SAMPLE SIZE TO TEST FOR INTERACTION BETWEEN A SPECIFIC EXPOSURE AND A SECOND RISK FACTOR IN A PAIR-MATCHED CASE-CONTROL STUDY

Peihua Qiu[1], Melvin L. Moeschberger[2], Glen E. Cooke[3] and Pascal J. Goldschmidt-Clermont[3]

School of Statistics[1], University of Minnesota, 206 Church St. SE, Minneapolis, MN 55455

Biostatistics Program[2], Ohio State University, 320 West 10th Ave., Columbus, OH 43210

Heart and Lung Institute[3], Ohio State University, 420 West 12th Ave., Columbus, OH 43210

Abstract

We discuss a sample size calculation for a pair-matched case-control study to test for interaction between a specific exposure and a second risk factor. The second risk factor could be either binary or continuous. An algorithm for the calculation of sample size is suggested which is based on a logistic regression model that relates the logarithm of the disease-exposure odds ratio to the second risk factor. This problem is motivated by a study comparing the prevalence of GP-IIIa Pl^{A2} polymorphism (the exposure) in individuals with and without myocardial infarction (case-control). One of the hypotheses in this study is whether or not there is an interaction between the prevalence of GP-IIIa Pl^{A2} polymorphism and a second risk factor such as smoking status and homocysteine level. We introduce the algorithm in detail with several numerical examples.

Key Words: Pair-matched case-control study; Sample size calculation; Logistic regression model; Conditional probability; Odds ratio; Interaction; Score test; Pl^{A2}; Platelet; Risk factors; Ischemic heart disease.

1 Introduction

Case-control studies provide a research method for scientists to investigate possible factors that are associated with disease[1,2]. In order to control for confounding factors, each case is often matched with one or more controls on the basis of similarity with respect to the confounding variables. These kind of designs are called pair-matched case-control studies.

[1]Author for correspondence. E-mail: qiu@stat.umn.edu

In a recent case-control study (71 cases and 68 controls), we found that the Pl^{A2} polymorphism of the platelet glycoprotein (GP)IIIa was associated with the myocardial infarction[3]. The prevalence of Pl^{A2} positivity in individuals admitted to the coronary care unit of the Johns Hopkins Hospital with a diagnosis of myocardial infarction was 40%, twice that of the control group admitted to the Johns Hopkins Hospital, but with no history of vascular disease, either coronary artery disease, stroke or peripheral vascular disease. This study is preliminary in that the sample size is small and possible confounding factors such as smoking status or cholesterol were not taken into account. We are currently considering to undertake a well-designed multiple-center study to confirm our initial findings. Based on some considerations such as the availability of the proper patients and other important factors that may affect the study, we have decided to adopt a pair-matched case-control study with cases and controls pair-matched for some possible confounding factors including age, race, gender, geographic area, smoking and total cholesterol. Coronary artery disease and coronary thromboembolic events result from the conspiracy of environmental factors such as smoking, a saturated fat rich diet, or a sedentary lifestyle, and genetic factors that predispose individuals for these events. It is therefore of great interest for the investigators to test for the interaction effects between Pl^{A2} (the exposure) and other risk factors such as smoking status, homocysteine level and other relevant polymorphisms.

Sample size determination is a fundamental component in the design of a pair-matched case-control study. When the exposure variable is binary, McNemar's[4] test of the disease-exposure association is classic. It is based only on the discordant pairs $(+-)$ and $(-+)$ where $(+-)$ denotes case exposed and control unexposed pairs and $(-+)$ denotes case unexposed and control exposed pairs. Sample size determination based on the McNemar's test is discussed by several authors including Miettinen[5], Duffy[6], Connett *et al.*[7], Dupont[8], Fleiss and Levin[9] and Lachin[10]. Royston[11] compared some of these methods and also gave some guidelines for practitioners in using these methods and in choosing the parameters of the algorithms.

In the Pl^{A2} heart disease study described in the introduction, it was desired to test the interaction between the exposure and a second risk factor. When the design is not pair-matched, several authors have discussed the sample size estimation based on an either additive or multiplicative model for joint effects[12,13,14]. When the design is pair-matched, however, we have not found a method that would help us with the sample size determination for detecting such an interaction.

We suggest an algorithm to determine the sample size for a pair-matched case-control study to detect the interaction between the exposure variable and another risk factor. The exposure variable considered is binary but the second risk factor could be either binary or continuous. The algorithm is first described in Section 2 for the case that the second risk factor is a matching variable. Determination of some parameters used in the algorithm is discussed in Section 3. In Section 4 we discuss the case when the second risk factor is not a matching variable. Some numerical examples are given in Section 5 to demonstrate the algorithm in detail. It is applied to the Pl^{A2} heart disease study in Section 6. Some remarks conclude the article in Section 7. Mathematical derivation of a formula is included in Appendix A.

2 When the Second Risk Factor is a Matching Variable

Let x_{10} and x_{01} denote the numbers of $(+-)$ and $(-+)$ pairs, respectively. Then $m = x_{10}+x_{01}$ is the number of discordant pairs. Conditioning on m discordant pairs, let π_{10}^c represent the conditional probability of a $(+-)$ pair. When cases and controls are matched on the second risk factor Y, we could use the following logistic regression model to relate π_{10}^c to Y:

$$\log(\frac{\pi_{10}^c}{1-\pi_{10}^c}|Y) = \delta + (Y - Y_0)\theta \tag{2.1}$$

In (2.1), Y_0 is the baseline level of Y. If the exposure status of a case is uncorrelated with its matched control, $\frac{\pi_{10}^c}{1-\pi_{10}^c}$ is just the disease-exposure odds ratio. The coefficient δ represents the log odds ratio when the risk factor Y is at its baseline level. One reason the baseline level of Y is used in (2.1) is that users often have some preliminary information about the disease-exposure odds ratio at some specific level of Y. Then this level could be used as the baseline level and that information could be used to estimate δ. The coefficient θ represents the change of the log odds ratio (that is the log ratio of the odds ratios) when Y increases one unit. When $\theta = 0$, the odds ratio does not depend on Y. In other words, there is no interaction effect between the exposure and the second risk factor Y on the disease. This can also be explained by Figure 2.1.

