Biometrical Journal **48** (2006) 1, 35–53 DOI: 10.1002/bimj.200510200

A Score Test for Determining Sample Size in Matched Case-Control Studies with Categorical Exposure

Samiran Sinha[1] and **Bhramar Mukherjee**[*,2]

[1] Department of Statistics, Texas A & M University, College Station, TX-77843, USA
[2] Department of Statistics, University of Florida, Gainesville, FL-32611, USA

Received 16 March 2005, revised 6 October 2005, accepted 21 October 2005

Summary

The paper considers the problem of determining the number of matched sets in $1:M$ matched case-control studies with a categorical exposure having $k+1$ categories, $k \geq 1$. The basic interest lies in constructing a test statistic to test whether the exposure is associated with the disease. Estimates of the k odds ratios for $1:M$ matched case-control studies with dichotomous exposure and for $1:1$ matched case-control studies with exposure at several levels are presented in Breslow and Day (1980), but results holding in full generality were not available so far. We propose a score test for testing the hypothesis of no association between disease and the polychotomous exposure. We exploit the power function of this test statistic to calculate the required number of matched sets to detect specific departures from the null hypothesis of no association. We also consider the situation when there is a natural ordering among the levels of the exposure variable. For ordinal exposure variables, we propose a test for detecting trend in disease risk with increasing levels of the exposure variable. Our methods are illustrated with two datasets, one is a real dataset on colorectal cancer in rats and the other a simulated dataset for studying disease-gene association.

Key words: Chi-squared test; Colon carcinogenesis; Discordant matched sets; Disease-gene association; Eigen values; Non-centrality parameter; Odds-ratio; Ordinal exposure, Power function; Score test; Trend effect.

1 Introduction

In popular epidemiologic study designs, such as case-control, case-cohort, and nested case-control design, often there is a significant amount of cost and time involved in sampling the units, especially for a rare disease. In such situations, one would like to determine the number of sampling units required to detect specific departures from the hypothesis of interest at the planning stages of the experiment (Schlesselman, 1974). For a individually matched case-control design, a case is matched with one or more controls with respect to a set of confounding variables. The design itself is hard to implement as it becomes necessary to determine the appropriate set of controls and collect information on them. For such designs one would like to know the number of matched sets needed to attain desired level of accuracy. In the present article, we propose a method for determining the number of matched sets for $1:M$ ($M \geq 1$) matched case-control studies with a categorical exposure variable having $k+1$, $k \geq 1$ categories.

Though stratification is often introduced in epidemiologic designs to control the effects of confounding, there is only a limited amount of literature on determining the number of matched sets for highly stratified studies with categorical exposure. Analyzing exposure data with the existing cate-

[*] Corresponding author: e-mail: mukherjee@stat.ufl.edu, Phone: +1 352 392 1941 ex 241, Fax: +1 352 392 5175

gories can throw some light on the underlying nature of disease-exposure association which may be obscured if one simply dichotomizes the categorical exposure variable. For example, in genetic association studies, interest lies in the association between a disease and a candidate gene. If one assumes a bi-allelic marker locus for the candidate gene, there would be three possible genotypes (three categories of the exposure variable), and sometimes dichotomization of the genotype may hide the true nature of association specially for the *complex diseases* which do not obey single-gene dominant or single-gene recessive Mendallian law (Jarvik, 1998).

Sample size determination is an important aspect in any comparative study. For individually matched case-control studies with binary exposure, standard formulas for sample size determination is available in Schlesselman (1982). Parker and Bregman (1986) proposed sample size determination strategy for $1:M$ matched case-control studies with a binary exposure variable taking into account the heterogeneity in exposure prevalence among the controls in different matched sets. Dupont (1988) considered sample size selection problems for $1:M$ matched case control studies taking the association for the exposure between matched case and control patients into account. Recently Nam (1992), Nam and Fears (1992a, b), and Nam (1997) considered the same problem for general case-control study designs. Breslow and Day (1980) contains other methods of sample size determination for polychotomous exposure variable in unmatched and $1:1$ matched case-control studies and for binary exposure variable in $1:M$ ($M > 1$) matched case-control study designs. However, to the best our knowledge, there is no work on sample size determination for $1:M$ matched case-control studies with a polychotomous exposure variable.

In order to determine sample size for matched case-control studies with categorical exposure variable one can use likelihood ratio (LR) test, Wald test, or score test (Rao, 1947). Both the LR test and Wald test require calculation of maximum likelihood estimates (MLE) of the model parameters under the alternative hypothesis which in turn needs substantial computational effort for solving a system of nonlinear equations. Therefore, we focus on score test for the hypothesis of our interest, and exploiting the power function of the test we derive the required sample size, that is the number of matched sets required to detect a specified departure from the null hypothesis with certain probability. When there is a natural ordering among the levels of the exposure variable, we propose a test for detecting the presence of a trend in disease risk with increasing levels of the exposure variable and derive the power function of the proposed test. Sample sizes for trend detection are based on this power function. The attractive features of using the score test are two-fold. First, it does not require computation of the MLE's of the parameters under the alternative hypothesis, and second, it has similar first order asymptotic property as Wald and Likelihood ratio test under both the null and Pitman type alternative hypothesis (Serfling, 1980, pp. 156). However one must recognize that inspite of the first order asymptotic equivalence, finite sample properties of these three types of test could be different, depending on the nature of the problem (Lusbader, Moolgavkar and Venzon, 1984). Due to difficulties involved in computing the MLE the score test seems to be a preferable alternative in this situation. The score test has many other attractive and well-known theoretical features and is also most powerful for local alternatives (Cox and Hinkley, 1974, p. 113). As one referee has pointed out, determining sample size by using the score test may serve as a good precursor even when other test statistics are used at the analysis stage.

Two datasets have been analyzed in this paper. We apply the proposed method to these datasets and then perform a small scale simulation study motivated by the examples to determine required sample sizes for various combinations of the odds ratio parameters. The first one is a real dataset on colorectal cancer of rats (Hong et al., 2001). We explore the association between apoptosis and proliferation in stem cells in rats, the data structure is decried in detail in Section 6.1. The second one is a simulated dataset following a real genetic data on allele frequencies of 12 marker loci in Buenos Aires metropolitan population (Sala et al., 1999). Among several marker loci, F13A is chosen as the candidate gene associated with a disease we simulated. We consider the most frequent allele at this marker locus to be the disease causing allele. Using the genotypes at this marker locus as a categorical exposure variable we generate a hypothetical matched case-control dataset and apply our method to test the hypothesis of no association between the disease and the candidate gene. The examples presented in this paper represent two different situations, one when the prevalence value of some exposure categories are very rare and when the prevalence values are moderate. In both the situations we calculate the test statistics and determine sample sizes from given prevalence values of the exposure variable. We also carry out the test for detecting trend in disease risk for both the examples. Both of these examples share a common feature that collecting data on study units is expensive (in example 1, it is through examination of colon cells, in example 2, through genotyping), and thus determining the right sample size can substantially save resources.

Most of the currently available literature on sample size determination in matched studies is for dichotomous exposure. In context of the two data examples we dichotomize the exposure variable, and then compare the sample sizes required by our method and the ad hoc method proposed by Schlesselman (1982) for a set of plausible values of the odds ratio.

The rest of the article is organized as follows. Model and notations are described in Section 2. Test procedures for a general polychotomous exposure is considered in Section 3, while Section 4 contains a special treatment of ordinal variables. In Section 5 we discuss the sample size determination problem. Section 6 contains numerical examples based on the real datasets and some simulation results. Section 7 contains concluding remarks. Details of some formulae are relegated to the Appendix.

Before we conclude this section we highlight some of the main features of this article. First, we propose a test of the hypothesis of independence of the exposure and the disease variable in a $1:M$ matched case-control study with polychotomous exposure variable. For an ordinal exposure variable we propose a test for detecting a trend in disease risk with increasing levels of the exposure variable. Second, we consider the problem of sample size determination for $1:M$ matched case-control studies with polychotomous exposure variable. Finally, one major advantage of the proposed method is that the test statistics are very easy to calculate simply by using the knowledge of the discordant matched sets. Using the distribution of the MLE's obtained from conditional logistic regression which is the classical method of analyzing matched case-control data, to determine the sample size will require more computation.

2 Model and Assumptions

Suppose we have N matched sets and each matched set consists of 1 case and M controls. Let X be a single exposure variable. Assume that the exposure variable has $k+1$ categories, denoted by 0, $1, \ldots, k$. Although we use numerical indices to denote the exposure levels, these categories are in nominal scale, and 0 is assumed to be the baseline category. Let p_{0h} and p_{1h} denote the prevalence of h-th level of the exposure variable among the control and case population respectively, i.e.,

$$p_{0h} = \Pr(X = h \mid D = 0) \quad \text{and} \quad p_{1h} = \Pr(X = h \mid D = 1) .$$

Suppose ψ_h be the ratio of odds of disease in exposure category h to that of the category 0. So

$$\psi_h = \frac{p_{1h} p_{00}}{p_{10} p_{0h}}, \quad \text{for} \quad h = 0, 1, \ldots, k . \tag{1}$$

By definition $\psi_0 = 1$. We assume that the odds ratios are the same across the matched sets. The goal is to test the hypothesis H_0: $\psi_1 = \psi_2 = \ldots = \psi_k = 1$, i.e., the exposure variable and the disease outcome are independent. The alternative hypothesis is that at least one of ψ_j, $j = 1, \ldots, k$ is different from 1.

