

where $\sum_l$ denotes $\sum_{l=1}^{\binom{n_i}{d_i}}$. The expected information matrix $I(\theta)$ then is

$$I(\theta) = \sum_{i=1}^N \left[\left(\frac{\sum_l s_{i(l)}^2 e^{s_i \theta}}{\sum_l e^{s_i \theta}} \right) - \left(\frac{\sum_l s_{i(l)} e^{s_i \theta}}{\sum_l e^{s_i \theta}} \right)^2 \right] = \sum_{i=1}^N V_{1i}. \quad (3)$$

Evaluating each under the null hypothesis H_0 : $\theta = 0$ yields

$$U(\theta_0) = \sum_{i=1}^N \left[s_i - \frac{\sum_l s_{i(l)}}{\binom{n_i}{d_i}} \right] = \sum_{i=1}^N [s_i - E(s_i|H_0)] = \sum_{i=1}^N [s_i - \bar{s}_{0i}], \quad (4)$$

and

$$I(\theta_0) = \sum_{i=1}^N \left[\left(\frac{\sum_l s_{i(l)}^2}{\binom{n_i}{d_i}} \right) - \left(\frac{\sum_l s_{i(l)}}{\binom{n_i}{d_i}} \right)^2 \right] = \sum_{i=1}^N V_{0i}. \quad (5)$$

Then the score test is provided by

$$Z = \frac{U(\theta_0)}{\sqrt{I(\theta_0)}} \quad (6)$$

that is distributed as standard normal under H_0 . The quantities s_i and $E(s_i|H_0)$ are readily computed for any quantitative or qualitative covariate. This yields a convenient test. Below the form of the test is shown for specific cases. Note that sets with no cases ($d_i = 0$) or no controls ($m_i = 0$) do not contribute to $U(\theta_0)$ or to $I(\theta_0)$ and can simply be discarded. This also applies to sets where all members share the same covariate value.

In the most general case of a study with variable numbers of cases and controls, let n_{dm} designate the number of matched sets with d cases, $1 \leq d \leq D$, and m controls, $1 \leq m \leq M$, and let x_{dmij} refer to the covariate value for the j th subject of the d th set with d cases and m controls. Then $s_{dmi} = \sum_{j=1}^d x_{dmij}$ and $s_{dmi(l)} = \sum_{j(l)=1}^d x_{dmij(l)}$ for the l th of $\binom{d+m}{d}$ combinations of d subjects within that set, and let

$$E[s_{dmi}|H_0] = \bar{s}_{0dmi} = \frac{\sum_l s_{dmi(l)}}{\binom{d+m}{d}}. \quad (7)$$

Then the score test can be expressed as

$$Z = \frac{\sum_{d=1}^D \sum_{m=1}^M \sum_{i=1}^{n_{dm}} [s_{dmi} - \bar{s}_{0dmi}]}{\left(\sum_{d=1}^D \sum_{m=1}^M \left[\left(\frac{\sum_l s_{dmi(l)}^2}{\binom{d+m}{d}} \right) - \left(\frac{\sum_l s_{dmi(l)}}{\binom{d+m}{d}} \right)^2 \right] \right)^{1/2}} \quad (8)$$

For the special case of a design with $D = 1$ case and m controls in each matched set, so that $n = m + 1$, then $s_i = x_{i1}$ and

$$\bar{s}_{0i} = \frac{\sum_l s_{il(l)}}{\binom{n_i}{d_i}} = \bar{x}_{0i} = \frac{\sum_{j=1}^{m+1} x_{ij}}{m+1} \quad (9)$$

so that

$$\begin{aligned} U(\theta_0) &= \sum_{i=1}^N \left[x_{i1} - \bar{x}_{0i} \right] = \sum_{i=1}^N \left[x_{i1} - \frac{\sum_{j=1}^{m+1} x_{ij}}{m+1} \right], \\ I(\theta_0) &= \sum_{i=1}^N \left[\frac{\sum_{j=1}^{m+1} (x_{ij} - \bar{x}_{0i})^2}{m+1} \right] = \sum_{i=1}^N V_{0i}. \end{aligned} \quad (10)$$

This yields the score test

$$Z = \frac{\sum_{i=1}^N \left[x_{i1} - \bar{x}_{0i} \right]}{\sqrt{\sum_{i=1}^N V_{0i}}}. \quad (11)$$

2.2 A Binary Covariate

Now consider a single binary covariate X to represent “exposed” versus not to some risk factor, where $x_{ij} = 1$ if exposed (or +), 0 if not (−). Let $n_{i+} = \#\{x_{ij} = 1\}$ and $n_{i-} = \#\{x_{ij} = 0\}$ where $n_i = n_{i-} + n_{i+}$. As above, for the matched set of n_i subjects, the members are ordered such that the first d_i are the cases and the remaining $m_i = n_i - d_i$ are the controls. Then $s_i = \sum_{j=1}^{d_i} x_{ij}$ is the number of cases positive for the covariate in the i th set, and s_{i0} is the like

sum for the first d_i observations in the i th combination of $\binom{n_i}{d_i}$ subjects in the i th set.