In Figure 2.1, triangles and squares denote the conditional probabilities of the $(-+)$ and the $(+-)$ pairs at $Y = y_1$ and $Y = y_2$ levels, respectively. Since the distances between $(-+)$ and $(+-)$ on the abscissa of Figures 2.1(a) and 2.1(b) are arbitrary, we can make them 1. In this case, the slopes of the lines are the odds ratios. In plot (a), the odds ratio depends on the level of the second

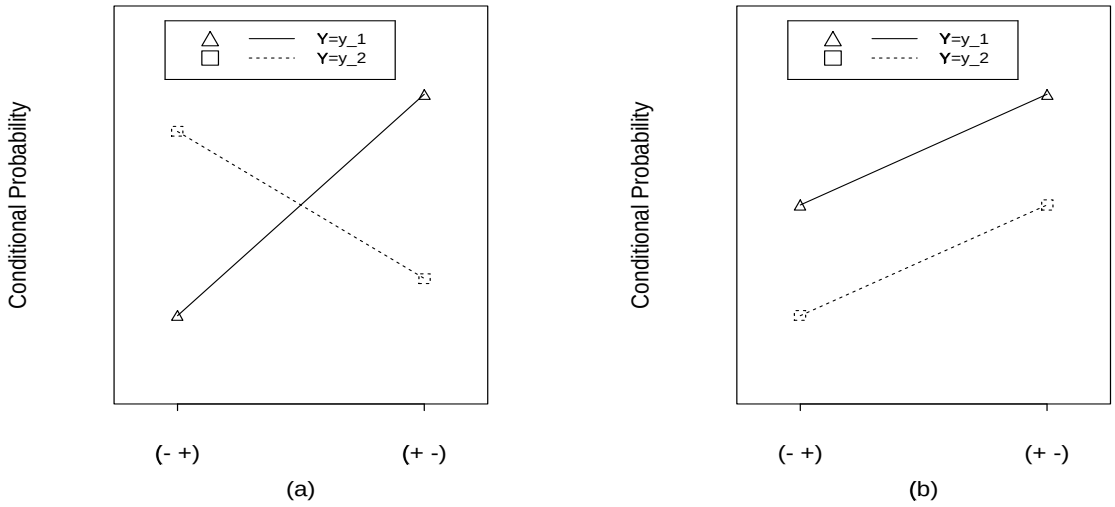

Figure 2.1: Conditional probabilities of the $(-+)$ and the $(+-)$ pairs conditioning on m discordant pairs. Triangles represent the conditional probabilities at $Y = y_1$. Squares denote the conditional probabilities at $Y = y_2$. (a) There is an interaction effect between the exposure and the second risk factor Y on the disease. (b) There is no interaction effect.

risk factor Y. That implies an interaction effect between the exposure and the second risk factor Y on the disease. Plot (b) depicts the case of no interaction effect.

Therefore θ is a measurement of the interaction and it may be interesting to test

$$H_0 : \theta = 0 \ \text{ vs } \ H_a : \theta > 0 \tag{2.2}$$

if we are interested in knowing if the risk factor Y will positively affect the exposure-disease association or not. Later, $H_a : \theta < 0$ and $H_a : \theta \neq 0$ will be discussed.

Suppose that the sample of Y corresponding to the m discordant pairs is $\{Y_1, \ldots, Y_{x_{10}}; Y_{x_{10}+1}, \ldots, Y_m\}$ and the conditional distributions (conditioning on the m discordant pairs) of Y in the $(+-)$ and $(-+)$ populations have means and variances (μ_1, σ_1^2) and (μ_0, σ_0^2), respectively. Then the Score test (see e.g., Section 1.3, Hosmer and Lemeshow[15]) rejects H_0 of (2.2) if

$$\frac{\sum_{i=1}^{x_{10}} Y_i - x_{10}\bar{Y}}{\sqrt{x_{10} S_Y^2 x_{01}/(m-1)}} > Z_\alpha \tag{2.3}$$

where $\bar{Y}$ and S_Y^2 are the sample mean and the sample variance of $\{Y_1, \ldots, Y_{x_{10}}; Y_{x_{10}+1}, \ldots, Y_m\}$. Using computations similar to those described by Lubin *et al.*[16], the power function of the score test is

$$power(\theta, x_{10}, m) = 1 - \beta = 1 - \Phi\left(\frac{Z_\alpha \sigma_Y \sqrt{x_{10} x_{01}/m}}{\tau} - \eta/\tau\right) \tag{2.4}$$

where

$$\begin{aligned}\eta &= \frac{x_{10}x_{01}}{m}(\mu_1-\mu_0)\\ \tau^2 &= \frac{x_{10}x_{01}}{m^2}(x_{01}\sigma_1^2+x_{10}\sigma_0^2)\\ \sigma_Y^2 &= \frac{1}{m}(x_{10}\sigma_1^2+x_{01}\sigma_0^2)+\frac{x_{10}x_{01}}{m^2}(\mu_1-\mu_0)^2.\end{aligned}$$

Please notice that although θ does not appear in the right hand side of (2.4), it is actually hidden in μ_1, μ_0, σ_1^2 and σ_0^2 (cf. equations (3.2) and (3.3) in Section 3 and the related discussions).

Conditioning on the m discordant pairs, x_{10} could be regarded as a random variable with $binomial(m, \pi_{10}^c)$ distribution. Then the expected power of (2.4) with respect to x_{10} is

$$avgpower(\theta, m) = E_{x_{10}}(power(\theta, x_{10}, m)) = \sum_{i=0}^{m} power(\theta, i, m) \times binomial(i, m, \pi_{10}^c) \tag{2.5}$$

As some authors already pointed out (e.g., Royston[11]), m itself can be regarded as a random variable with $binomial(n, \pi_d)$ distribution where n is the sample size (total number of pairs) and π_d is the probability of discordant pairs. We then can define the expected power function with respect to both x_{10} and m as

$$Eavgpower(\theta, n) = E_m(avgpower(\theta, m)) = \sum_{i=0}^{n} avgpower(\theta, i) \times binomial(i, n, \pi_d) \tag{2.6}$$

Then, for given power $1-\beta$ and alternative $\theta = \theta_1 > 0$, the sample size could be chosen as the smallest n such that

$$Eavgpower(\theta_1, n) \geq 1-\beta. \tag{2.7}$$

We denote this sample size as n_{UC}.