3 Methods

As the formulation for general k is very complicated, we first derive the case for $k = 2$. Following the same structure of argument, we derive the formula for any general k in Appendix 1. We focus our attention only to the discordant matched sets as the concordant matched sets are noninformative regarding the parameters of our interest (Breslow and Day (1980), pp 164).

For $1:M$ $(M \ge 1)$ matched study with polychotomous exposure variable the concordant sets are those where case and controls are exposed to the same level of exposure variable, and the discordant sets are defined as the matched sets with at least two distinct levels of the exposure variable being present.

The probability that a matched set has m_1 subjects exposed at the level 1 and m_2 subjects exposed at the level 2 along with the restrictions $m_1 \ge 1$, $m_2 \ge 1$, and $m_1 + m_2 \le M$, is given by,

$$\begin{aligned} p_{m_1,m_2} &= \Pr(\text{case is exposed at the level 0, and } m_1 \text{ and } m_2 \text{ controls are exposed at the}\\ &\qquad \text{level 1 and 2 respectively}) + \Pr(\text{case is exposed at the level 1, and } m_1 - 1\\ &\qquad \text{and } m_2 \text{ controls are exposed at the level 1 and 2 respectively})\\ &\quad + \Pr(\text{case is exposed at the level 2, and } m_1 \text{ and } m_2 - 1 \text{ controls are exposed}\\ &\qquad \text{at the level 1 and 2 respectively})\\ &= (1 - p_{11} - p_{12}) \frac{M!}{m_1! m_2! (M - m_1 - m_2)!} \, p_{01}^{m_1} p_{02}^{m_2} (1 - p_{01} - p_{02})^{M - m_1 - m_2}\\ &\quad + p_{11} \frac{M!}{(m_1 - 1)! \, m_2! (M - m_1 - m_2 + 1)!} \, p_{01}^{m_1 - 1} p_{02}^{m_2} (1 - p_{01} - p_{02})^{M - m_1 - m_2 + 1}\\ &\quad + p_{12} \frac{M!}{m_1! (m_2 - 1)! \, (M - m_1 - m_2 + 1)!} \, p_{01}^{m_1} p_{02}^{m_2 - 1} (1 - p_{01} - p_{02})^{M - m_1 - m_2 + 1} . \end{aligned} \tag{2}$$

Let $T_{m_1,m_2} = n^0_{m_1,m_2} + n^1_{m_1,m_2} + n^2_{m_1,m_2}$ be the total number of matched sets where m_1 subjects are exposed at the level 1 and m_2 are exposed at the level 2 of the exposure variable while $n^j_{m_1,m_2}$ denotes the number of matched sets out of T_{m_1,m_2} where case is exposed at the level j, $j = 0, 1, 2$. Here m_1, $m_2 = 1, 2, \ldots, M$ and $m_1 + m_2 \le M$. Thus conditionally,

$$(n^1_{m_1,m_2}, n^2_{m_1,m_2}) \mid T_{m_1,m_2} \sim \text{Multinomial}\,(T_{m_1,m_2}; p_{1\,|\,m_1,m_2}, p_{2\,|\,m_1,m_2}), \tag{3}$$

where $p_{1|m_1,m_2} = \Pr(\text{case is exposed at the level } 1 \mid m_1 \text{ and } m_2 \text{ subjects are exposed at the levels 1 and 2 respectively}) = m_1\psi_1/\{m_1\psi_1 + m_2\psi_2 + (M - m_1 - m_2 + 1)\}$ and similarly the other conditional probability is $p_{2\,|\,m_1,m_2} = m_2\psi_2/\{m_1\psi_1 + m_2\psi_2 + (M - m_1 - m_2 + 1)\}$.

Let $T_m^{(i,j)}$ be the number of matched sets where m subjects are exposed at the level i and the rest are exposed at the level j of the exposure variable. Here i and j are different and $i, j = 0, 1, 2$; and $m = 1, 2, \ldots, M$. So the total number of discordant matched sets is

$$N_D = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} I(m_1 + m_2 \le M)\, T_{m_1,m_2} + \sum_{m=1}^{M} T_m^{(1,0)} + \sum_{m=1}^{M} T_m^{(2,0)} + \sum_{m=1}^{M} T_m^{(1,2)} , \tag{4}$$

where $I(\cdot)$ is an indicator function. We write $T_m^{(i,j)} = n_m^{i(i,j)} + n_m^{j(i,j)}$, where $n_m^{i(i,j)}$ denotes the number of matched set out of $T_m^{(i,j)}$ where case is exposed at the level i. Further define

$$\begin{aligned} p_m^{(i,j)} &= \Pr(\text{a matched set has } m \text{ subjects exposed at the level } i, \text{ and the rest are}\\ &\qquad \text{exposed at the level } j)\\ &= \Pr(\text{ the case is exposed at the level } i, m - 1 \text{ controls are exposed at the level } i,\\ &\qquad \text{and the rest are exposed at the level } j) + \Pr(\text{ the case is exposed at the level } j,\\ &\qquad m \text{ controls are exposed at the level } i, \text{ and the rest are exposed at the level } j)\\ &= p_{1i} \binom{M}{m-1} p_{0i}^{m-1} p_{0j}^{M-m+1} + p_{1j} \binom{M}{m} p_{0i}^{m} p_{0j}^{M-m} . \end{aligned} \tag{5}$$

Hence conditional on $T_m^{(i,j)}$, $n_m^{i(i,j)}$ follows a binomial distribution with success probability

$$p_m^{i(i,j)} = \frac{p_{1i}\begin{pmatrix} M \\ m-1 \end{pmatrix} p_{0i}^{m-1} p_{0j}^{M-m+1}}{p_{1i}\begin{pmatrix} M \\ m-1 \end{pmatrix} p_{0i}^{m-1} p_{0j}^{M-m+1} + p_{1j}\begin{pmatrix} M \\ m \end{pmatrix} p_{0i}^{m} p_{0j}^{M-m}}$$
$$= \frac{m\psi_i}{m\psi_i + (M-m+1)\psi_j}\,, \quad \text{for} \quad i,j = 0,1,2\,. \tag{6}$$

The conditional probability given in (6) turns into Eq. (5.13) of Breslow and Day (1980) if one assumes only two levels of the exposure variable.

The likelihood function of the odds ratios for the matched case-control data is

$$L \propto \prod_{m_1,\,m_2} (p_{0\,|\,m_1,\,m_2})^{n^0_{m_1,m_2}} (p_{1|m_1,\,m_2})^{n^1_{m_1,m_2}} (p_{2|m_1,m_2})^{n^2_{m_1,m_2}}$$
$$\times \prod_m (p_m^{1(1,2)})^{n_m^{1(1,2)}} (p_m^{2(1,2)})^{n_m^{2(1,2)}} \prod_m (p_m^{1(1,0)})^{n_m^{1(1,0)}} (p_m^{0(1,0)})^{n_m^{0(1,0)}}$$
$$\times \prod_m (p_m^{2(2,0)})^{n_m^{2(2,0)}} (p_m^{0(2,0)})^{n_m^{0(2,0)}}\,. \tag{7}$$

The proposed score statistic for testing H_0 is,

$$S = \boldsymbol{Z}^T D_0^{-1} \boldsymbol{Z}\,, \tag{8}$$

where

$$\boldsymbol{Z}^T = \left.\frac{\partial \log L}{\partial \boldsymbol{\psi}}\right|_{\boldsymbol{\psi}=\mathbf{1}} = (Y_1 - \mu_{10}, Y_2 - \mu_{20})$$

and

$$D_0 = \left(-\left.\frac{\partial^2 \log L}{\partial \boldsymbol{\psi}\,\partial \boldsymbol{\psi}^\tau}\right|_{\boldsymbol{\psi}=\mathbf{1}}\right) = \begin{pmatrix} \sigma_{10}^2 & \sigma_{120} \\ \sigma_{120} & \sigma_{20}^2 \end{pmatrix}. \tag{9}$$

Where

$$Y_1 = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} I(m_1 + m_2 \le M)\, n^1_{m_1,m_2} + \sum_{m=1}^{M} n_m^{1(1,2)} + \sum_{m=1}^{M} n_m^{1(1,0)} \tag{10}$$

and,

$$Y_2 = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} I(m_1 + m_2 \le M)\, n^2_{m_1,m_2} + \sum_{m=1}^{M} n_m^{2(1,2)} + \sum_{m=1}^{M} n_m^{2(2,0)}\,, \tag{11}$$

represent the total number of discordant matched sets where the case is exposed at the level 1 and 2 of the exposure respectively. μ_{10}, μ_{20}, σ_{10}^2, σ_{20}^2 are the mean and variance of Y_1 and Y_2 evaluated under H_0, and σ_{120} is the covariance between Y_1 and Y_2 under H_0. Exact expressions for these quantities are collected in Appendix 2. Note that we can write $\boldsymbol{Y} = (Y_1, Y_2)$ as a sum of independent random variables i.e., $\boldsymbol{Y} = \sum_{i=1}^{N} \boldsymbol{Q}_i$, where $\boldsymbol{Q}_i = (Q_{i1}, Q_{i2})$ and $Q_{ij} = I(i$-th matched set is a discordant set and the case is exposed at the level $j)$, $j = 1, 2$. For finite M, $\boldsymbol{Y}$ asymptotically follows a bivariate normal distribution and under H_0 it has mean (μ_{10}, μ_{20}) and variance covariance matrix D_0. Hence, under the null hypothesis, S asymptotically follows a central Chi-square distribution with 2 degrees of freedom.