For the simple 1: m design with a single case matched to m controls with $n = m + 1$ subjects per set, from (10) it follows that the score equation under H_0 is

$$\begin{aligned} U(\theta_0) &= \sum_{i=1}^N \left[x_{i1} - \bar{x}_{0i} \right] = s - \sum_{i=1}^N \left[\frac{n_{i+}}{n} \right] \\ &= s - \sum_{i=1}^N p_{0i} = s - e_0 \end{aligned} \quad (12)$$

where p_{0i} is the probability of a positive covariate in the i th set under H_0 , and e_0 is the sum of the probability positive for all cases, or the expected number under H_0 . The corresponding Information function is

$$\begin{aligned} I(\theta_0) &= \sum_{i=1}^N \left[\frac{\sum_{j=1}^{n_i} x_{ij}^2}{n} - \bar{x}_i^2 \right] = \sum_{i=1}^N \left[\frac{n_{i+}}{n} - \left(\frac{n_{i+}}{n} \right)^2 \right] \\ &= \sum_{i=1}^N [p_{0i}(1 - p_{0i})] = \sum_{i=1}^N V_{0i}. \end{aligned} \quad (13)$$

Then this yields the score test

$$Z = \frac{U(\theta_0)}{\sqrt{I(\theta_0)}} = \frac{s - e_0}{\sqrt{\sum_{i=1}^N V_{0i}}} = \frac{s - e_0}{\sqrt{\sum_{i=1}^N [p_{0i}(1 - p_{0i})]}}. \quad (14)$$

Miettinen's [4] test of H_0 for the 1: m design is readily shown to be equivalent. Walter [5] presents a generalization of Miettinen's test for the 1: m_i design with variable numbers of controls matched to each case, with only a single case per set. His test is also a special case of (8).

3. SAMPLE SIZE AND POWER

3.1 Distribution Under the Alternative

As employed by Schoenfeld [15], and also Hsieh and Lavori [16], under the alternative hypothesis $H_1: \theta \neq 0$, the score equation $U(\theta_0)$ in (4) can be expressed as

$$U(\theta_0) = \sum_{i=1}^N [s_i - E(s_i|H_1)] + \sum_{i=1}^N [E(s_i|H_1) - E(s_i|H_0)]. \quad (15)$$

The first term is simply the score function $U(\theta)$ in (2) and from basic principles [3], $U(\theta) \sim N[0, I(\theta)]$. The second term then is

$$\sum_{i=1}^N [E(s_i|H_1) - E(s_i|H_0)] = \sum_{i=1}^N \left[\frac{\sum_l s_{il(l)} e^{s_{il}\theta}}{\sum_l e^{s_{il}\theta}} - \frac{\sum_l s_{il(l)}}{\binom{n_i}{d_i}} \right] \quad (16)$$

Applying a Taylor's expansion for θ in the neighborhood of $\theta_0 = 0$, it can then be shown that

$$\sum_{i=1}^N [E(s_i|H_1) - E(s_i|H_0)] \cong \theta I(\theta_0) = \theta \sum_{i=1}^N V_{0i} \quad (17)$$

Combining the two results, it follows that

$$U(\theta_0)|H_1 \sim N[\theta I(\theta_0), I(\theta)] \quad (18)$$

Thus

$$Z|H_1 \sim N\left[\theta\sqrt{I(\theta_0)}, \left(\frac{I(\theta)}{I(\theta_0)}\right)\right]. \quad (19)$$

For a specific true value θ under the alternative hypothesis, the basic equation for the power of the test [3] yields

$$\begin{aligned} |\theta|\sqrt{I(\theta_0)} &= \frac{Z_{1-\alpha} + Z_{1-\beta}\sqrt{I(\theta)/I(\theta_0)}}{\frac{|\theta|\sqrt{I(\theta_0)} - Z_{1-\alpha}}{\sqrt{I(\theta)/I(\theta_0)}}} \quad (20) \end{aligned}$$

where $Z_{1-\alpha}$ is the critical value for the test, using $Z_{1-\alpha/2}$ for a two-sided test, and the power of the test is provided by $\Phi(Z_{1-\beta})$, $\Phi(z)$ being the standard normal *cdf* at z . Approximately $I(\theta) \cong I(\theta_0)$, as under a local alternative, in which case the expression for power simplifies to