From (2.5), we can determine the number of discordant pairs as the smallest m such that $avgpower(\theta_1, m) \geq 1-\beta$. Then the sample size can be defined as $n_C = m/\pi_d$. The value of n_C corresponds to the conditional formula of the sample size of a pair-matched case-control study with one risk factor (the exposure) only. The value of n_{UC} is related to the unconditional formula. In the usual pair-matched case-control study without considering the interaction, Royston[11] recommended the unconditional formula based on his simulation results. We will compare n_{UC} and n_C in the presence of interaction with a simulation example in Section 5.

If users are interested in testing $H_a : \theta < 0$ instead of the one in (2.2), then (2.4) should be

changed to

$$power(\theta, x_{10}, m) = \Phi(\frac{-Z_\alpha \sigma_Y \sqrt{x_{10}x_{01}/m}}{\tau} - \eta/\tau).$$

If the two-sided alternative hypothesis $H_a : \theta \neq 0$ is considered, then (2.4) needs to be modified to the following equation:

$$power(\theta, x_{10}, m) = 1 - \Phi(\frac{Z_{\alpha/2} \sigma_Y \sqrt{x_{10}x_{01}/m}}{\tau} - \eta/\tau) + \Phi(\frac{-Z_{\alpha/2} \sigma_Y \sqrt{x_{10}x_{01}/m}}{\tau} - \eta/\tau).$$

In such cases, other equations (2.5)-(2.7) are kept unchanged.

3 Several Notes about the Sample Size Calculation

In this section we make some notes about the selection of some parameters used in the sample size calculation. Firstly, when Y is at its baseline level Y_0, δ in (2.1) has several equivalent expressions.

$$\delta = \log(\frac{\pi_{10}^c}{1-\pi_{10}^c}|Y_0) = \log(\frac{\pi_{10}}{\pi_{01}}|Y_0) = \log(\psi_{Y_0}) \tag{3.1}$$

where π_{10} and π_{01} denote unconditional probabilities of a $(+-)$ pair and a $(-+)$ pair, respectively, and ψ_{Y_0} represents the disease-exposure odds ratio at the baseline level of Y, Y_0. Therefore δ can be determined as long as one of $\pi_{10}^c, \pi_{10}/\pi_{01}$ and ψ_{Y_0} is known at $Y = Y_0$.

In (2.4), the power function $power(\theta, x_{10}, m)$ depends on the parameters μ_0, σ_0^2, μ_1 and σ_1^2, which are unknown in many cases and therefore need to be estimated beforehand. Model (2.4) is similar to the model (3) of Lubin *et al.*[16] although the latter paper described a totally different situation in that the exposure was continuous and no other risk factors were involved. Lubin *et al.*[16] made some suggestions on calculating similar parameters used in their algorithm. Here we adapt two of these suggestions to our case in order to estimate $\mu_0, \sigma_0^2; \mu_1, \sigma_1^2$.

(i) If one knows the conditional distribution of Y, $F(y)$, and its density $f(y)$ conditioning on the m discordant pairs, then for $i = 0, 1$,

$$\mu_i = \frac{\int y f(y) P(D=i|y)\, dy}{\int f(y) P(D=i|y)\, dy} \tag{3.2}$$

$$\sigma_i^2 = \frac{\int y^2 f(y) P(D=i|y)\, dy}{\int f(y) P(D=i|y)\, dy} - \mu_i^2 \tag{3.3}$$

where $(D = 1)$ denotes $(+-)$ pair and $(D = 0)$ denotes $(-+)$ pair. $P(\cdot)$ is the conditional probability conditioning on the m discordant pairs. Hence $P(D = 1|y) = \pi_{10}^c$ given $Y = y$

and $P(D = 0|y) = \pi_{01}^c$ given $Y = y$. The conditional probability π_{10}^c given $Y = y$ could be obtained from model (2.1).

(ii) If $F(y)$ is unknown but we have a random sample of the second risk factor, $\{Y_1, \ldots, Y_r\}$, corresponding to r discordant pairs, then the integrations in (3.2) and (3.3) can be replaced by the corresponding sample means. For example, the numerator of (3.2) can be replaced by $\{\sum_{j=1}^r Y_j P(D = i|Y_j)\}/r$.

In (2.5), $avgpower(\theta_1, m)$ depends on the parameter π_{10}^c which needs to be specified. From (2.1),

$$\pi_{10}^c = \frac{\exp[\delta + (Y - Y_0)\theta_1]}{1 + \exp[\delta + (Y - Y_0)\theta_1]} \tag{3.4}$$

which depends on the value of Y. To get rid of this dependence on Y, a natural way is to replace it by its expectation with respect to Y. That is, π_{10}^c in (3.4) is estimated by either

$$\int \frac{\exp[\delta + (y - Y_0)\theta_1]}{1 + \exp[\delta + (y - Y_0)\theta_1]} f(y)\, dy \tag{3.5}$$

if the density $f(y)$ of Y is known; or

$$\frac{1}{r} \sum_{i=1}^{r} \frac{\exp[\delta + (Y_i - Y_0)\theta_1]}{1 + \exp[\delta + (Y_i - Y_0)\theta_1]} \tag{3.6}$$

if we have a sample $\{Y_1, \ldots, Y_r\}$ of Y instead.

Finally, π_d in (2.6) which is the probability of a discordant pair needs to be specified. We first notice that model (2.1) relates the *ratio* of the conditional probabilities of the $(+-)$ and $(-+)$ pairs to the risk factor Y. It describes how the two groups of discordant pairs are distributed among the total m discordant pairs. This model is not helpful in estimating π_d. If there is some preliminary information about π_d at various Y levels, then this information can be used to estimate π_d in the same spirit as (3.2) and (3.3). Often this preliminary information is not available, therefore π_d needs to be estimated in the conventional way as follows.