We propose the following test function for testing H_0 at level of significance α,

$$\phi = \begin{cases} 1 & \text{if} \quad S > \chi^2_{2,\alpha} \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where $\chi^2_{2,\alpha}$ is the upper α percentile point of central chi-square distribution with 2 degrees of freedom.

Remark 1 Note that for binary exposure variable the test statistic S reduces to

$$S = \frac{\left\{\sum\limits_{m=1}^{M} n_m^{1(1,0)} - \sum\limits_{m=1}^{M} T_m^{(1,0)} m/(M+1)\right\}^2}{\sum\limits_{m=1}^{M} T_m^{(1,0)} m(M-m+1)/(M+1)^2} \tag{13}$$

which is the same as Eq. (5.19) of Breslow and Day (1980) without continuity correction.

As mentioned earlier, maximum likelihood estimators of the odds ratio parameters are obtained by solving a set of nonlinear equations in two variables and hence are not easy to obtain. Therefore, we provide a set of computationally simpler estimates following Mantel and Haenszel (1959),

$$\hat{\psi}_{1MH} = \frac{\sum\limits_{m_1,m_2} (M - m_1 - m_2 + 1)\, n^1_{m_1,m_2} + \sum\limits_{m} (M - m + 1)\, n_m^{1(1,0)}}{\sum\limits_{m_1,m_2} m_1 n^0_{m_1,m_2} + \sum\limits_{m} m n_m^{0(1,0)}}, \tag{14}$$

$$\hat{\psi}_{2MH} = \frac{\sum\limits_{m_1,m_2} (M - m_1 - m_2 + 1)\, n^2_{m_1,m_2} + \sum\limits_{m} (M - m + 1)\, n_m^{2(2,0)}}{\sum\limits_{m_1,m_2} m_2 n^0_{m_1,m_2} + \sum\limits_{m} m n_m^{0(2,0)}}. \tag{15}$$

The potential drawback of the Mantel-Haenszel estimator is that it does not use the information contained in the matched sets where the baseline category (level 0) does not appear, which may entail loss of efficiency.

4 Trend test for Ordinal Exposures

The proposed test statistic (8) is derived for categorical exposure variable with nominal levels. However, if there is some natural ordering in the exposure categories, specially for a quantitative variable, one may want to detect a linear trend in disease risk with increasing levels of the exposure variable, that means, we test $H_0 : \gamma = 0$ against $H_a : \gamma \neq 0$, where $\psi_l = \exp(\gamma x_l)$ and x_l is the value of the exposure variable at l-th level. Without loss of generality one may assume that $x_0 = 0$ and $x_1 < x_2 < \ldots < x_l$. This is equivalent to assuming a prospective model for the disease probability $\operatorname{logit} P(D = 1|x_l) = \alpha_i + \gamma x_l$, where α_i defines the effect of i-th matched set on the disease probability. The above hypothesis can be restated in terms of the odds ratios of the exposure variable as $H_0 : \psi_1 = \ldots = \psi_k = 1$ against $H_a : \psi_k > \ldots > \psi_1 > 1$ (increasing linear trend) or $\psi_k < \ldots < \psi_1 < 1$ (decreasing linear trend). The trend test statistic first appear in the work of Armitage (1955). The score test for detecting presence of a linear trend in (log) of odds ratios ψ_l with increasing levels of the exposure variable is based on the statistic $\sum\limits_{h=0}^{2} x_h N^i_{1h}$ for every matched set, where N^i_{1h} denotes the number of cases exposed at the level x_h of the exposure variable in the i-th matched set. The score statistic for testing H_0 is

$$\chi^2 = \frac{\left(\sum\limits_{i=1}^{N} \sum\limits_{h=0}^{2} x_h N^i_{1h} - e_{0|T}\right)^2}{v_{0|T}} = \frac{(Y_1 x_1 + Y_2 x_2 - e_{0|T})^2}{v_{0|T}}, \tag{16}$$

where

$$
\begin{aligned}
e_{0\,|\,T} = & \frac{1}{M+1}\left\{ \sum_{m_1, m_2} (m_1 x_1 + m_2 x_2)\, T_{m_1, m_2} + x_1 \sum_m m T_m^{(1,0)} + x_2 \sum_m m T_m^{(2,0)} \right. \\
& \left. + \sum_m (m x_1 + (M - m + 1)\, x_2)\, T_m^{(1,2)} \right\}, \\
v_{0\,|\,T} = & \frac{1}{(M+1)}\left\{ \sum_{m_1, m_2} (m_1 x_1^2 + m_2 x_2^2)\, T_{m_1, m_2} + x_1^2 \sum_m m T_m^{(1,0)} + x_2^2 \sum_m m T_m^{(2,0)} \right. \\
& \left. + \sum_m (m x_1^2 + (M - m + 1)\, x_2^2)\, T_m^{(1,2)} \right\} \\
& - \frac{1}{(M+1)^2}\left\{ \sum_{m_1, m_2} T_{m_1, m_2} (m_1 x_1 + m_2 x_2)^2 + x_1^2 \sum_m T_m^{(1,0)} m^2 + x_2^2 \sum_m T_m^{(2,0)} m^2 \right. \\
& \left. + \sum_m (m x_1 + (M - m + 1)\, x_2)^2\, T_m^{(1,2)} \right\}.
\end{aligned}
$$

Note that $e_{0|T}$ and $v_{0|T}$ are the expectation and variance of $(Y_1 x_1 + Y_2 x_2)$ evaluated under H_0, conditioned on $\boldsymbol{T} = (T_{m_1, m_2}, T_m^{(1,2)}, T_m^{(1,0)}, T_m^{(2,0)})$.

Under H_0, the statistic asymptotically follows a central chi-squared distribution with 1 degrees of freedom and may be used to detect trend effects of the exposure on the disease risk.

5 Sample Size Determination

5.1 Categorical exposure variable: detecting association

To determine the power of the test as given in (8) we need to determine the distribution of the test statistic S under the alternative hypothesis.

The mean $\boldsymbol{D}$ and the variance-covariance matrix $\boldsymbol{\mu}$ of $\boldsymbol{Z}$ are

$$
\boldsymbol{\mu} = (E(E_\psi(Y_1) - \mu_{10}),\, E(E_\psi(Y_2) - \mu_{20})) \tag{17}
$$

$$
\boldsymbol{D} = \begin{pmatrix} E(\mathrm{Var}_\psi\,(Y_1)) & E(\mathrm{Cov}_\psi\,(Y_1, Y_2)) \\ E(\mathrm{Cov}_\psi\,(Y_1, Y_2)) & E(\mathrm{Var}_\psi\,(Y_2)) \end{pmatrix}, \tag{18}
$$

where the outer expectation denotes the expectation with respect to $\boldsymbol{T}$, and $E(\boldsymbol{T}) = N(p_{m_1, m_2}, p_m^{(1,2)}, p_m^{(1,0)}, p_m^{(2,0)})$. $E_\psi(Y_1)$, $E_\psi(Y_2)$, $\mathrm{Var}_\psi\,(Y_1)$, $\mathrm{Var}_\psi\,(Y_2)$, and $\mathrm{Cov}_\psi\,(Y_1, Y_2)$ denote the conditional mean, variance, and covariance of Y_1 and Y_2 given $\boldsymbol{T}$. Let ζ_i's be the eigen vectors of D_0 with eigen values λ_i's, then we have the following result.

Theorem *Under the alternative hypothesis, S has approximate Chi square distribution with degrees freedom*

$$
\nu = \max\left\{ 1, \frac{2\sum_i \lambda_i w_i (1 + \delta_i) - 1}{\sum_i \lambda_i^2 w_i^2 (1 + 2\delta_i)} \right\}
$$

and non-centrality parameter $\delta = \max\left\{ 0, \nu\{\sum_i \lambda_i w_i (1 + \delta_i) - 1\} \right\}$, *where* $\delta_i = (\zeta_i^T \boldsymbol{\mu})^2$ *and* $w_i = \zeta_i^T \boldsymbol{D} \zeta_i$, *and the expression for* $\boldsymbol{\mu}$ *and* $\boldsymbol{D}$ *are given in (17) and (18).*

Note that all the quantities are function of N as indicated in the appendix. Proof of this result and a guideline of how to use it in practice are given in Appendix 4. For given type-II error probability, β, one can now find the required sample size by satisfying,

$$
\inf\left\{ N : \mathrm{pr}\,(\chi_\nu^2(\delta) \geq \chi_{2,\alpha}^2) \geq 1 - \beta \right\},
$$

where $\chi^2_\nu(\delta)$ denotes chi-square distribution with ν degrees of freedom and non-centrality parameter δ.

Note that both the non-centrality parameter and degrees of freedom of the chi-squared distribution as specified above are functions of the number of matched sets N, the exposure prevalences in the control population, and the odds ratios. Therefore to determine the sample size for given values of exposure prevalences in the control population, the odds ratios, type-I error probability α, and type-II error probability β, we first calculate pr $(\chi^2_\nu(\delta) \geq \chi^2_{\alpha,2})$ for smallest possible value of N, and then we keep on increasing N by 1 until the above probability exceeds $(1-\beta)$. The value of N when we stop is reported as the required sample size.