$$Z_{1-\beta} = |\theta|\sqrt{I(\theta_0)} - Z_{1-\alpha}. \quad (21)$$

3.2 A Quantitative Covariate

3.2.1 As a Function of θ —Assume that X is a quantitative covariate with variance σ_i^2 within the i th matched set with d_i cases and m_i controls, $n_i = d_i + m_i$, so that within that set $V(s_i) = V(s_{i(l)}) = d_i \sigma_i^2$. Noting that V_{0i} in (5) is similar to the expression for $V(s_i)$ but without the usual “−1” correction in the denominator, it follows that

$$V_{0i} = \left[\frac{\binom{n_i}{d_i} - 1}{\binom{n_i}{d_i}} \right] V(s_i) = d_i \sigma_i^2 \left[\frac{\binom{n_i}{d_i} - 1}{\binom{n_i}{d_i}} \right] = d_i \sigma_i^2 \left[\frac{n_i! - d_i! m_i!}{n_i!} \right] \quad (22)$$

Under some model that provides a set of values $\{n_i, d_i, \sigma_i^2\}$ for all N matched sets, this then provides the value of the variance term for each set, and thus the value of $I(\theta_0) = \sum_{i=1}^N V_{0i}$.

Under the model that generated the $\{n_i, d_i, \sigma_i^2\}$, the power of the test in (6) is provided by the expression for $Z_{1-\beta}$ in (21). Under an appropriate model that provides an arbitrary number of such sets, such as sampling from a distribution for (n, d, σ^2) , it would also be possible to solve for the number of such sets N that would provide a desired level of power.

When there is a fixed number of cases and controls for all matched sets with a common variance among sets, i.e. $n_i = n$, $d_i = d$, and $\sigma_i^2 = \sigma^2$ for all $i = 1, \dots, N$, then

$$I(\theta_0) = Nd\sigma^2 \left[\frac{\binom{n}{d} - 1}{\binom{n}{d}} \right] = Nd\sigma^2 \left[\frac{n! - d!m!}{n!} \right]. \quad (23)$$

This result also applies to the heteroscedastic case for suitably large N such that $\sum_i \sigma_i^2 \rightarrow N\sigma^2$ where $E(\sigma_i^2) = \sigma^2$. Substituting into (21), the power of the test is provided by

$$Z_{1-\beta} = |\theta| \sigma \sqrt{Nd} \left[\frac{n! - d!m!}{n!} \right]^{1/2} - Z_{1-\alpha}. \quad (24)$$

Thus, for a given n and d , the number of matched sets (N) required to provide a desired level of power $1 - \beta$ using a test at level α (or $\alpha/2$ two-sided) is provided by

$$N = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\theta^2 \sigma^2 d \left[\frac{n! - d!m!}{n!} \right]} \quad (25)$$

When there is a single case ($d = 1$) and m controls for each of the N sets, where $n = m + 1$, then $I(\theta_0) = Nm\sigma^2/(m + 1)$. The power of the test is then provided by

$$Z_{1-\beta} = |\theta| \sigma \sqrt{\frac{Nm}{m+1}} - Z_{1-\alpha} \quad (26)$$

and the number of cases required to provide the desired level of power to detect a coefficient θ is

$$N = \left[\frac{Z_{1-\alpha} + Z_{1-\beta}}{\theta\sigma} \right]^2 \left(\frac{m+1}{m} \right) = K \left(\frac{m+1}{m} \right). \quad (27)$$

Note that N is a decreasing function of m . For $m = 1$, $N = 2K$. For $m = 2$, $N = 1.5K$. For $m = 3$, $N = K4/3$, etc.

In the above expressions, $\theta = \ln(O)$ is the log odds ratio (hazard ratio in the Cox PH model) per unit change in X . In some cases it may be more convenient to specify the odds ratio per c unit change in X , say $O_c = \exp(c\theta)$, from which O is obtained as $O = O_c^{(1/c)}$. The effect size may also be described in terms of the odds ratio per standard deviation unit, say O_σ , assuming a common variance within sets. In this case the term $\theta\sigma$ is replaced by $\ln(O_\sigma) = \ln(O^\sigma)$ in the above expressions. The units change could also be specified in terms of a fraction of the within-set standard deviation. For example, assume that $\sigma = 10$ and it is desired to detect an odds ratio of 2 per $c = \sigma/2 = 5$ unit difference in X . The inverse implies that the odds ratio per unit change in X is $O = 2^{(1/5)} = 1.1487$ and $\theta = \ln(O) = 0.139$.

The above expressions for $Z_{1-\beta}$ and N are similar to those of Hsieh and Lavori [16] for a quantitative covariate in the Cox PH model. At first glance these equations are counterintuitive, implying that as the variance of the covariate increases, the required number of subjects with the event decreases, and that for any number of events, power increases. The reason, however, is that the higher the σ , the greater the variation in risk over the range of the covariate. For a given O and σ , the odds ratio per standard deviation change in the covariate is O^σ , and the odds ratio comparing the extremes of the distribution of X ($\pm 3\sigma$) is $O^{6\sigma}$. Thus for a given O , the range of risk increases as σ increases, and so also does power.

