Published in final edited form as:

Stat Med. 2012 January 13; 31(1): 29–44. doi:10.1002/sim.4414.

Estimation and testing of the relative risk of disease in case-control studies with a set of k matched controls per case with known prevalence of disease^{x†}

Barry Kurt Moser^{*,†} and Susan Halabi

Department of Biostatistics and Bioinformatics and the CALGB Statistical Center, Duke University Medical Center, Box 2717, Durham, NC 27705, USA

Abstract

The analysis of case-control studies with matched controls per case is well documented in the medical literature. Of primary interest is the estimation of the relative risk of disease. Matched case-control studies fall into two scenarios: the probability of exposure is constant within each of the case and control groups, or the probability of exposure varies within each group. Numerous estimation procedures have been developed for both scenarios. Often these procedures are developed under the rare disease assumption, where the relative risk of disease is approximated by the odds ratio. In this paper, without making the rare disease assumption, we develop consistent estimators of the relative risk of disease for both scenarios. Exact derivations of the relative risk of disease are provided. Estimators, confidence intervals, and test statistics for the relative risk of disease are developed. We then make the following observations based on extensive simulations. First, our estimators are as close or closer to the relative risk of disease than other estimators. Second, our estimators produce mean square errors for the relative risk of disease that are as good as or better than these other estimators. Third, our confidence intervals provide accurate coverage probabilities. Therefore, these new estimators, confidence intervals, and test statistics can be used to either estimate or test the relative risk of disease in matched case-control studies.

Keywords

disease prevalence; consistent estimator; constant and varying probabilities of exposure

1. Introduction

Epidemiologists and other medical researchers commonly use case-control studies when estimating the relative risk of disease. The objective of such studies is to determine if there is a disease-exposure association among a specified population of patients. More specifically, the disease-exposure association is determined by estimating the relative risk of disease. Throughout the paper we focus on a class of case-control studies where k controls are matched to each case. Generally, the assigned matching of the control(s) to the case is based on factors such as age, family history, gender, race, etc. Note that cases are interchangeable with persons who have disease, and controls are interchangeable with persons who have no disease. The terms 'disease' and 'no-disease' can be viewed as generic

[†]Both authors contributed equally to this work.

Copyright © 2011 John Wiley & Sons, Ltd.

*Correspondence to: Barry Kurt Moser, Department of Biostatistics and Bioinformatics and the CALGB Statistical Center, Duke University Medical Center, Box 2717, Durham, NC 27705, USA..

[†]moser004@mc.duke.edu

terms that describe any relevant dichotomous outcome. For example, the presence of a genetic marker can be used to identify a case (disease) and the absence of the marker used to describe a control (no-disease).

Throughout the paper, we discuss two different matched case-control scenarios, which we subsequently refer to as Scenarios I and II. In Scenario I the probability of exposure is assumed to be constant for all cases and controls. Scenario I is sometimes referred to as marginal estimation. In Scenario II the probability of exposure is allowed to vary across cases and controls. Scenario II is sometimes referred to as conditional estimation. For both scenarios the proposed estimators of the relative risk of disease share three main components: (i) an estimate of the odds ratio; (ii) an estimate of the relative risk of exposure, which is constant for Scenario I and varies for Scenario II; and (iii) a known prevalence of disease.

A rich literature exists on cohort designs, prospective sampling, and nested case-control studies including Refs [1–4]. In addition, the literature on matched case-control studies is extensive. We make no effort to provide a complete list of matched case-control articles here. We do however enumerate some of the major papers [5–23] that address the three main components of the proposed estimators that are listed in the preceding paragraph. In so doing we explain that our proposed estimators are not only related to existing estimators, but they also provide new approaches not available in the current literature.