5.2 Ordinal Categorical Exposure Variable: Detecting Trend

To determine required sample size to detect linear trend in disease risk with ordinal levels of exposure variable, we need to calculate the power function of the test defined in (16). Let, $e_0 = E(e_{0\,|\,T})$, $v_0 = E(v_{0\,|\,T})$, where these expectations are evaluated at $\boldsymbol{\psi} = \mathbf{1}$. Let the expectations evaluated at any value of $\boldsymbol{\psi}$ in the alternative space be denoted by, $e_1 = E\left\{E\left(\sum_{i=1}^{N}\sum_{h=0}^{2} x_h N^i_{1h} \mid \boldsymbol{T}\right)\right\}$ and $v_1 = E\left\{\mathrm{Var}\left(\sum_{i=1}^{N}\sum_{h=0}^{2} x_h N^i_{1h} \mid \boldsymbol{T}\right)\right\}$. The exact expressions for e_1 and v_1 are collected in Appendix 5. Then using standard asymptotics we write

$$\begin{aligned}
\text{Power} &= 1 - P\left(e_0 - z_{\alpha/2} v_0^{1/2} \leq \sum_{i=1}^{N}\sum_{h=0}^{2} x_h N^i_{1h} \leq e_0 + z_{\alpha/2} v_0^{1/2}\right)\\
&= 1 - P\left(\frac{e_0 - z_{\alpha/2} v_0^{1/2} - e_1}{v_1^{1/2}} \leq \frac{\sum_{i=1}^{N}\sum_{h=0}^{2} x_h N^i_{1h} - e_1}{v_1^{1/2}} \leq \frac{e_0 + z_{\alpha/2} v_0^{1/2} + e_1}{v_1^{1/2}}\right)\\
&= 1 - \Phi\left(\frac{e_0 + z_{\alpha/2} v_0^{1/2} - e_1}{v_1^{1/2}}\right) + \Phi\left(\frac{e_0 - z_{\alpha/2} v_0^{1/2} - e_1}{v_1^{1/2}}\right),
\end{aligned} \tag{19}$$

where $\Phi(\cdot)$ stands for cumulative probability function for standard normal distribution. We use this power function to obtain the required sample size.

Remark 2 When the Type II error probability β is small, the last term on the right-hand side of (22) is negligible. In this case,

$$-z_\beta = \frac{e_0 - e_1 + z_{\alpha/2} v_0^{1/2}}{v_1^{1/2}} = \frac{N^{1/2}(e_0^* - e_1^*) + z_{\alpha/2} v_0^{*1/2}}{v_1^{*1/2}}, \tag{20}$$

where, from the expressions of e_i as collected in the appendix, we can write, $e_i = Ne_i^*$ and $v_i^{1/2} = N^{1/2} v_i^{*1/2}$, where e_i^* and v_i^* are free of N. It follows that,

$$N = \frac{(z_\beta v_1^{*1/2} + z_{\alpha/2} v_0^{*1/2})^2}{(e_0^* - e_1^*)^2}. \tag{21}$$

6 Example and Simulation

6.1 Example 1: colorectal cancer in rats

The dataset we are considering is part of a large study conducted by a team of Texas A & M University researchers (Hong et al., 2001) to investigate the effect of dietary fat on the development of colon carcinogenesis in rats. In brief, colon cells replicate and spend their entire life cycle within crypts, finger like structures that grow into the wall of colon. The function of the normal crypts is to produce cells that line the colon. The entire study consists of 6 rats, and within each rat 20 crypts were considered, identified by the numbers 1 through 20. Here we focused on particular stem cells and noted if the cell has undergone apoptosis. Apoptosis is termed as programmed cell death or cell suicide in response to a variety of stimuli. This is a normal process in multicellular organisms. In our dataset a cell is considered as a case if it has undergone apoptosis and considered as a control cell otherwise. Cell proliferation is defined as growth in the number of cells due to reproduction and division of cells in a multicellular organism. For each stem cell in our dataset one has information on proliferation status which is categorized as no proliferation, medium proliferation, and high proliferation and would be considered as our exposure variable. There were a total of 214 stem cells which underwent apoptosis and we randomly selected three controls for each case using crypt as a matching variable. So, we formed a 1 : 3 matched case-control study with 214 total matched sets. In the following, we outline the computational steps for calculating the odds ratio and the proposed test statistic for this particular dataset.

Note that here $M = 3$. From Table 1 we get all the quantities needed to compute the estimates in (14) and (15), and we obtain $\hat{\psi}_{1MH} = 2.0975$ and $\hat{\psi}_{1MH} = 11.20$. In order to test the null hypothesis of no association between proliferation status and apoptosis, we calculate our test statistic in (8) using the following two steps.

Step 1. Using Table 1 we calculate Y_1 and Y_2 as described in (10) and (11) and they are obtained as $Y_1 = 30$ and $Y_2 = 20$. Similarly following the similar type of calculations we obtain $\mu_{20} = 6.75$.

Using the formulas given in the Appendix, we obtain $\mu_{10} = 20.75$ and $\mu_{20} = 6.75$, whereas the elements of D_0 are obtained as, $\sigma_{10}^2 = 13.4375$, $\sigma_{20}^2 = 5.625$, and $\sigma_{120} = -0.6875$.

Step 2. Using formula (8), we obtain $S = 31.7608$. So according to (12) we reject the null hypothesis of no association between proliferation and apoptosis at 5% level of significance.

Since there is a natural ordering among the proliferation status, we also perform a test of trend on the risk of apoptosis with increasing levels of proliferation as given in (16). For this dataset, we take $x_1 = 1$ and $x_2 = 2$, and $e_{0|T} = 34.25, v_{0|T} = 30.9375$. Hence by (16), the test statistics is $\chi^2 = 41.3111$. Comparing with the chi-squared distribution (with d$f = 1$) cut-off value, we reject the null hypothesis of no trend effect at 5% level of significance. This finding, in fact conforms with the biological association expected between apoptosis and cell proliferation.

To do a realistic simulation we mimic the colon carcinogenesis dataset in hand. We note that among the controls, the prevalence of the different categories of the exposure variable are 0.9101,

Table 1 Calculation of the relevant quantities as defined on page 6 for the discordant matched sets for example 1, where $\sum_{i=0}^{2} n^{i}_{m_1,m_2} = T_{m_1,m_2}$, $\sum_{i=0,1} n_m^{i(1,0)} = T_m^{(1,0)}$, and $\sum_{i=0,2} n_m^{i(2,0)} = T_m^{(2,0)}$.

(m_1, m_2)	$n^0_{m_1,m_2}$	$n^1_{m_1,m_2}$	$n^2_{m_1,m_2}$	T_{m_1,m_2}
(1, 1)	1	0	4	5
(1, 2)	0	0	0	0
(2, 1)	1	2	0	3

m	$n_m^{0(1,0)}$	$n_m^{1(1,0)}$	$T_m^{(1,0)}$	$n_m^{0(2,0)}$	$n_m^{2(2,0)}$	$T_m^{(2,0)}$	$n_m^{1(1,2)}$	$n_m^{2(1,2)}$	$T_m^{(1,2)}$
1	29	24	53	3	16	19	0	0	0
2	3	2	5	0	0	0	0	0	0
3	1	2	3	0	0	0	0	0	0

Table 2 Required sample size for detecting association in example 1, 1:3 matched study, where ψ_1 and ψ_2 are the two odds ratios. The type I error is set at 5% while the power is set at 80%.

ψ_2	ψ_1				
	1.5	2.0	2.5	3.0	5.0
1.5	1048	798	638	535	355
2.0	550	496	445	400	300
2.5	356	342	322	304	253
3.0	258	254	246	238	216
5.0	116	118	119	120	124

0.0799, and 0.01. With the above specifications of the exposure prevalence p_{00}, p_{01}, and p_{02} respectively, we calculate the required sample sizes for different values of ψ_1 and ψ_2. All the sample sizes are calculated after setting the Type I error probability $\alpha = 0.05$ and Type II error probability $\beta = 0.20$. As expected one needs more sample for smaller values ψ_1 and ψ_2 compared to the higher values of the odds ratios as in the latter case, departures from the null hypothesis is more pronounced. The results are presented in Table 2.

In addition, we compute the required sample sizes for detecting trend with 3 ordered levels of the exposure variable using the power function (19) for different values of γ. As expected, as γ increases, the required number of matched sets decreases.

6.2 Example 2: association between disease and *A* genetic factor

In this example we first generate a prototype case-control dataset for exploring disease-gene association by mimicking a real data on allele frequencies of 12 marker loci in the Buenos Aires metropolitan population as described in Sala et al. (1999), Table 3. We chose the most frequent allele at marker locus F13A as the disease causing allele. This selection is arbitrary. Sala et al. (1999) reports the frequency of this allele in the population to be $p = 0.312$. Assuming the population is in Hardy–Weinberg (HW) equilibrium we calculate the prevalence of three possible genotypes. If A denotes the disease causing allele, then $P(AA) = p^2 = 0.0974$, $P(Aa) = 2p(1-p) = 0.4293$, and $P(aa) = (1-p)^2 = 0.473$. Let g_0, g_1, and g_2 denote the genotypes *aa*, *Aa*, and *AA*. Note that though the bi-allelic gene is assumed to be in HW equilibrium in the general population, among the case and control population the genotype frequencies may not be in HW equilibrium. We assume that the log-odds ratios of the disease among persons with no copy (*aa*), one copy (*Aa*), and two copies (*AA*) of the disease causing allele to be 0, 1, and 1, i.e., $\beta_0 = 0$, $\beta_1 = 1$ and $\beta_2 = 1$. In other words, the disease risk is elevated if you have at least one copy of the disease causing allele. For generating the disease status data, we assume the prospective stratified logistic regression model for a matched case

Table 3 Required sample sizes for trend test in example 1, 1:3 matched study, where $\psi_1 = e^{\gamma}$ and $\psi_2 = e^{2\gamma}$ are the two odds ratios, and N denotes the required number of matched sets. The type I error is set at 5% while the power is set at 80%.