3.2.2 As a Function of the Mean Difference—It may be difficult for investigators to specify the minimal treatment effect of interest for a quantitative covariate in terms of the odds ratio or relative hazard per unit change in X or per standard deviation. In this situation, it might be more relevant to describe the power of a study in terms of the distributions of the covariate among the cases and controls within sets. Assume that the quantitative covariate is distributed as $X \sim f_1(\mu_{i1}, \sigma_i^2)$ among the cases and $X \sim f_2(\mu_{i2}, \sigma_i^2)$ among the controls within the i th set, for some distributions f_1 and f_2 with average variance $E(\sigma_i^2) = \sigma^2$ and average mean difference $E(\mu_{i1} - \mu_{i2}) = (\mu_1 - \mu_2) = \Delta \neq 0$ under H_1 over all sets.

First consider the design with d cases and m controls for all of the N sets. Let $E[U_i(\theta_0)|H_1]$ denote the expectation of the contribution of the i th set to the score equation $U(\theta_0)$ in (4) with respect to the distribution of the covariate values under H_1 . Then for N sets (N suitably large) $E[U(\theta_0)|H_1] = NE\{E[U_i(\theta_0)|H_1]\}$ that is a function of (μ_1, μ_2) . This yields

$$\begin{aligned} E[U(\theta_0)|H_1] &= NE\{E[s_i|H_1]\} - NE\{E[(s_i H_0)|H_1]\} \\ &= Nd\mu_1 - N \binom{n}{d}^{-1} \left[d\mu_1 + \binom{m}{1} [(d-1)\mu_1 + \mu_2] \right. \\ &\quad \left. + \binom{m}{2} [(d-2)\mu_1 + 2\mu_2] + \dots + \binom{m}{d} d\mu_2 \right]. \end{aligned} \quad (28)$$

Simultaneously, from (18) above, where $\sum_i \sigma_i^2 \rightarrow N\sigma^2$, the expected value can be expressed as a function of θ to yield

$$E[U(\theta_0)|H_1] = \theta I(\theta_0) = \theta N d \sigma^2 \left[\frac{\binom{n}{d} - 1}{\binom{n}{d}} \right]. \quad (29)$$

Equating (28) to (29) one can then solve for θ as a function of $\mu_1, \mu_2, \sigma^2, d$, and m . For example, for $d = 2$ and $m = 3$ the resulting equations yield

$$N \left[2\mu_1 - \frac{2\mu_1 + 6[\mu_1 + \mu_2] + (3)(2)\mu_2}{10} \right] = \theta N 2 \sigma^2 \left[\frac{\binom{5}{2} - 1}{\binom{5}{2}} \right] \quad (30)$$

$$\left[\frac{12(\mu_1 - \mu_2)}{18\sigma^2} \right] = \theta$$

so that the model parameter θ is a function of the mean difference divided by the variance times a constant depending on d and m .

The power of the study is then provided by substituting the above value of θ , and the value of $I(\theta_0)$ in (23) into the expression for $Z_{1-\beta}$ in (24). The number of such sets required to provide the desired level of power is then provided by (25).

The above developments could also be applied to the case where there are (d_i, m_i) cases and controls within the i th set in which case $E[U(\theta_0)|H_1] = NE\{E[U(\theta_0)|H_1]\}$ and $I(\theta_0) = NE(V_{0i})$, each with respect to the distribution of (d_i, m_i) . Substituting into (29) one could solve for θ as a function of μ_1, μ_2, σ^2 .

For the special case where $d = 1$ and fixed m and $n = m + 1$, the above reduces to

$$E[U(\theta_0)|H_1] = NE \left\{ E \left[(x_{i1} - \bar{x}_{0i}) | H_1 \right] \right\}$$

$$= NE \left[E(x_{i1}) - \frac{\sum_{j=1}^{m+1} E(x_{ij})}{m+1} \right] = \theta N \left(\frac{m\sigma^2}{m+1} \right) \quad (31)$$

which implies that

$$N\mu_1 - N \left[\frac{\mu_1 + m\mu_2}{m+1} \right] = N\theta \left[\frac{m\sigma^2}{m+1} \right] \quad (32)$$

$$\frac{\mu_1 - \mu_2}{\sigma^2} = \theta$$

for any m . Substituting this expression for θ in (24), the power of a study with N cases each matched to m controls is provided by

$$Z_{1-\beta} = \sqrt{\frac{Nm}{m+1}} \left[\frac{\mu_1 - \mu_2}{\sigma} \right] - Z_{1-\alpha}. \quad (33)$$

Likewise, from (27), the number of matched sets required to provide a given level of power to detect a difference $\Delta = (\mu_1 - \mu_2)$ is provided by

$$N = \left[\frac{(Z_{1-\alpha} + Z_{1-\beta})\sigma}{\Delta} \right]^2 \left(\frac{m+1}{m} \right) \quad (34)$$

Compared to the expressions for power in (26) and number of sets in (27) in terms of θ , we now see that power increases (sample size decreases) as the variance decreases for a given mean difference between cases and controls.