When the disease-exposure association is our only concern and no other risk factors are involved, there has been much discussion in the literature about the determination of π_d. If the exposure status of a case is assumed to be independent of the exposure status of its matched control, then

$$\pi_d = p_0 q_1 + p_1 q_0 \tag{3.7}$$

where $p_0(=1-q_0)$ denotes the estimated proportion of exposed controls in the target population and $p_1(=1-q_1)$ is the estimated proportion of exposed cases (cf. e.g., Section 6.6, Schlesselman[2]). The quantity p_1 can be calculated by $p_1 = \psi p_0/(q_0 + \psi p_0)$ where ψ is the disease-exposure odds ratio. As in (3.1), $\psi = \pi_{10}^c/(1-\pi_{10}^c)$ and π_{10}^c can be calculated by (3.4)-(3.6).

If the exposure status of the case is correlated with the exposure status of the matched control and we know the odds ratio ω measuring the association between a case's and the matched control's exposure status (ω is defined by $\omega = (\pi_{11}\pi_{00})/(\pi_{10}\pi_{01})$), then Fleiss and Levin[9] suggested the following formula to calculate π_d.

$$\pi_d = (p_0 q_1 + p_1 q_0)\frac{\sqrt{1+4(\omega-1)p_1 q_1}-1}{2(\omega-1)p_1 q_1}.$$

Fleiss and Levin[9] pointed out that (3.7) leads to an underestimation of the required total sample size when the exposure status of the case and its matched control are correlated. A similar modification was given by Dupont[8]. Lachin[10] has some related discussions on this topic.

4 When the Second Risk Factor is not a Matching Variable

Model (2.1) is valid only when Y is a matching variable. Consequently, the sample size calculation is only valid for detecting interaction between the exposure variable and the matching variable. If the second risk factor is not a matching variable, then we could replace Y in (2.1) by $Y^{(1)} - Y^{(0)}$ where $Y^{(1)}$ and $Y^{(0)}$ denote the second risk factor for case and the matched control, respectively. The idea to use the difference between a case and its matched control for a risk factor in a logistic regression model was first suggested by Breslow *et al.*[17] when they tried to estimate multiple relative risk functions for matched case-control studies. The resulting model can be used when the interaction between the exposure variable and $Y^{(1)} - Y^{(0)}$ is our main concern. If our major concern is to detect an interaction between the exposure variable and $(Y^{(1)}, Y^{(0)})$, then we suggest generalizing (2.1) in the following way:

$$\log(\frac{\pi_{10}^c}{1-\pi_{10}^c}|Y^{(1)}, Y^{(0)}) = \delta + (Y^{(1)} - Y_0^{(1)})\theta^{(1)} + (Y^{(0)} - Y_0^{(0)})\theta^{(0)} \tag{4.1}$$

where $Y_0^{(1)}$ and $Y_0^{(0)}$ are the baseline levels of $Y^{(1)}$ and $Y^{(0)}$, respectively. As in (2.1), the coefficient $\theta^{(1)}$ is a measurement of the interaction between the exposure variable and $Y^{(1)}$, $\theta^{(0)}$ measures the interaction between the exposure variable and $Y^{(0)}$, and these two interaction effects are assumed to be additive.

Therefore it is interesting to test

$$H_0 : \theta^{(1)} = \theta^{(0)} = 0 \text{ vs } H_a : \text{at least one of } (\theta^{(1)}, \theta^{(0)}) \text{ is not } 0 \tag{4.2}$$

if we want to know whether the interaction between the exposure variable and $(Y^{(1)}, Y^{(0)})$ is significant or not. Suppose that the sample of $(Y^{(1)}, Y^{(0)})$ corresponding to the m discordant pairs is $\{(Y_1^{(1)}, Y_1^{(0)}), \ldots, (Y_{x_{10}}^{(1)}, Y_{x_{10}}^{(0)}); (Y_{x_{10}+1}^{(1)}, Y_{x_{10}+1}^{(0)}), \ldots, (Y_m^{(1)}, Y_m^{(0)})\}$. Then the multivariate score test (see e.g., Chen[18]) rejects H_0 of (4.2) if

$$(0, \sum_{i=1}^{x_{10}} Y_i^{(1)} - x_{10}\overline{Y^{(1)}}, \sum_{i=1}^{x_{10}} Y_i^{(0)} - x_{10}\overline{Y^{(0)}}) \frac{(\Lambda'\Lambda)^{-1}}{x_{10}(1-x_{10})/m^2} \begin{pmatrix} 0 \\ \sum_{i=1}^{x_{10}} Y_i^{(1)} - x_{10}\overline{Y^{(1)}} \\ \sum_{i=1}^{x_{10}} Y_i^{(0)} - x_{10}\overline{Y^{(0)}} \end{pmatrix} \geq \chi^2_{2,\alpha} \tag{4.3}$$

where $\overline{Y^{(1)}} = \frac{1}{m}\sum_{i=1}^m Y_i^{(1)}$, $\overline{Y^{(0)}} = \frac{1}{m}\sum_{i=1}^m Y_i^{(0)}$, $\Lambda' = \begin{pmatrix} 1 & \ldots & 1 & 1 & \ldots & 1 \\ Y_1^{(1)} & \ldots & Y_{x_{10}}^{(1)} & Y_{x_{10}+1}^{(1)} & \ldots & Y_m^{(1)} \\ Y_1^{(0)} & \ldots & Y_{x_{10}}^{(0)} & Y_{x_{10}+1}^{(0)} & \ldots & Y_m^{(0)} \end{pmatrix}$, and $\chi^2_{2,\alpha}$ is a $1-\alpha$ quantile of the central χ^2_2 distribution.