γ	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.2
ψ_1	1.35	1.49	1.65	1.82	2.01	2.22	2.46	3.32
ψ_2	1.82	2.22	2.72	3.32	4.05	4.95	6.05	11.02
N	1140	845	406	219	127	78	49	15

control study (Breslow et al. (1978)),

$$P(D = 1 \mid S_i, G = g_m) = \frac{\exp(\alpha_i + \beta_m g_m)}{1 + \exp(\alpha_i + \beta_m g_m)}\,, \quad m = 0, 1, 2\,,$$

where S_i denotes the unmeasured or measured stratum specific covariates and α_i is effect of the i-th matched set on the disease probability, $i = 1, \ldots, N$ (N is the number of matched sets).

To simulate a 1 : 2 matched case-control data with 100 matched sets, we first generate the genetic exposure variable G for each of the three subjects in every matched set according to the prevalence values mentioned before. Let G_{i1}, G_{i2} and G_{i3} denote the genetic exposure corresponding to the three subjects in the i-th matched set, $i = 1, \ldots, 100$. Given these exposure values we determine the disease status D_{ij} of the j-th subject in the i-th stratum following a conditional logistic regression model (Breslow et al. (1978)), conditioning on the event that the number of cases in each matched set is pre-fixed by the study design to be 1. First, We generate a Bernoulli random variable D_{i1} with success probability

$$p = P\left(D_{i1} = 1 \mid G_{i1}, G_{i2}, G_{i3}, \sum_{j=1}^{3} D_{ij} = 1\right) = \frac{\exp(\beta_{G_{i1}})}{\exp(\beta_{G_{i1}}) + \exp(\beta_{G_{i2}}) + \exp(\beta_{G_{i3}})}\,,$$

where

$$\beta_{G_{ij}} = \begin{cases} \beta_0 & \text{if} \quad G_{ij} = g_0 \\ \beta_1 & \text{if} \quad G_{ij} = g_1 \\ \beta_2 & \text{if} \quad G_{ij} = g_2\,. \end{cases}$$

If D_{i1} equals 1, set D_{i2} and D_{i3} equal to zero. Otherwise, we generate D_{i2} from a Bernoulli distribution with success probability

$$p = P\left(D_{i2} = 1 \mid G_{i1}, G_{i2}, G_{i3}, \sum_{j=1}^{3} D_{ij} = 1, D_{i1} = 0\right) = \frac{\exp(\beta_{G_{i2}})}{\exp(\beta_{G_{i2}}) + \exp(\beta_{G_{i3}})}\,.$$

Now, if D_{i2} equals 1, set $D_{i3} = 0$, otherwise set $D_{i3} = 1$. Following the above scheme we generate $N = 100$ matched sets with one case and three controls. Table 4 gives the summary of discordant matched sets in the simulated dataset.

Note that here $M = 2$. As in example 1, we first calculate the two odds ratio estimates given in (14) and (15), and they are, $\hat{\psi}_{1MH} = 3.0$ and $\hat{\psi}_{2MH} = 4.5$. Next we calculate the test statistic to test the null hypothesis of no disease-exposure association by using the following two steps.

Step 1. Using Table 4, Eqs. (10) and (11), and the formulas given in the Appendix we calculate $Y_1 = 53$, $Y_2 = 10$, $\mu_{10} = 37.3333$, and $\mu_{20} = 9.3333$. The elements of D_0 are obtained as: $\sigma_{10}^2 = 16.8889$, $\sigma_{20}^2 = 5.8889$, and $\sigma_{120} = -3.8889$.

Table 4 Calculation of the relevant quantities as defined on page 6 for the discordant matched sets for example 2, where $\sum_{i=0}^{2} n^i_{m_1,m_2} = T_{m_1,m_2}$, $\sum_{i=0,1} n_m^{i(1,0)} = T_m^{(1,0)}$, $\sum_{i=0,2} n_m^{i(2,0)} = T_m^{(2,0)}$, and $\sum_{i=1,2} n_m^{i(1,2)} = T_m^{(1,2)}$.

(m_1, m_2)	$n^0_{m_1,m_2}$	$n^1_{m_1,m_2}$	$n^2_{m_1,m_2}$	T_{m_1,m_2}
(1,1)	1	6	2	9

m	$n_m^{0(1,0)}$	$n_m^{1(1,0)}$	$T_m^{(1,0)}$	$n_m^{0(2,0)}$	$n_m^{2(2,0)}$	$T_m^{(2,0)}$	$n_m^{1(1,2)}$	$n_m^{2(1,2)}$	$T_m^{(1,2)}$
1	12	19	31	1	3	4	0	0	0
2	4	19	23	0	1	1	9	4	13

Table 5 Required sample size for detecting association in example 2, 1:2 matched study, where ψ_1 and ψ_2 are the two odds ratios. The type I error is set at 5% while the power is set at 80%.

ψ_1	ψ_2				
	1.5	2.0	2.5	3.0	5.0
1.5	158	115	86	72	51
2.0	84	98	91	79	54
2.5	55	70	78	79	58
3.0	42	53	63	69	63
5.0	23	28	33	39	59

Step 2. Plugging all the values in (8) we obtain the value of the test statistic $S = 18.19$ with a P-value of 0.00014. So there is significant association between the disease and the candidate gene.

We also calculate the test statistic in (16) to see if there is any trend in the disease risk with increasing levels of the exposure variable (the three possible genotypes in this case with zero, one or two copies of the allele A respectively). The value of this test statistic turns out to be 15.434 which is statistically significant when compared with the chi-squared cut off value (with d$f = 1$) at 5% level of significance.

For the above scenario we calculate the required sample sizes to detect the departures from the null hypothesis of no association for given prevalence values and various choices of ψ_1 and ψ_2. For the simulation, we assume that the exposure prevalence values in the control population are the same as in the overall population. This is equivalent to assuming the disease to be rare. Therefore we take $p_{01} = 0.4293$ and $p_{02} = 0.0974$. The results are presented in Table 5.

We also calculate sample sizes for detecting trend with the ordered categorical exposure variable. The results are presented in Table 6. All the sample sizes are calculated after setting $\alpha = 0.05$ and $\beta = 0.20$.

Figure 1 presents a plot of the power function against the required sample size when one wants to detect a given difference in ψ_1 and ψ_2 with a pre-specified probability (power) for a 1:2 matched study. The exposure prevalences are chosen as in the simulated dataset. With the same prevalence values of the exposure variable, we also calculate the power of the ordinal trend test for different sample sizes. Figure 2 presents two such power curves with $\gamma = 0$ and $\gamma = 1$ for a 1:2 matched case-control design.

Table 6 Required sample size for detecting trend in example 2, 1:2 matched study, where $\psi_1 = e^{\gamma}$ and $\psi_2 = e^{2\gamma}$ are the two odds ratios, and N denotes the required number of matched sets. The type I error is set at 5% while the power is set at 80%.

γ	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.2
ψ_1	1.35	1.49	1.65	1.82	2.01	2.22	2.46	3.32
ψ_2	1.82	2.22	2.72	3.32	4.05	4.95	6.05	11.02
N	292	137	76	47	32	22	16	8