For the design with variable numbers of controls matched to each case, with a single case per set, Walter [5] presents a test for the difference in the means of a quantitative covariate, assumed normally distributed among cases and controls with different means and variances. His test is computationally different from the expression (8) above. He also presents an expression for the power of his test and the computation of sample size. Both his test and his equations for sample size and power are computationally different from the above. The expressions herein are distribution-free, only requiring specification of the means and variance of the covariate, and also allow for any number of cases and controls within each matched set.

3.3 For a Binary Covariate

In the most general case with variable numbers of cases (d_i) and controls (m_i) within each of N matched sets, under the alternative hypothesis $H_1: \theta \neq 0$, the expected value of the score statistic from (18) is $E[U(\theta_0)|H_1] = \theta I(\theta_0)$ where the information function has elements with V_{0i} as in (5). Since s_i is a sum of d_i Bernoulli variables, then $V(s_i) = d_i \sigma_{0i}^2$ where $\sigma_{0i}^2 = \pi_{0i}(1 - \pi_{0i})$ and $\pi_{0i} = E(p_{0i})$ is the probability of a positive covariate value under H_0 in the i th set. For a collection of N sets with d_i cases, m_i controls and null probability π_{0i} ($i = 1, \dots, N$), the power of the study to detect a coefficient θ is provided by (21).

For the case where there are d cases and m controls for all N matched sets, then it follows that

$$I(\theta_0) = d \left[\frac{\binom{n}{d} - 1}{\binom{n}{d}} \right] \sum_{i=1}^N \sigma_{0i}^2 = Nd \sigma_0^2 \left[\frac{n! - d!m!}{n!} \right]. \quad (35)$$

where σ_0^2 is the average Bernoulli variance over all sets under H_0 . Alternately, under H_0 it can be assumed that there is a common probability of exposure $E(X) = \pi_0$ among the cases and controls in all sets with the common Bernoulli variance $\sigma_0^2 = \pi_0(1 - \pi_0)$. Then the power of the study is provided by (24) and the sample size by (25).

Power and sample size can also be expressed in terms of the differences in the probabilities under the alternative hypothesis by again applying the result in (15). Under H_1 assume common (or average) probabilities of exposure $E(x_{1i}) = \pi_1$ among the cases in each set and $E(x_{ij}) = \pi_2$ for the controls. The resulting expression for $E[U(\theta_0)|H_1]$ is as in shown in (28) with (π_1, π_2) substituted for (μ_1, μ_2) , and also in (29) using the null Bernoulli variance σ_0^2 . It follows that

$$\theta = C \left[\frac{\pi_1 - \pi_2}{\sigma_0^2} \right] \quad (36)$$

where C is a function of d and m as shown in the example above for a quantitative covariate (30), where $C = 1$ when $d = 1$ for any value m .

For the design with m controls matched to each case, the resulting expressions for power and sample size as a function of θ are the same as presented in (26) and (27), and as a

function of the difference in probabilities positive for the covariate among cases and controls under the alternative hypothesis, $\Delta = \pi_1 - \pi_2$, in (33) and (34), each using the Bernoulli variance σ_0^2 .

Miettinen [4] presents an expression for the power of his test for a binary covariate in a 1: m design with m controls matched to a single case in each set. Walter [5] presents a generalization of Miettinen's results to a 1: m_i design with variable numbers of controls matched to a single case within each set, that includes Miettinen's original derivation as a special case. His expression for the power of his test is obtained from

$$Z_{1-\beta(W)} = \frac{|\Delta| \sum_{i=1}^N m_i}{\left[(\psi^*/2) \sum_{i=1}^N m_i (m_i + 1) \right]^{1/2}} - Z_{1-\alpha} \quad (37)$$

where ψ^* is the probability that a random pair of controls is discordant that is an approximation to the probability (ψ) that a case and one of its controls are discordant under H_0 . Conversely, from the derivations herein, the power for a study with a single case and variable numbers of controls in each set, with a common probability positive over all sets, is

$$Z_{1-\beta} = |\theta| \left(\pi_0 (1 - \pi_0) \sum_{i=1}^N \left[\frac{n_i - 1}{n_i} \right] \right)^{1/2} - Z_{1-\alpha} \quad (38)$$

where $(n_i! - d_i! m_i!)/n_i!$ reduces to $(n_i - 1)/n_i$ for $d_i = 1$ and $m_i = n_i - 1$. Whereas Miettinen [4] and Walter [5] describe the power of the test as a function of a common difference in probabilities positive for cases and controls among all matched sets, the methods herein assume a common odds ratio for all matched sets in accordance with the underlying assumptions of the conditional logistic regression model on which the test is based.