- (a) The odds ratio: For Scenario I, Mantel and Haenszel [5], Miettinen [6], Pike and Morrow [8], and Ejigou and McHugh [9, 10] all discussed estimation of the odds ratio. For Scenario II, Greenhouse [11] cites Cornfield [12] as the first author to suggest using the logistic function to estimate the odds ratio in a case-control setting. Other authors who contributed to the case-control logistic regression literature are Anderson [13], Prentice and Pike [14], Breslow and Powers [15], and Breslow [16]. Conditional logistic regression is a common approach for analyzing matched case-control data for Scenario II. The use of conditional logistic regression in this context was discussed by Breslow *et al.* [17], Holford *et al.* [18], and Breslow and Day [19]. In most of the previously cited literature, the odds ratio is estimated under the rare disease assumption where the relative risk of disease is approximated by the odds ratio. In contrast to these authors, we do not make the rare disease assumption when we develop our estimators of the relative risk of disease for either Scenario I or II.
- (b) The relative risk of exposure: Any procedure that makes the rare disease assumption, and thus approximates the relative risk of disease with the odds ratio, also inherently implies that the relative risk of disease can be estimated by the relative risk of exposure. The contribution we make in Scenario I is to formulate a completely new estimate of the relative risk of exposure without making the rare disease assumption. In Scenario II we directly estimate the conditional probability of exposure given the disease status using a conditional logistic model. Then for both Scenarios I and II, we use Bayes theorem to combine the prevalence of disease, the relative risk of exposure, and the odds ratio to estimate the relative risk of disease.
- (c) The prevalence of disease: Known prevalence of disease has also been utilized in recent literature to estimate the relative risk of disease in matched case-control studies. For Scenario II, methods proposed by van der Laan [7], and discussed by Rose and van der Laan [20], use weighted logistic regression to estimate the relative risk of disease. The weight is based on known disease prevalence. The van der Laan estimator and our proposed estimator of the

relative risk of disease both employ known prevalence of disease, albeit in a different manner. We demonstrate by simulation and example however that the two approaches produce very similar estimates.

As a final comment on the existing literature, Zhang and Yu [21] proposed a simple method to estimate the relative risk of disease using the incidence of disease to correct the odds ratio. McNutt *et al.* [22] however observed that under Scenario II the correction proposed by Zhang produces biased estimates and biased confidence intervals (CIs) on the relative risk of disease because the correction does not account for any variability in the probability of exposure in neither the cases nor the controls. Zhang's biased results are also noted by Greenland [23]. Our proposed estimator of the relative risk of disease is similar to Zhang's estimator in one regard, both estimators are functions of the odds ratio. However, unlike Zhang's correction, our estimator accounts for any variability in the probability of exposure in both the cases and the controls. Furthermore, we demonstrate in a simulation study that our estimator and our confidence interval on the relative risk of disease are unbiased.

To clearly define the experimental framework for our methodology, we next describe the sampling scheme for Scenarios I and II. The framework resembles the sampling scheme described by van der Laan [7]. First, a categorical matching variable(s) is specified. A case is then sampled from the marginal distribution of the cases given the categorical matching variable. Then k controls are also sampled from the marginal distribution of the controls given the same value of the categorical matching variable. In addition, for Scenario II the case and the k controls are also identified by a vector of covariates z , which are believed to affect the probability of exposure for the cases and controls. A case is then sampled from the conditional distribution of the cases given a categorical matching variable and conditioned on a vector of covariates. Similarly, each of k controls is then sampled from the conditional distribution of the controls given the same categorical matching variable and conditioned on a vector of covariates. Therefore, the analysis unit is identified by a $(k+1)$ -tuple variable of a case and k controls that is linked by a categorical matching variable and contains a vector of covariates when the probability of exposure varies. For both Scenarios I and II the cases and the controls are assumed to be independent.