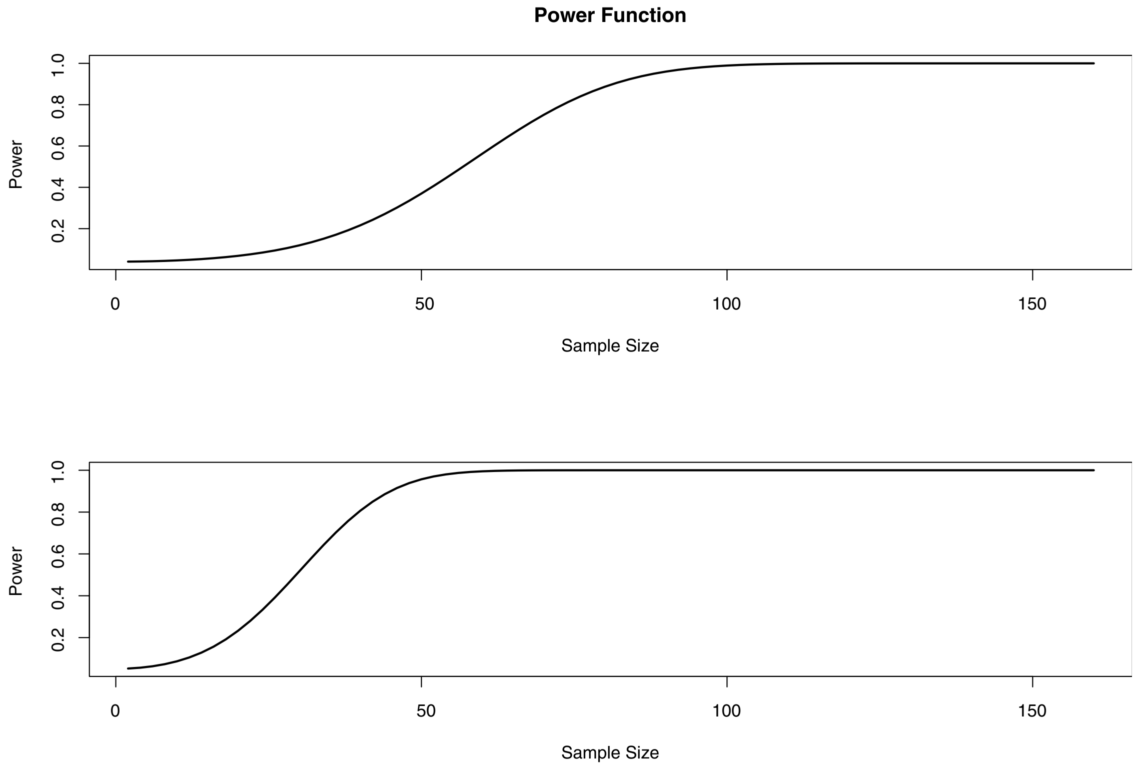

Figure 1 Power function for testing association with polytomous exposure variable is plotted against sample size for $p_{01} = 0.4293$ and $p_{02} = 0.0974$ for 1 : 2 matched case-control studies. The top figure is for $\psi_1 = \psi_2 = \exp(1) = 2.72$, and the bottom figure is for $\psi_1 = \exp(1) = 2.72$ and $\psi_2 = \exp(1.5) = 4.48$.

6.3 Binary exposure variable

In conjunction with our two examples, in this sub-section, we explore the performance of our method if the exposure levels are dichotomized. We compare our sample size recipe with the standard formula provided by Schlesselman (1982). For 1 : 1 matched pair data with binary exposure, the required sample size (Schlesselman, 1982, Eqs. 6–20 and 6–23, Parker and Bregman, 1986, Eq. (3.1), modified for a two-tailed test) is given as,

$$N_S^* = \frac{\left\{\dfrac{Z_{\alpha/2}(1+\psi) + 2Z_\beta \psi^{1/2}}{\psi - 1}\right\}^2}{\dfrac{(\psi+1)\,p_{01}(1-p_{01})}{1+(\psi-1)\,p_{01}}}\,. \tag{22}$$

Where ψ is the odds ratio and p_{01} is the exposure prevalence in the control population. This formula for sample size for pair-matched studies is based on the normal approximation to testing a single binomial proportion and a crude estimate of the probability of a discordant exposure pair.

For $1:M$ matched case-control study with binary exposure variable the number of matched sets required to attain a certain power of the test is the fraction $(M+1)/2M$ of the sample size required for 1:1 matched case-control design. According to Ury (1975) this fraction is the reciprocal of the asymptotic relative efficiency of a $1:M$ to 1 : 1 matched study design. Therefore for

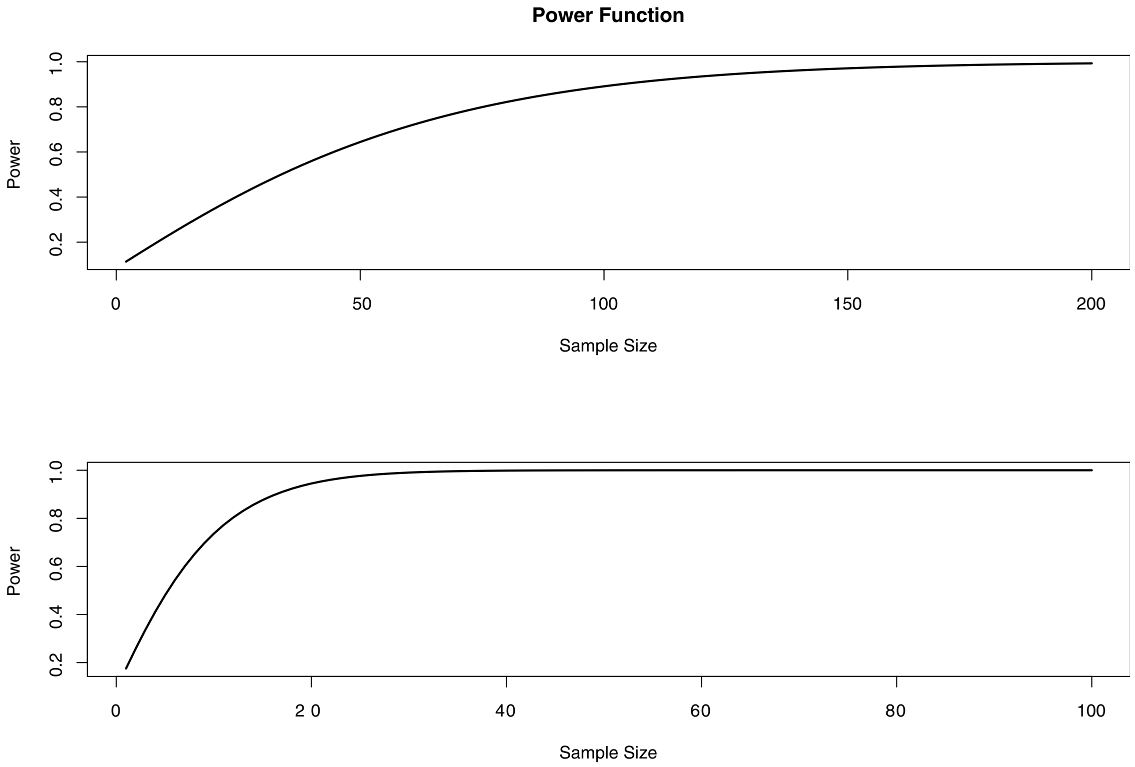

Figure 2 Power function for detecting trend with ordinal categorical exposure variable is plotted against sample size for $p_{01} = 0.4293$ and $p_{02} = 0.0974$ for 1:2 matched case-control studies. The top figure is for $\gamma = 0.50$, and the bottom figure is for $\gamma = 1$.

$1\!:\!M$ matched design the needed sample size is $N_S = (M+1)\,N_S^*/2M$. In contrast, according to our method the required sample size for binary exposure variable is derived based on the power function of the test statistic in (13). When the Type II error probability is small (set at 20% for our simulations), for binary exposure one could alternatively use (21) with $e_0^* = \sum_m m p_m^{(1,0)}$, $v_0^* = \sum_m m(M-m+1)\,p_m^{(1,0)}/(M+1)^2$, $e_1^* = \sum_m p_m^{(1,0)} m\psi/(m\psi + M - m + 1)$, and $v_1^* = \sum_m p_m^{(1,0)} m\psi(M-m+1)/(m\psi+M-m+1)^2$, and $p_m^{(1,0)} = p_{11}\binom{M}{(m-1)} p_{01}^{m-1} p_{00}^{M-m+1} + p_{10}\binom{M}{m} p_{01}^m p_{00}^{M-m}$. Here $\psi = p_{11}p_{00}/p_{10}p_{01}$, where $p_{11} + p_{10} = 1$ and $p_{01} + p_{00} = 1$. Recall that for the binary exposure case our methods are essentially identical to that of Breslow and Day (1980, Chapter 5).

For the setting of the first example with 1:3 matching, in order to dichotomize the exposure, we collapse categories 1 and 2 of the exposure variable and call it category 1. Therefore the exposure prevalence among the control becomes $p_{01} = 0.0799 + 0.01 = 0.0899$. The sample size results are presented in Table 7. Note that, the number of matched sets required by our method, is typically larger than those obtained by Schlesselman's formula. One of the reasons for Schlesselman's method to underestimate the sample size could be that obtaining the sample size formula for $1\!:\!M$ studies by multiplying the sample size for 1:1 matched design with the factor of $(M+1)/2M$ may not always be optimal. The ad hoc estimate of probability of discordant exposure pairs used in Schlesselman's formula is often quite crude. Also note that the exposure prevalence is low in this setting as a result the required number of matched sets in both methods is relatively large.

Table 7 Required sample sizes for 1:3 matched case-control study with binary exposure by Schlesselman's method and the method proposed in the current paper. Here ψ denotes the odds ratio. Exposure prevalence p_{01} is set at 0.0899. The type I error is set at 5% while the power is set at 80%.

ψ	1.5	2	2.5	3	3.5	4
N_S	660	202	107	70	51	40
N	751	251	142	98	75	61

Table 8 Required sample sizes for 1 : 2 matched case-control study with binary exposure by Schlesselman's method and the method proposed in the current paper. Here ψ denotes the odds ratio. Exposure prevalence p_{01} is set at 0.5267. The type I error is set at 5% while the power is set at 80%.

ψ	1.5	2	2.5	3	3.5	4
N_S	295	105	62	45	36	30
N	288	99	58	41	32	26

In the setting of the second example with 1 : 2 matched design for detecting disease-gene association, the prevalence of the genetic exposure after collapsing the categories $G = g_1$ and $G = g_2$ is given by $p_{01} = 0.4293 + 0.0974 = 0.5267$. The results comparing our method with those of Schlesselman are presented in Table 8. In this case, our sample sizes are marginally larger than those of Schlesselman but both the methods furnish very similar sample sizes. Note that with larger prevalence of the genetic factor, the sample sizes for both the methods are smaller when compared with those obtained in Table 7. Based on these simulation, the matching ratio, as well as the the exposure prevalence seems to play a role in determining the relative performances of these methods.