4. SOME EXAMPLES

4.1 A Quantitative Covariate

The study of the Epidemiology of Diabetes Interventions and its Complications (EDIC) is a long term follow-up study of the cohort of 1441 subjects originally enrolled in the Diabetes Control and Complications Trial. The DCCT/EDIC [17] showed that lower levels of glucose markedly reduced the risk of cardiovascular disease (CVD) in this cohort. The study is now planning a nested case-control study to measure "biomarkers" of oxidative stress, one of the possible mechanisms through which prolonged hyperglycemia could lead to CVD. Recently Blankenberg, et al., [18] reported that the biomarker soluble intercellular adhesion molecule had a hazard ratio of 1.46 per SD unit difference for risk of CVD. It is anticipated that 125 subjects will have experienced CVD in the DCCT/EDIC, all of whom will be employed as cases. With $m = 2$ controls per case, the 250 time-matched controls will provide 85% power to detect an odds ratio of 1.39 per SD difference using a score test at the 0.05 level, two-sided. In the matched case-control study the odds ratio is approximately equal to the hazard ratio. This odds ratio was obtained by solving (27) for θ with $\sigma = 1$. Using (26), this study would provide power of 93% to detect an odds ratio of 1.46 per SD.

One of the biomarkers of primary interest is the 8-isoprostane/creatinine ratio with a standard deviation of $\sigma = 8.41$ mg/ng from preliminary data. It follows that an odds ratio of 1.39 per SD corresponds to an odds ratio of $(1.39)^{1/8.41} = 1.04$ per mg/ng difference in the biomarker, with corresponding $\theta = 0.038$. From (32), this in turn implies that the study would have 85% power to detect a mean difference between cases and controls of $\mu_1 - \mu_2 = \theta \sigma^2 = 2.69$ ng/mg.

Hosmer and Lemeshow [19] present data from a multiply-matched case control study of factors associated with the risk of a pregnancy culminating in an infant with low birth weight (cf. [3], p. 301 et seq.). The study consisted of 17 matched sets with the following number of sets (w) with d cases and m controls:

d	1	1	1	1	2	2	2	3	4	5	5	6	8
m	4	5	6	8	1	7	11	13	4	7	8	9	10
w	1	1	3	1	1	1	1	1	1	2	2	1	1

(39)

One of the risk factors to be assessed was maternal body weight with a standard deviation of $\sigma = 32$ among women with normal birth weight infants (the controls). Given the above distribution of cases and controls, based on (22), $I(\theta_0)$ is provided by

$$I(\theta_0) = \sum_{i=1}^N V_{0i} = \sigma^2 \sum_{i=1}^N d_i \left[\frac{\binom{n_i}{d_i} - 1}{\binom{n_i}{d_i}} \right] = (32)^2 (51.2615) = 52491.8 \quad (40)$$

and the power of the study to detect a log odds ratio (θ) per lb maternal weight difference is provided by (21). Examination over a range of values for θ shows that the study provided 90% power to detect an odds ratio of 0.986 per unit change in weight or $\theta = \ln(0.986) = -0.0141$.

4.2 A Qualitative Covariate

Another variable assessed by Hosmer and Lemeshow [19] is a maternal history of urinary tract infection. Approximately 10% of women with normal births have a history of UTI ($\pi_2 = 0.1$). Assume that it was desired to detect a doubling of this risk among the cases who give birth to low birth-weight infants ($\pi_1 = 0.2$), such that $\pi_0 = 0.15$. Thus

$\sigma_0^2 = (0.15)(0.85) = 0.1275$. Given the mixtures of (d_i, m_i) presented in (39), and the sum of the combinatorial terms in (40), then $I(\theta_0) = (0.1275)(51.2615) = 6.536$. Examination over a range of values for θ shows that the study provided 90% power to detect an odds ratio of 3.55 for those with a history of UTI versus not, or $\theta = \ln(3.55) = 1.27$.

Walter [5] presents an example of a matched case control study with 104 matched sets, each with a single case, 27 containing a single control and 77 containing two controls. Using the sample estimate of a discordant control pair from the sets with two controls, it is estimated that $\hat{\psi}^* = 0.385$. For a design with 104 matched sets with $\sum_i m_i = 181$ and $\sum_i m_i^2 = 355$, Walter's (37) provides an estimate of power = 0.57 to detect a difference of $\Delta = 0.1$ in the probability of a positive covariate value among cases versus controls.