Under the assumption that the conditional distributions of $(Y^{(1)}, Y^{(0)})$ in the $(+,-)$ and $(-,+)$ populations have the same covariance matrix $\Sigma = \begin{pmatrix} (\sigma^{(1)})^2 & \rho\sigma^{(1)}\sigma^{(0)} \\ \rho\sigma^{(1)}\sigma^{(0)} & (\sigma^{(0)})^2 \end{pmatrix}$ and different means $(\mu_1^{(1)}, \mu_1^{(0)})$ and $(\mu_0^{(1)}, \mu_0^{(0)})$ conditioning on the m discordant pairs, the power function of the score test (4.3) is derived in Appendix A, which is defined by

$$power(\theta^{(1)}, \theta^{(0)}, x_{10}, m) = P(S > \chi^2_{2,\alpha} \cdot \Delta), \tag{4.4}$$

where S is a $\chi^2_2(\phi)$ random variable with noncentrality parameter

$$\phi = \frac{x_{10}(m-x_{10})}{m}(\mu_1^{(1)} - \mu_0^{(1)}, \mu_1^{(0)} - \mu_0^{(0)})\Sigma^{-1} \begin{pmatrix} \mu_1^{(1)} - \mu_0^{(1)} \\ \mu_1^{(0)} - \mu_0^{(0)} \end{pmatrix},$$

$$\Delta = \frac{\eta^{(1)}\eta^{(0)} + 2\nu^{(1)}\nu^{(0)}\eta^{(10)} - (\nu^{(0)})^2\eta^{(1)} - (\nu^{(1)})^2\eta^{(0)} - (\eta^{(10)})^2}{(1-\rho^2)(\sigma^{(1)})^2(\sigma^{(0)})^2}$$

and

$$\begin{aligned} \nu^{(1)} &= [x_{10}\mu_1^{(1)} + (m-x_{10})\mu_0^{(1)}]/m \\ \nu^{(0)} &= [x_{10}\mu_1^{(0)} + (m-x_{10})\mu_0^{(0)}]/m \\ \eta^{(1)} &= (\sigma^{(1)})^2 + [x_{10}(\mu_1^{(1)})^2 + (m-x_{10})(\mu_0^{(1)})^2]/m \end{aligned}$$

$$\eta^{(0)} = (\sigma^{(0)})^2 + [x_{10}(\mu_1^{(0)})^2 + (m - x_{10})(\mu_0^{(0)})^2]/m$$
$$\eta^{(10)} = \rho\sigma^{(1)}\sigma^{(0)} + [x_{10}\mu_1^{(1)}\mu_1^{(0)} + (m - x_{10})\mu_0^{(1)}\mu_0^{(0)}]/m.$$

We would like to point out that the notes made in Section 3 are also valid here to determine the related parameters used in the algorithm. For example, if we have a sample of r discordant pairs and the corresponding observations of the second risk factor are $\{(Y_1^{(1)}, Y_1^{(0)}), \ldots, (Y_r^{(1)}, Y_r^{(0)})\}$, then $(\mu_1^{(1)}, \mu_1^{(0)}), (\mu_0^{(1)}, \mu_0^{(0)})$ and Σ could be estimated by:

$$\widehat{\mu_i^{(1)}} = \frac{\sum_{j=1}^r Y_j^{(1)} P(D = i|Y_j^{(1)}, Y_j^{(0)})}{\sum_{j=1}^r P(D = i|Y_j^{(1)}, Y_j^{(0)})}, \quad i = 1, 0$$

$$\widehat{\mu_i^{(0)}} = \frac{\sum_{j=1}^r Y_j^{(0)} P(D = i|Y_j^{(1)}, Y_j^{(0)})}{\sum_{j=1}^r P(D = i|Y_j^{(1)}, Y_j^{(0)})}, \quad i = 1, 0$$

$$(\widehat{\sigma^{(i)}})^2 = \frac{1}{r}\sum_{j=1}^r (Y_j^{(i)})^2 - \frac{1}{r}\left[(\widehat{\mu_1^{(i)}})^2 \sum_{j=1}^r P(D = 1|Y_j^{(1)}, Y_j^{(0)}) + (\widehat{\mu_0^{(i)}})^2 \sum_{j=1}^r P(D = 0|Y_j^{(1)}, Y_j^{(0)})\right],$$
$$i = 1, 0$$

$$\hat{\rho} = \frac{\sum_{j=1}^r Y_j^{(1)} Y_j^{(0)} - \widehat{\mu_1^{(1)}}\widehat{\mu_1^{(0)}} \sum_{j=1}^r P(D = 1|Y_j^{(1)}, Y_j^{(0)}) - \widehat{\mu_0^{(1)}}\widehat{\mu_0^{(0)}} \sum_{j=1}^r P(D = 0|Y_j^{(1)}, Y_j^{(0)})}{r\widehat{\sigma^{(0)}}\widehat{\sigma^{(1)}}}.$$

From the above discussion, it is not difficult to include more terms in the right hand side of (4.1) to model interactions between the exposure variable and more than one other risk factors or to model 3-way or higher-way interactions.

5 Numerical Examples

In this section we give some numerical examples of the algorithm for sample size calculation presented in Section 2. We assume that $Y_0 = 0, Y \sim N(0, 1), \psi_{Y_0} = 3.0, \alpha = 0.05$, $1 - \beta = 0.8$ or 0.9, $\theta_1 = \log(2.0)$ or $\log(2.5)$, $p_0 = 0.01$, 0.05, 0.1, 0.2, 0.3, 0.4 or 0.5. For each combination of $1 - \beta, \theta_1$ and p_0, Table 5.1 presents the sample sizes n_C and n_{UC}, where the probability of a discordant pair π_d is calculated from (3.7) (under the assumption that the exposure status of pair is independent of the exposure status of its matched control) and the number of discordant pairs m is determined from (2.5) such that m is the smallest positive integer with $avgpower(\theta_1, m) \geq 1 - \beta$.

Table 5.1 shows that for larger values of θ_1 the sample size needs to be larger. This is reasonable since more information is definitely needed to detect a smaller interaction effect. Table 5.1 also

Table 5.1: Sample sizes n_C and n_{UC}, the probability of discordant pair π_d and the number of discordant pairs m.