Parker and Bregman (1986) take Schlesselman's standard formula in (22) as a starting basis and incorporate the new dimension that the exposure prevalence p_{01} could vary in each matched set. In presence of such heterogeneity, (22) normally underestimates the required sample size (Parker and Bregman, 1986), thus the sample sizes obtained by Parker and Bregman will be typically larger than N_S. Since the principal goal of this paper is to address exposure variables with multiple categories in a 1 : M matched study, for the binary exposure, we take Schlesselman's formula as a precursor for comparison purposes and refrain from specifically exploring the effect of stratum heterogeneity on sample size determination along the lines of Parker and Bregman (1986).

7 Discussion and Final Comments

In this paper we furnish an easy to use recipe for determining the number of matched sets in a 1 : M matched case-control study with multiple levels of the exposure variable. In the process we derive a test statistic for the independence of the exposure and the disease variable when the categories are nominal and when they are ordinal. The advantage of this method is that it does not require computation of the MLE's of the parameters to perform the test, rather we can do the test by simply constructing tables of discordant matched sets. One may also derive a score test for testing homogeneity of the k odds ratios, i.e. $H_0 : \psi_1 = \psi_2 = \ldots = \psi_k$, using the likelihood function we have used in this paper.

We briefly indicate that estimation of the common odds ratio following the classical ideas of Mantel and Haenszel (1959) but do not furnish standard error formulae. For interval estimation of the odds ratio and derivation of the standard errors, one may resort to modern resampling techniques like the bootstrap as an alternative to establishing closed-form approximation formulae. However the goal of this paper is more modest. We focus on the designing issue when one wants to detect exposure-disease association in a matched case-control study with categorical exposure. Extension of the results to a set of multivariate categorical exposures which may be associated among themselves poses interesting technical challenges one might undertake. Admittedly, the approach presented in this paper is not the natural framework to deal with continuous exposure variables. In the most general setting, when one has a mixed set of continuous and discrete exposure variables, the problem needs to be recast in very different ways than proposed in this paper. A more general strategy to attack the design problem could be adopted by following simulation based design ideas as described in Müller (1999) under a Bayesian paradigm. The *R* code for calculating the sample size and the datasets used in this paper are available on http://stat.tamu.edu/~sinha.

Appendix

1 Formulation of the test statistic (8) for general *k*

Let $n^i_{m_{r_1},m_{r_2},\dots,m_{r_j}}$ be the number of matched sets where case is exposed at the level i and m_{r_s} subjects are exposed at the level r_s, $r_s = r_1, \dots, r_j$ and $i \in (r_1, r_2, \dots, r_j)$. Here $m_{r_1} + m_{r_2} + \dots + m_{r_j} = M + 1$. Let S_j be the set of all possible combinations of any j levels out of $k+1$ levels, $j = 2, \dots, k+1$. Then $S = \bigcup_{j=2}^{k+1} S_j$ is the collection of all possible combinations of at least two levels taken at a time from $k+1$ levels. Define

$$Y_i = \sum_{j=2}^{k+1} \sum_{(r_1,\dots,r_j)\in S_j} \sum_{m_{r_1},\dots,m_{r_j}} n^i_{m_{r_1},\dots,m_{r_j}} I(i \in (r_1,\dots,r_j)) \quad \text{for} \quad i = 1, 2, \dots, k\,.$$

Let $= (\boldsymbol{\mu}_1, \dots, \mu_k)$, where

$$\mu_i = E(Y_i) = \sum_{j=2}^{k+1} \sum_{(r_1,\dots,r_j)\in S_j} \sum_{m_{r_1},\dots,m_{r_j}} T_{m_{r_1},\dots,m_{r_j}} \frac{m_i \psi_i I(i \in (r_1,\dots,r_j))}{m_{r_1}\psi_{r_1} + \dots + m_{r_j}\psi_{r_j}}$$

and $T_{m_{r_1},\dots,m_{r_j}} = \sum_{i=0}^{k} n^i_{m_{r_1},\dots,m_{r_j}} I(i \in (r_1,\dots,r_j)$. Under H_0, the mean μ_i's will be

$$\sum_{j=2}^{k+1} \sum_{(r_1,\dots,r_j)\in S_j} \sum_{m_{r_1},\dots,m_{r_j}} T_{m_{r_1},\dots,m_{r_j}} \frac{m_i}{\sum_{s=1}^{j} m_{r_s}} I(i \in (r_1,\dots,r_j))\,.$$

Further assume that $\boldsymbol{D}_{k\times k} = (\sigma_{pq})$ be the variance-covariance matrix of $\boldsymbol{Y} = (Y_1, \dots, Y_k)$, where

$$\sigma_{pq} = -\sum_{j=2}^{k+1} \sum_{(r_1,\dots,r_j)\in S_j} \sum_{m_{r_1},\dots,m_{r_j}} T_{m_{r_1},\dots,m_{r_j}} \frac{m_p m_q \psi_p \psi_q I((p,q) \in (r_1,\dots,r_j))}{(m_{r_1}\psi_{r_1} + \dots + m_{r_j}\psi_{r_j})^2} \quad \text{if} \quad p \neq q\,.$$

The test statistics to test H_0 is

$$\boldsymbol{Y}(-\boldsymbol{\mu}_0)^T \boldsymbol{D}_0^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}_0)\,,$$

where $\boldsymbol{\mu}_0$ and $\boldsymbol{D}_0$ are the mean and variance-covariance matrix of $\boldsymbol{Y} = (Y_1, \dots, Y_k)$ evaluated under H_0. Using the analogous argument that we used in the main text, one can derive the asymptotic distribution of the above statistics, and under H_0 this statistic approximately follows central chi-square distribution with k degrees of freedom.

2 Expressions for the moments of Y_1 and Y_2

$$E_\psi(Y_1) = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} I(m_1 + m_2 \le M)\, T_{m_1,m_2} \frac{m_1\psi_1}{\{m_1\psi_1 + m_2\psi_2 + (M - m_1 - m_2 + 1)\}}$$
$$+ \sum_{m=1}^{M} T_m^{(1,2)} \frac{m\psi_1}{\{m\psi_1 + (M - m + 1)\,\psi_2\}} + \sum_{m=1}^{M} T_m^{(1,0)} \frac{m\psi_1}{\{m\psi_1 + (M - m + 1)\}}$$

$$E_\psi(Y_2) = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} I(m_1 + m_2 \le M)\, T_{m_1,m_2} \frac{m_2\psi_2}{\{m_1\psi_1 + m_2\psi_2 + (M - m_1 - m_2 + 1)\}}$$
$$+ \sum_{m=1}^{M} T_m^{(1,2)} \frac{(M - m + 1)\,\psi_2}{\{m\psi_1 + (M - m + 1)\,\psi_2\}} + \sum_{m=1}^{M} T_m^{(2,0)} \frac{m\psi_2}{\{m\psi_2 + (M - m + 1)\}}$$

$$\mathrm{Var}_\psi\,(Y_1) = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} I(m_1 + m_2 \le M)\, T_{m_1,m_2} \frac{(m_1\psi_1)\,\{m_2\psi_2 + (M - m_1 - m_2 + 1)\}}{\{m_1\psi_1 + m_2\psi_2 + (M - m_1 - m_2 + 1)\}^2}$$
$$+ \sum_{m=1}^{M} T_m^{(1,2)} \frac{(M - m + 1)\,\psi_2 m\psi_1}{\{m\psi_1 + (M - m + 1)\,\psi_2\}^2} + \sum_{m=1}^{M} T_m^{(1,0)} \frac{m\psi_1(M - m + 1)}{\{m\psi_1 + (M - m + 1)\}^2}$$

$$\mathrm{Var}_\psi\,(Y_2) = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} I(m_1 + m_2 \le M)\, T_{m_1,m_2} \frac{(m_2\psi_2)\{m_1\psi_1 + (M - m_1 - m_2 + 1)\}}{\{m_1\psi_1 + m_2\psi_2 + (M - m_1 - m_2 + 1)\}^2}$$
$$+ \sum_{m=1}^{M} T_m^{(1,2)} \frac{(M - m + 1)\,\psi_2 m\psi_1}{\{m\psi_1 + (M - m + 1)\,\psi_2\}^2} + \sum_{m=1}^{M} T_m^{(2,0)} \frac{m\psi_2(M - m + 1)}{\{m\psi_2 + (M - m + 1)\}^2}$$

$$\mathrm{Cov}_\psi\,(Y_1, Y_2) = -\sum_{m_1=1}^{M} \sum_{m_2=1}^{M} I(m_1 + m_2 \le M)\, T_{m_1,m_2} \frac{(m_2\psi_2)\, m_1\psi_1}{\{m_1\psi_1 + m_2\psi_2 + (M - m_1 - m_2 + 1)\}^2}$$
$$- \sum_{m=1}^{M} T_m^{(1,2)} \frac{(M - m + 1)\,\psi_2 m\psi_1}{\{m\psi_1 + (M - m + 1)\,\psi_2\}^2}$$