To examine the agreement with the more general expressions herein, power was calculated using (38) and Walter's (37) over a range of values $0.15 \leq \pi_1 \leq 0.55$ with corresponding $\pi_2 = \pi_1 - 0.1$, and $\pi_0 = \pi_1 - 0.05$. For all of these cases, since $\Delta = 0.1$, Walter's equation yields power of 0.57. The following are the corresponding odds ratios for each value of π_0 , and the corresponding level of power:

π_0	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
OR	3.353	2.250	1.889	1.714	1.615	1.556	1.519	1.500	1.494
Power	0.899	0.754	0.657	0.593	0.550	0.521	0.502	0.492	0.488

(41)

As π_0 decreases below 0.25 (or increases above 0.75) the odds ratio for a fixed $\Delta = 0.1$ increases exponentially and so does power. Walter's method does poorly in this case.

Similar results are obtained for other values of Δ , Walter's method underestimating power for $\pi_0 < 0.25$ (> 0.75), and increasingly overestimating power as π_0 approaches 0.5.

5. DISCUSSION

This paper describes the power of the score test in the conditional logistic regression model, or the Cox proportional hazards model with the Cox adjustment for ties and stratified sets, that can be applied to a study with any numbers of cases and controls or to the case where sets are sampled from a distribution of cases and controls within sets. For example, suppose that the distribution of (d_i, m_i) for the Hosmer-Lemeshow study presented in (39) represents the population of possible sets of cases and controls. From (40), the mean contribution among sets to $I(\theta_0)$ is $51.2615/17 = 3.0154$. Then the above expressions could be generalized to provide the power of a study with N sets sampled at random from this distribution. For example, for the test of the effect of a history of urinary tract infection, then the value of $I(\theta_0)$ with $N = 25$ such sets would be $I(\theta_0) = 25 * (0.1275)(3.0154) = 9.6115$ that would provide 90% power to detect an odds ratio of 2.845. This strategy could be employed when the distribution of numbers of cases and controls is unknown but the mean numbers, and approximately the mean factor, can be specified.

In many cases, such as the study described by Hosmer and Lemeshow, the objective is to evaluate multiple risk factors simultaneously, or the effect of one risk factor adjusted for other covariates. In this case, the score test is a C_a test. In principle, the power function for this test could also be derived for a given covariance matrix of the parameter estimates evaluated under the null hypothesis for the one factor of interest. This would be algebraically and computationally tedious.

Alternately, it seems reasonable to employ a deflation factor in conjunction with a univariate assessment. Let $\widehat{\beta}_x$ denote the estimate of the coefficient for the risk factor of interest X in a univariate unadjusted model with information function $I(\widehat{\beta}_x)$ and corresponding variance of the estimate $V(\widehat{\beta}_x)$. Then let $\widehat{\beta}_{x|z}$ denote the coefficient estimate in a model adjusted for a vector of other covariates $\mathbf{Z}$, where $R^2_{x|z}$ is the coefficient of determination for the regression of X on $\mathbf{Z}$. Hsieh and Lavori [16] then obtain the covariate adjusted variance as

$$V(\widehat{\beta}_{x|z}) = V(\widehat{\beta}_x) [1 - R^2_{x|z}]^{-1} = (I(\widehat{\beta}_x) [1 - R^2_{x|z}])^{-1}. \quad (42)$$

Likewise, the expressions herein would be modified to employ $I(\theta_0) [1 - R^2_{x|z}]$. This approach was also recently shown by Bernardo et al. [20] to apply to the power function of the score test in an exponential regression model for survival data.

For example, in the Hosmer-Lemeshow study, there is 90% power to detect an odds ratio of 0.986 per unit increase in weight in a univariate analysis. In the multivariate analysis in [3], $R^2 = 0.0634$ for the regression of maternal weight on the other covariates. Then (40) would be modified to yield

$$I(\theta_0) = (32)^2 (51.2615) (1 - 0.0634) = 49163.8 \quad (43)$$

that in turn yields 87.8% power to detect an odds ratio of 0.986 per unit increase in weight in the adjusted analysis. If weight had a higher intercorrelation with other covariates in the model, such as $R^2_{x|z} = 0.3$, power would be reduced to 77%.

Recently Schoenfeld and Borenstein [21] describe calculation of the precise non-centrality parameter for the Wald test of a subset of coefficients in a multivariate logistic or PH model that employs a vector of quantitative covariates that are assumed to be distributed as multivariate normal. In simulations they show that the above variance inflation adjustment is not accurate when one or more of the adjusting covariates has a strong effect on the response. However, in a case-control study this adjustment may be fine if it can be assumed that any covariate with a large effect will be controlled by the matching.