		$\theta_1 = \log(2.0)$				$\theta_1 = \log(2.5)$			
	p_0	n_C	n_{UC}	π_d	m	n_C	n_{UC}	π_d	m
$1-\beta = 0.8$	0.01	2132	2154	0.03612	77	1385	1403	0.03468	48
	0.05	475	478	0.16244	77	307	310	0.15672	48
	0.1	271	273	0.28508	77	174	175	0.27705	48
	0.2	175	175	0.44241	77	111	112	0.43386	48
	0.3	150	151	0.51487	77	95	95	0.50897	48
	0.4	146	147	0.52872	77	92	92	0.52606	48
	0.5	154	155	0.5	77	96	97	0.5	48
$1-\beta = 0.9$	0.01	2935	2961	0.03612	106	1904	1916	0.03468	66
	0.05	653	658	0.16244	106	422	423	0.15672	66
	0.1	372	375	0.28508	106	239	239	0.27705	66
	0.2	240	241	0.44241	106	153	153	0.43386	66
	0.3	206	207	0.51487	106	130	130	0.50897	66
	0.4	201	201	0.52872	106	126	126	0.52606	66
	0.5	212	213	0.5	106	132	132	0.5	66

shows that larger sample sizes are needed to detect a fixed degree of interaction effect with more power. As indicated in Table 5.1, the value of m does not depend on p_0. This can also be checked from the formulas given in Section 2. From there we can see that the value of p_0 affects the sample sizes only through the value of π_d. From the table the relationship between π_d and p_0 is not monotonic or symmetric about $p_0 = 0.5$ as we might imagine. It is skewed depending on the values of ψ_{Y_0} and θ_1. When $\psi_{Y_0} = 3$ and $\theta_1 = \log(2.0)$, this relationship is described by Figure 5.1.

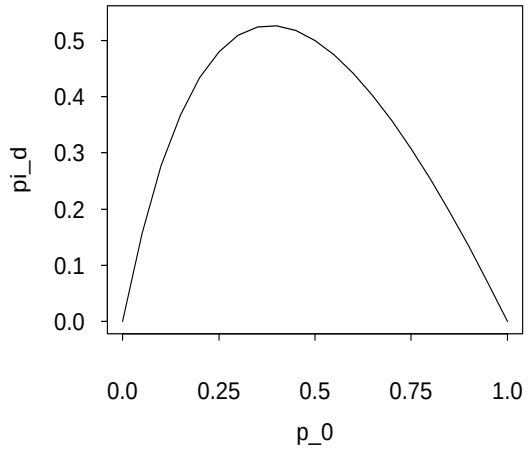

Figure 5.1: The relationship between π_d and p_0 when $\psi_{Y_0} = 3$ and $\theta_1 = \log(2.0)$.

Table 5.1 shows that n_C is smaller than n_{UC} when $p_0 = 0.01$. Since n_{UC} is the smallest sample

size such that the overall averaged power (cf. (2.7)) is bigger than or equal to the specified power, this result is consistent with the conclusion of Royston[11] that the conditional formula of the sample size calculation often underestimates the necessary sample size. However Table 5.1 also shows that the difference between n_C and n_{UC} is not large. Actually when $p_0 \geq 0.05$, the values of n_C and n_{UC} are almost the same.

6 Application to the Pl^{A2} Heart Disease Study

We now return to the Pl^{A2} myocardial infarction study introduced in Section 1. In this study we are interested in testing the interaction effect on the myocardial infarction outcome between the prevalence of the Pl^{A2} polymorphism (the exposure) and a second risk factor. We use smoking status and homocysteine level as two examples of the second risk factor to show the relevance of the method for sample size calculation discussed in this paper.

We first study sample size calculation when the smoking status (Y) is considered as a second risk factor. Since Y is a matching variable, formulas from Sections 2 and 3 will be used. From a preliminary study[3], we know that about 63.7% people in a general population are current or former smokers and the disease-exposure odds ratio is about 2.8. The proportion of exposed cases is between .3 and .5. Based on this information, we use $\psi_{\bar{Y}} = 2.8$, $p_1 = .3$ or .5 and $Y \sim binomial(1, .637)$ in our algorithm. If we are interested in detecting a minimum interaction of $\theta_1 = \log(2.0)$ (that is, the disease-exposure odds ratio of the current or former smokers is 2 times the odds ratio of non-smokers), then the necessary sample sizes determined by our procedure are presented in Figure 6.1. For example if we wish to have 80% or 90% powers with $p_1 = .5$, then the required sample sizes are about 550 and 780, respectively.

Next we consider homocysteine level as a second risk factor which is not a matching variable in this study. By the experimental results presented in Genest *et al.*[19] and Schwartz *et al.*[20], the homocysteine level of cases ($Y^{(1)}$) has mean 13.66 nmol/ml and standard deviation 6.44 nmol/ml. For controls, the mean and standard deviation of the homocysteine levels ($Y^{(0)}$) are 10.93 nmol/ml and 4.92 nmol/ml, respectively. Distributions of $Y^{(1)}$ and $Y^{(0)}$ are slightly skewed to the right (see Genest *et al.*[19]). We found that log-normal distributions $LN(2.514, .448)$ and $LN(2.299, .430)$ provided a reasonable fit to the homocysteine level distributions, where the two values 2.514 and .448 in $LN(2.514, .448)$ are the log-normal parameters corresponding to the mean and standard deviation

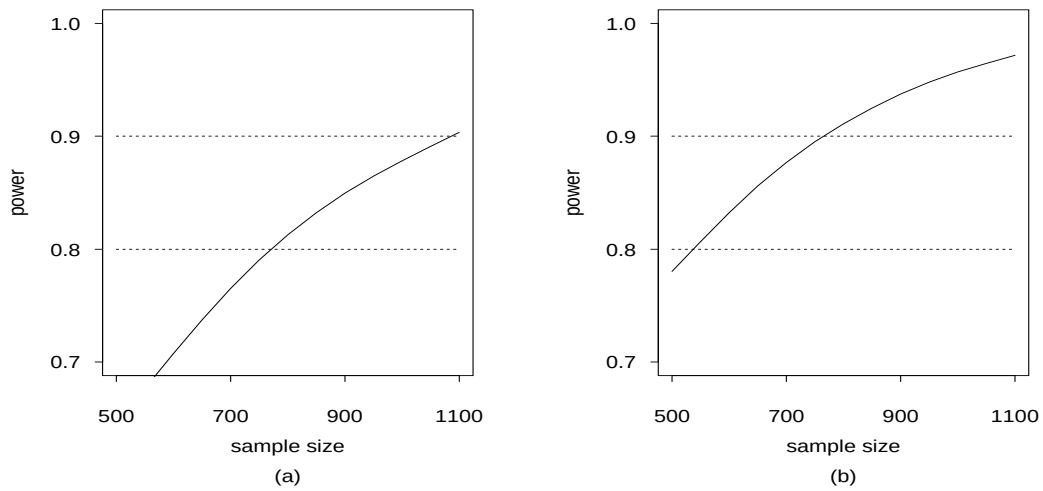

Figure 6.1: Calculated sample sizes for different powers to detect an interaction of $\theta_1 = \log(2.0)$ between Pl^{A2} and the smoking status. (a) $p_1 = .3$; (b) $p_1 = .5$.