The paper proceeds in the following order: In Section 2 we provide the notation and a general formula for the relative risk of disease for a known prevalence of disease. In Section 3 we derive an estimate of the relative risk of disease for Scenario I when $k = 1$. We then provide specific formulae for the estimators, variances, confidence intervals, and test statistics of the relative risk of disease when $k = 1, 2$. In Section 4 we develop an estimate of the relative risk of disease for Scenario II when $k = 1$. The estimation procedure for Scenario II uses the conditional logistic regression model with a set of covariates z . Again, we provide specific formulae for the estimators, variances, confidence intervals, and test statistics of the relative risk of disease when $k = 1$. In Section 5, to study the performance of the new estimators for both Scenarios I and II, we report Monte Carlo simulation results when $k = 1, 2$. In Section 6 we provide a real-life example of a matched case-control study where we perform the numerical calculations for our estimation procedure. In Section 7, we discuss the results of the new estimators and their implications. Finally, the Appendices are provided online at the link: <http://www.duke.edu/~shalabi/SIM/Appendices.pdf>.

2. Notation and a general form for the relative risk of disease

The following notation will be used for Scenario I: Let ρ_D = the relative risk of disease, ρ_E = the relative risk of exposure, γ = the known prevalence of disease, ω_D = the odds ratio of disease, and ω_E = the odds ratio of exposure. For Scenario II, when the probability of exposure varies across cases and controls as a function of covariates z , the preceding terms are designated as $\rho_D(z)$, $\rho_E(z)$, $\gamma(z)$, $\omega_D(z)$, and $\omega_E(z)$.

The relative risk of disease ρ_D is defined as the ratio of the probability of disease given exposure to the probability of disease given no exposure. The relative risk of exposure ρ_E is defined as the ratio of the probability of exposure given disease to the probability of exposure given no disease. The odds ratio of disease ω_D is the probability of disease given exposure over one minus the probability of disease given exposure, divided by the probability of disease given no exposure over one minus the probability of disease given no exposure. The odds ratio of exposure ω_E is the probability of exposure given disease over one minus the probability of exposure given disease, divided by the probability of exposure given no disease over one minus the probability of exposure given no disease.

Note that in Scenario I, $\omega_D = \omega_E = \omega$. Similarly, in Scenario II when the logistic model is used to represent the probability of disease or the probability of exposure, we also obtain $\omega_D(z) = \omega_E(z) = \omega$. For further elaboration on this latter relationship see the discussion after Equation (18). Therefore, at times we simply refer to the odds ratio, with the understanding that the odds ratio of disease is equivalent to the odds ratio of exposure.

Using the preceding notation, the following general form of the relative risk of disease is derived in Appendix 1.

$$\rho_D(z) = \frac{\gamma(z)\rho_E(z) + [1 - \gamma(z)]\omega_E(z)}{\gamma(z)\rho_E(z) + [1 - \gamma(z)]}. \quad (1)$$

From (1), as the prevalence of disease $\gamma(z)$ tends to 0, the relative risk of disease $\rho_D(z)$ tends to the odds ratio $\omega_D(z) = \omega_E(z) = \omega$. That is, (1) is consistent with the fact that under the rare disease assumption, the odds ratio and the relative risk of disease are equivalent. More importantly, (1) provides a formulation of the relative risk of disease for any known disease prevalence.

For Scenario II, Equation (1) is presented in its most general form; that is, the prevalence of disease, the relative risk of exposure, and the odds ratio of exposure vary across cases and controls as a function of the covariate z . Without loss of generality, when the known prevalence of disease is a constant across cases and controls, $\gamma(z)$ can be simplified with γ . However, if the prevalence of disease is known to vary as a function of the covariates z , then all subsequent formulas can be reset to the designation $\gamma(z)$.

Finally, Equation (1) also holds under Scenario I. This simplified form is

$$\rho_D = \frac{\gamma\rho_E + (1 - \gamma)\omega}{\gamma\rho_E + (1 - \gamma)}. \quad (2)$$

3. The relative risk of disease for Scenario I

3.1. Joint and marginal distributions of the case-control combinations

In this section we illustrate the concepts described in the preceding paragraphs by using a real-life example of a matched case-control study reported in [24]. Cases are randomly sampled from women who developed a venous, arterial embolism or deep venous thrombosis while taking adjuvant tamoxifen for stage I, II or IIIA breast cancer identified from 34 Cancer and Leukemia Group B (CALGB) institutions. For the purpose of case eligibility, catheter-associated thrombosis sufficient to require removal of the catheter is considered a thrombotic event (TE). For each randomly selected case, two controls are randomly sampled from women who are enrolled at the same CALGB institution, whose age

at breast cancer diagnosis is within 5 years of the selected case, are also taking adjuvant tamoxifen, and do not experience TE. That is, the CALGB institution and the age of breast cancer diagnosis are the matching variables. This example can also be used to illustrate Scenario II with the addition of the confounding factors (covariates): family history of TE and smoking history.