An alternative approach to the design of a case-control study embedded within a prospective cohort study is a case-cohort design, originally proposed by Prentice [22], wherein the cases are compared to a pre-selected random sub-cohort of the main cohort. In this case, however, the standard PH model can not be applied. Self and Prentice [23] describe a model that provides unbiased estimates of the coefficients and the variance of the estimates. Therneau and Li [24] show that standard programs for the Cox PH model can be “tricked” into providing the Self-Prentice estimates, among others that have been proposed. There has been no derivation of the expressions for the power of this analysis as a function of the number of cases and the size of the sub-cohort. Chen and Lo [25], however, evaluated the relative efficiency of the case-cohort design versus the standard case-control design and concluded that the two designs with equivalent total numbers of cases and controls (i.e. the sub-cohort) were equally efficient. This suggests that the computations presented herein for a nested case-control study can also be used to evaluate the sample size and power for a case-cohort study with the same number of cases and a sub-cohort with the same size as the combined set of controls.

Acknowledgments

This work was partially supported by funding from the National Institute of Diabetes, Digestive and Kidney Diseases.

REFERENCES

1. Breslow NE. Covariance adjustment of relative-risk estimates in matched studies. *Biometrics*. 1982; 38:661–672. [PubMed: 7171694]
2. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. 1947; 12:153–157. [PubMed: 20254758]
3. Lachin, JM. *Biostatistical Methods: The Assessment of Relative Risks*. Wiley; New York: 2000.
4. Miettinen OS. Individual matching with multiple controls in the case of all-or-none responses. *Biometrics*. 1969; 25:339–55. [PubMed: 5794104]
5. Walter SD. Matched case-control studies with a variable number of controls per case. *Applied Statistics*. 1980; 29:172–9.
6. Ury HK. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. *Biometrics*. 1975; 31:643–9. [PubMed: 1100136]
7. Taylor JMG. Choosing the number of controls in a matched case-control study, some sample size, power and efficiency considerations. *Statistics in Medicine*. 1986; 5:29–36. [PubMed: 3961313]
8. Lui K-J. Estimation of sample sized in case-control studies with multiple controls per case: Dichotomous data. *American Journal of Epidemiology*. 1988; 127:1064–1070. [PubMed: 3358407]
9. Ejigou A. Power and sample size for matched case-control studies. *Biometrics*. 1996; 52:925–33.
10. Sinha S, Mukherjee. A score test for determining sample size in matched case-control studies with categorical exposure. *Biometrical Journal*. 2006; 48:35–53. [PubMed: 16544811]
11. Liddell FDK, McDonald JC, Thomas DC. Methods for cohort analysis: Appraisal by application to asbestos mining. *Journal of the Royal Statistical Society, Series A*. 1977; 140:469–483.

12. Thomas DC, Liddel FDK, McDonald JC, Thomas DC. Methods for cohort analysis: Appraisal by application to asbestos mining. *Journal of the Royal Statistical Society, Series A*. 1977; 140:483–485. Addendum to the paper by.
13. Cox DR. Regression models and life-tables (with discussion). *J. Roy. Statist. Soc., B*. 1972; 34:187–220.
14. Goldstein L, Langholz B. Asymptotic theory for nested case control sampling in the Cox regression model. *Annals of Statistics*. 1992; 20:1903–28.
15. Schoenfeld D. Sample-size formula for the proportional-hazards regression model. *Biometrics*. 1983; 39:499–503. [PubMed: 6354290]
16. 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]
17. Nathan DM, Cleary PA, Backlund JC, Genuth SM, Lachin JM, Orchard TJ, Raslin P, Zinman B. Intensive Diabetes Treatment and Cardiovascular Disease in Patients with Type 1 Diabetes. *The New England Journal of Medicine*. 2005; 353:2643–2653. [PubMed: 16371630]
18. Blankenberg S, et al. for the HOPE Study Investigators. Comparative impact of multiple biomarkers and N-Terminal Pro-Brain Natriuretic Peptide in the context of conventional risk factors for the prediction of recurrent cardiovascular events in the Heart Outcomes Prevention Evaluation (HOPE) Study. *Circulation*. 2006; 114:201–208. [PubMed: 16831981]
19. Hosmer, DW.; Lemeshow, S. *Applied Logistic Regression*. John Wiley; New York: 1989.
20. Bernardo MPV, Lipsitz SR, Harrington DP, Catalano PJ. Sample size calculation for failure time random variables in non-randomized studies. *The Statistician*. 2000; 49:31–40.
21. Schoenfeld DA, Borenstein M. Calculating the power or sample size for the logistic and proportional hazards models. *J Stat Computation and Simulation*. 2005; 75:771–85.
22. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
23. Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann Statist*. 1988; 16:64–81.
24. Therneau TM, Li H. Computing the Cox model for case-cohort designs. *Lifetime Data Analysis*. 1999; 5:99–112. [PubMed: 10408179]
25. Chen K, Lo S-H. Case-cohort and case-control analysis with Cox's model. *Biometrika*. 1999; 86:755–764.