$$\mu_{10} = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} I(m_1 + m_2 \le M)\, T_{m_1,m_2} \frac{m_1}{M + 1} + \sum_{m=1}^{M} T_m^{(1,2)} \frac{m}{M + 1} + \sum_{m=1}^{M} T_m^{(1,0)} \frac{m}{M + 1}$$

$$\mu_{20} = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} I(m_1 + m_2 \le M)\, T_{m_1,m_2} \frac{m_2}{M + 1} + \sum_{m=1}^{M} T_m^{(1,2)} \frac{(M - m + 1)}{M + 1} + \sum_{m=1}^{M} T_m^{(2,0)} \frac{m}{M + 1}$$

$$\sigma_{10}^2 = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} I(m_1 + m_2 \le M)\, T_{m_1,m_2} \frac{m_1(M - m_1 + 1)}{(M + 1)^2} + \sum_{m=1}^{M} T_m^{(1,2)} \frac{(M - m + 1)\, m}{(M + 1)^2}$$
$$+ \sum_{m=1}^{M} T_m^{(1,0)} \frac{m(M - m + 1)}{(M + 1)^2}$$

$$\sigma_{20}^2 = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} I(m_1 + m_2 \le M)\, T_{m_1,m_2} \frac{m_2(M - m_2 + 1)}{(M + 1)^2} + \sum_{m=1}^{M} T_m^{(1,2)} \frac{(M - m + 1)\, m}{(M + 1)^2}$$
$$+ \sum_{m=1}^{M} T_m^{(2,0)} \frac{m(M - m + 1)}{(M + 1)^2}$$

$$\sigma_{120} = -\sum_{m_1=1}^{M} \sum_{m_2=1}^{M} I(m_1 + m_2 \le M)\, T_{m_1,m_2} \frac{m_2 m_1}{(M + 1)^2} - \sum_{m=1}^{M} T_m^{(1,2)} \frac{(M - m + 1)\, m}{(M + 1)^2}$$

3 Proof of the theorem

By spectral decomposition one can write

$$S = \boldsymbol{Z}^T D_0^{-1} \boldsymbol{Z} = \boldsymbol{Z}^T \sum_i \lambda_i \zeta_i \zeta_i^T \boldsymbol{Z} \tag{A.1}$$

where $0 < \lambda_1 \leq \lambda_2$ are the eigen values of D_0^{-1} with corresponding orthogonal eigenvectors ζ_1 and ζ_2. Rewrite (A.1) as

$$S = \boldsymbol{Z}^T \sum_i \lambda_i \zeta_i \zeta_i^T \boldsymbol{Z} = \sum_i \lambda_i \boldsymbol{Z}^T \zeta_i \zeta_i^T \boldsymbol{Z} = \sum_i \lambda_i U_i^2 \,, \tag{A.2}$$

where $U_i = \zeta_i^T \boldsymbol{Z}$. Therefore U_i has expectation $\zeta_i^T \boldsymbol{\mu}$ and variance $\zeta_i^T \boldsymbol{D} \zeta_i$. Under the usual regularity condition U_i^2s are independent and approximately

$$U_i^2 \sim w_i \chi_1^2(\delta_i) \,, \tag{A.3}$$

where $\delta_i = (\zeta_i^T \boldsymbol{\mu})^2$ and $w_i = \zeta_i^T \boldsymbol{D} \zeta_i$. Hence S is a linear combination of non-central chi-square distributions. Although originally S has a very complicated distribution, the Satterthwaite (1959) approximation works well in this situation.

Assume that

$$S \overset{\text{approx}}{\sim} \frac{1}{\nu} \chi_\nu^2(\delta) \,, \tag{A.4}$$

where $\chi_\nu^2(\delta)$ is a non-central chi-square distribution with ν degrees of freedom and non-centrality parameter δ. Equating first two moments of both sides of (A.4) one gets

$$E(S) = \sum_i \lambda_i w_i (1 + \delta_i) = 1 + \frac{\delta}{\nu} \,, \tag{A.5}$$

$$\mathrm{Var}\,(S) = \sum_i \lambda_i^2 w_i^2 (2 + 4\delta_i) = \frac{2}{\nu} + 4\,\frac{\delta}{\nu^2} \,. \tag{A.6}$$

Using Eqs. (A.5), (A.6) and applying the restriction that $\nu \geq 1$ and $\delta \geq 0$, one obtains

$$\nu = \max \left\{ 1, \frac{2 \sum_i \lambda_i w_i (1 + \delta_i) - 1}{\sum_i \lambda_i^2 w_i^2 (1 + 2\delta_i)} \right\} \quad \text{and} \quad \delta = \max \left\{ 0, \nu \left\{ \sum_i \lambda_i w_i (1 + \delta_i) - 1 \right\} \right\} . \tag{A.7}$$

This completes the proof of the theorem.

4 Expressions for e_1 and v_1 in (19)

$$\begin{aligned} e_1 = \sum_{m_1, m_2} E(T_{m_1, m_2}) \frac{m_1 \psi_1 x_1 + m_2 \psi_2 x_2}{m_1 \psi_1 + m_2 \psi_2 + (M - m_1 - m_2 + 1)} + x_1 \sum_m \frac{E(T_m^{(1,0)})\, m\psi_1}{m\psi_1 + (M - m + 1)} \\ + x_2 \sum_m \frac{E(T_m^{(2,0)})\, m\psi_2}{m\psi_2 + (M - m + 1)} + \sum_m E(T_m^{(1,2)}) \frac{m\psi_1 x_1 + (M - m + 1)\, \psi_2 x_2}{m\psi_1 + (M - m + 1)\, \psi_2} \end{aligned} \tag{A.8}$$

$$\begin{aligned} v_1 = \sum_{m_1, m_2} E(T_{m_1, m_2}) \frac{(M - m_1 - m_2 + 1)\, (m_1 \psi_1 x_1^2 + m_2 \psi_2 x_2^2) + m_1 m_2 \psi_1 \psi_2 (x_1 - x_2)^2}{\{m_1 \psi_1 + m_2 \psi_2 + (M - m_1 - m_2 + 1)\}^2} \\ + x_1^2 \sum_m E(T_m^{(1,0)}) \frac{m\psi_1 (M - m + 1)}{(m\psi_1 + M - m + 1)^2} + x_2^2 \sum_m E(T_m^{(2,0)}) \frac{m\psi_2 (M - m + 1)}{(m\psi_2 + M - m + 1)^2} \\ + (x_1 - x_2)^2 \sum_m E(T_m^{(1,2)}) \frac{m\psi_1 (M - m + 1)\, \psi_2}{\{m\psi_1 + (M - m + 1)\, \psi_2\}^2} \,. \end{aligned} \tag{A.9}$$

Acknowledgements We like to thank associate editor and the three referees for constructive remarks which improved the presentation of this article.

References

Armitage, P. (1955). Test for linear trend in proportions and frequencies. *Biometrics* **11**, 375–386.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research: Volume I. The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L., and Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology* **108**, 299–307.

Cox, D. R. and Hinkley, D. V. (1974). *Theoritical Statistics*. London: Chapman and Hall.

Dupont, W. D. (1988). Power Calculations for Matched Case-Control Studies. *Biometrics* **44**, 1157–1168.

Hong, M. Y., Chapkin, R. S., Morris, J. S., Wang, N., Carroll, R. J., Turner, N. D., Chang, W. C. L., Davidson, F. A., and Lupton, J. R. (2001). Anatomical site-specific response to DNA damage is related to later tumor development in the rat AOM colon carcinogenesis model. *Carcinogenesis* **22**, 1831–1835.

Jarvik, G. P. (1998). Complex segregation analyses: Uses and limitations. *American Journal of Human Genetics* **63**, 942–946.

Lusbader, E. D., Moolgavkar, S., and Venzon, D. J. (1984). Test of the null hypothesis in case-control studies. *Biometrics* **40**, 1017–1024.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute* **22**, 719–748.

Müller, P. (1999). Simulation-based optimal design. *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting* 459–474.

Nam, J.-M. (1992). Sample size determination for case-control studies and the comparison of stratified and unstratified analyses. *Biometrics* **48**, 389–395.

Nam, J.-M. (1997). Sample size determination for designing a strata-matched case-control study to detect multiple risk factors. *Biometrical Journal* **39**, 441–454.

Nam, J.-M. and Fears, T. R. (1992a). Optimum sample size determination in stratified case-control studies with cost considerations. *Statistics in Medicine* **11**, 547–556.

Nam, J.-M. and Fears, T. R. (1992b). Control sample size when cases are given in constant ratio stratummatched case-control studies. *Statistics in Medicine* **11**, 1759–1766.

Parker, R. A. and Bregman, D. J. (1986). Sample size for individually matched case-control studies. *Biometrics* **42**, 919–926.

Rao, C. R. (1947). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society* **44**, 50–57.

Sala, A., Penacino, G., Carnese, R., and Corach, D. (1999). Reference database of hypervariable genetic markers of Argentina: Application for molecular anthropology and forensic casework. *Electrophoresis* **20**, 1733–1739.

Satterthwaite, F. E. (1959). Random balance experimentation. *Technometrics* **1**, 111–137.

Schlesselman, J. J. (1974). Sample size requirements in cohort and case-control studies of disease. *American Journal of Epidemiology* **99**, 381–384.

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

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley.

Ury, H. K. (1975). Efficiency of case-control studies with multiple controls per case: Continuous or dichotomous data. *Biometrics* **31**, 643–649.