of the distribution of $Y^{(1)}$, 13.66 and 6.44, respectively. Similarly, 2.299 and .430 in $LN(2.299, .430)$ corresponds to the mean and standard deviation of $Y^{(0)}$. Based on these informations, we use $\delta = \log(2.8)$, $p_1 = .3$ or .5, $Y^{(1)} \sim LN(2.514, .448)$ and $Y^{(0)} \sim LN(2.299, .430)$ in our algorithm. If we are interested in detecting a minimum interaction of $\theta_1^{(1)} = \log(\sqrt{2})/6.44$ and $\theta_1^{(0)} = \log(\sqrt{2})/4.92$ (that is, the disease-exposure odds ratio increases 2 times if the homocysteine level in both cases and controls increases one standard deviation), then the necessary sample sizes from our procedure are presented in Figure 6.2.

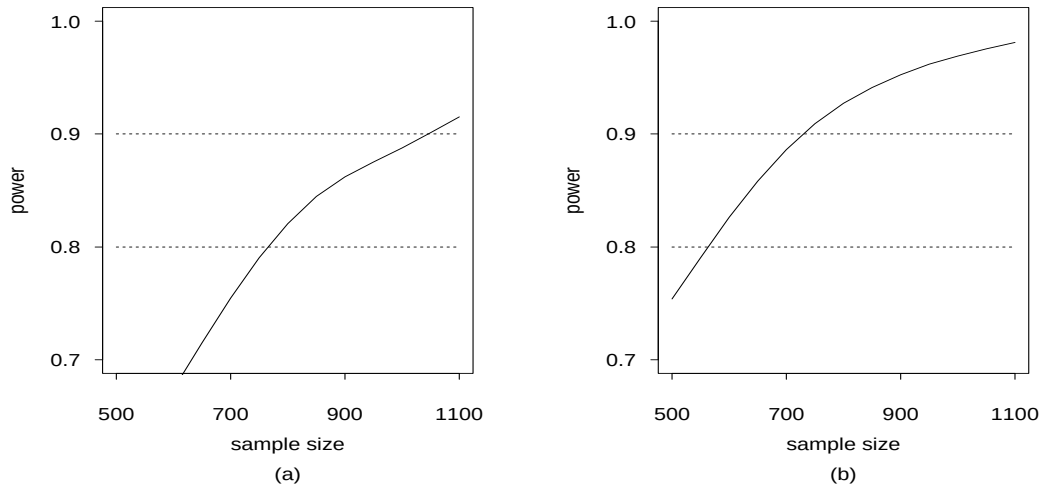

Figure 6.2: Calculated sample sizes for different powers to detect interaction of $\theta_1^{(1)} = \log(\sqrt{2})/6.44$ and $\theta_1^{(0)} = \log(\sqrt{2})/4.92$ between Pl^{A2} and the homocysteine level. (a) $p_1 = .3$; (b) $p_1 = .5$.

7 Concluding Remarks

We have presented an algorithm to calculate the sample size of a pair-matched case-control study to test for interaction between an exposure and a second risk factor. The second risk factor could be a matching variable or a non-matching variable, binary or continuous. This method could also be generalized to include more than one risk factor other than the exposure. Simulations have shown that it gives a reasonable solution.

We are currently concerned about the following problems. First, this article discusses 1-to-1 matching only. In many situations, 1-to-R matching might be preferable. Schlesselman[2] provided an approximate formula for this purpose. He suggested using sample size $(R+1)n/(2R)$ for a 1-to-R matched study if the sample size was calculated to be n based on a 1-to-1 matched pair study. Further research is needed, however, to generalize our algorithm to calculate the exact sample size of a 1-to-R matched case-control study. Second, the exposure variable is assumed to be binary in this paper and this property is used in our modelling. It would be advantageous if the algorithm could be generalized to include the case of a continuous exposure. Third, it will require further research to define methods to accommodate confounding variables in the sample size calculation process.

Appendix

A Derivation of Formula (4.4)

We first notice that

$$(\sum_{i=1}^{x_{10}} Y_i^{(1)} - x_{10}\overline{Y^{(1)}}, \sum_{i=1}^{x_{10}} Y_i^{(0)} - x_{10}\overline{Y^{(0)}}) \sim N(\frac{x_{10}(m-x_{10})}{m}(\mu_1^{(1)} - \mu_0^{(1)}, \mu_1^{(0)} - \mu_0^{(0)}), \frac{x_{10}(m-x_{10})}{m}\Sigma)$$

where

$$\Sigma = \begin{pmatrix} (\sigma^{(1)})^2 & \rho\sigma^{(1)}\sigma^{(0)} \\ \rho\sigma^{(1)}\sigma^{(0)} & (\sigma^{(0)})^2 \end{pmatrix}.$$

Secondly,

$$(0, \sum_{i=1}^{x_{10}} Y_i^{(1)} - x_{10}\overline{Y^{(1)}}, \sum_{i=1}^{x_{10}} Y_i^{(0)} - x_{10}\overline{Y^{(0)}}) \frac{(\Lambda'\Lambda)^{-1}}{x_{10}(1-x_{10})/m^2} \begin{pmatrix} 0 \\ \sum_{i=1}^{x_{10}} Y_i^{(1)} - x_{10}\overline{Y^{(1)}} \\ \sum_{i=1}^{x_{10}} Y_i^{(0)} - x_{10}\overline{Y^{(0)}} \end{pmatrix}$$

$$= (\sum_{i=1}^{x_{10}} Y_i^{(1)} - x_{10}\overline{Y^{(1)}}, \sum_{i=1}^{x_{10}} Y_i^{(0)} - x_{10}\overline{Y^{(0)}}) \frac{|\Sigma|(\Sigma x_{10}(m-x_{10})/m)^{-1}}{\begin{vmatrix} 1 & \overline{Y^{(1)}} & \overline{Y^{(0)}} \\ \overline{Y^{(1)}} & \overline{(Y^{(1)})^2} & \overline{Y^{(1)}Y^{(0)}} \\ \overline{Y^{(0)}} & \overline{Y^{(1)}Y^{(0)}} & \overline{(Y^{(0)})^2} \end{vmatrix}} \begin{pmatrix} \sum_{i=1}^{x_{10}} Y_i^{(1)} - x_{10}\overline{Y^{(1)}} \\ \sum_{i=1}^{x_{10}} Y_i^{(0)} - x_{10}\overline{Y^{(0)}} \end{pmatrix}.$$

Thirdly, we know that

$$(\sum_{i=1}^{x_{10}} Y_i^{(1)} - x_{10}\overline{Y^{(1)}}, \sum_{i=1}^{x_{10}} Y_i^{(0)} - x_{10}\overline{Y^{(0)}})(\Sigma x_{10}(m-x_{10})/m)^{-1} \begin{pmatrix} \sum_{i=1}^{x_{10}} Y_i^{(1)} - x_{10}\overline{Y^{(1)}} \\ \sum_{i=1}^{x_{10}} Y_i^{(0)} - x_{10}\overline{Y^{(0)}} \end{pmatrix} \sim \chi_2^2(\phi)$$

and $\overline{Y^{(1)}} \sim \nu^{(1)}$, $\overline{Y^{(0)}} \sim \nu^{(0)}$, $\overline{(Y^{(1)})^2} \sim \eta^{(1)}$, $\overline{(Y^{(0)})^2} \sim \eta^{(0)}$, and $\overline{Y^{(1)}Y^{(0)}} \sim \eta^{(10)}$. By combining all the above arguments, we can get (4.4).

REFERENCES

1. Breslow, N.E. and Day, N.E. *Statistical Methods in Cancer Research: Volume I. The Analysisof Case-Control Studies,* IARC, Lyon, 1980.

2. Schlesselman, J.J. *Case-Control Studies - Design, Conduct, Analysis,* Oxford University Press, New York, 1982.

3. Weiss, E.J., Bray, P.F., Tayback, M., Schulman, S.P., Kickler, T.S., Becker, L.C., Weiss, J.L., Gerstenblith, G. and Goldschmidt-Clermont, P.J. 'A polymorphism of a platelet glycoprotein receptor as an inherited risk factor for coronary thrombosts', *The New England Journal of Medicine*, **334,** 1090-1094 (1996).

4. McNemar, Q. 'Note on the sampling error of the differences between correlated proportions or percentages', *Psychometrika*, **12,** 153-157 (1947).

5. Miettinen, O.S. 'Individual matching with multiple controls in the case of all-or-none response', *Biometrics*, **25**, 339-355 (1969).

6. Duffy, S.W. 'Asymtotic and exact power for the McNemar test and its analogue with r controls per case', *Biometrics*, **40**, 1005-1015 (1984).

7. Connet, J.E., Smith, J.A. and McHugh, R.B. 'Sample size and power for the pair-matched case-control staudies', *Statistics in Medicine*, **6**, 53-59 (1987).

8. Dupont, W.D. 'Power calculations for matched case-control studies', *Biometrics*, **44**, 1157-1168 (1988).

9. Fleiss, J.L. and Levin, B. 'Sample size determination in studies with matched pairs', *Journal of Clinical Epidemiology*, **41**, 727-730 (1988).

10. Lachin, J.M. 'Power and sample size evaluation for the McNemar test with application to matched case-control studies', *Statistics in Medicine*, **11**, 1239-1251 (1992).

11. Royston, P. 'Exact conditional and unconditional sample size for pair-matched studies with binary outcome: a practical guide', *Statistics in Medicine*, **12**, 699-712 (1993).

12. Greenland, S. 'Tests for interaction in epidemiologic studies: a review and a study of power', *Statistics in Medicine*, **2**, 243-251 (1983).

13. Smith, P.G. and Day, N.E. 'The design of case-control studies: the influence of confounding and interaction effects', *International Journal of Epidemiology*, **13**, 356-365 (1984).

14. Lubin, J.H. and Gail, M.H. 'On power and sample size for studying features of the relative odds of disease', *American Journal of Epidemiology*, **131**, 552-566 (1990).

15. Hosmer, D.W. and Lemeshow, S. *Applied Logistic Regression,* John Wiley and Sons, New York, 1989.

16. Lubin, J.H., Gail, M.H. and Ershow, A.G. 'Sample size and power for case-control studies when exposures are continuous', *Statistics in Medicine*, **7**, 363-376 (1988).

17. Breslow, N.E., Day, N.E., Halvorsen, K.T., Prentice, R.L. and Sabai, C. 'Estimation of multiple relative risk functions in matched case-control studies', *American Journal of Epidemiology*, **108**, 299-307 (1978).

18. Chen, C.F. 'Score tests for regression models," *Journal of the American Statistical Association*, **78**, 158-161 (1983).

19. Genest, J.J., McNamara,J.R., Salem, D.N., Wilson, P.W.F., Schaefer, E.J. and Malinow, M.R. 'Plasma homocysteine levels in men with premature coronary artery disease', *Journal of the American College of Cardiology*, **16,** 1114-1119 (1990).

20. Schwartz, S.M., Siscovick, D.S., Malinow, M.R., Rosendaal, F.R., Beverly, R.K., Hess, D.L., Psaty, B.M., Longstreth, W.T., Koepsell, T.D., Raghunathan, T.E. and Reitsma, P.H. 'Myocardial infarction in young women in relation to plasma total homocysteine, folate, and a common variant in the methylenetetrahydrofolate reductase gene', *Circulation*, **96,** 412-417 (1997).