The exposure variable is factor V Leiden among breast cancer TE cases and among breast cancer non-TE matched controls. The presence and absence of factor V Leiden for the case or k controls is designated by (+) and (−), respectively. The term $p_i(+)$ denotes the probability of observing a matched set in which the case is exposed (+) and exactly i of the k controls are exposed for $i = 0, 1, \dots, k$. Similarly, the term $p_i(-)$ denotes the probability of observing a matched set in which the case is not exposed (−) and exactly i of the k controls are exposed. The sum of all $2(k + 1)$ probabilities equals one (i.e., $\sum_{i=0}^k [p_i(+) + p_i(-)] = 1$). The $2k$ combinations associated with the probabilities $p_0(+), \dots, p_{k-1}(+), p_1(-), \dots, p_k(-)$ refer to the discordant combinations. In the discordant combinations: (i) if the case is exposed then at least one the matched controls are not exposed or (ii) if the case is not exposed then at least one of the matched controls is exposed. The proportion of discordant combinations is therefore given by $0 \left(p = \sum_{i=0}^{k-1} p_i(+) + \sum_{i=1}^k p_i(-) \right) < 1$.

The joint distribution of each case and the k matched controls can be represented through two random variables:

$$X_1 = \begin{cases} 1 & \text{if the case is exposed} \\ 0 & \text{if the case is not-exposed} \end{cases} \quad (3)$$

and

$$X_2 = \begin{cases} k & \text{if all } k \text{ controls are exposed} \\ k-1 & \text{if exactly } k-1 \text{ controls are exposed} \\ \vdots & \vdots \\ 1 & \text{if exactly one control is exposed} \\ 0 & \text{if zero controls are exposed} \end{cases}.$$

The variables X_1, X_2 represent the number of exposed among cases and controls, respectively, and are assumed to be independent. Q_1 is the probability of a single case being exposed (constant for all cases), whereas Q_2 is the probability of a single control being exposed (constant for all controls). Under these conditions, $X_1 \sim \text{Bernoulli}(Q_1)$, $X_2 \sim \text{Binomial}(Q_2, k)$, and the joint distribution of X_1, X_2 equals

$$p_i(+) = P(X_1=1, X_2=i) = k! Q_1 Q_2^i (1 - Q_2)^{k-i} / [i! (k-i)!] \quad (4)$$

and

$$p_i(-) = P(X_1=0, X_2=i) = k! (1 - Q_1) Q_2^i (1 - Q_2)^{k-i} / [i! (k-i)!]$$

for $i = 0, 1, 2, \dots, k$. Because cases have the disease and controls do not, Q_1 and Q_2 can also be understood as the probability of exposure given disease and the probability of exposure given no disease, respectively. Therefore, the relative risk of exposure is given by $p_E = Q_1/Q_2$ and the odds ratio $\omega = [Q_1/(1 - Q_1)]/[Q_2/(1 - Q_2)]$.

3.2. Derivation of the relative risk of disease ρ_D for $k > 1$

We now derive the relative risk of disease ρ_D in terms of the $2k$ discordant probabilities $p_0(+), \dots, p_{k-1}(+), p_1(-), \dots, p_k(-)$ when $k > 1$. From (4), for any even k , the odds of exposure for a control are given by

$$\theta = Q_2 / (1 - Q_2) = 2k^{-1} \sum_{j=0}^{\frac{(k-2)}{2}} \frac{a_{2j+1}^{-1} p_{2j+1}(+) + a_{k-2j}^{-1} p_{k-2j}(-)}{a_{2j}^{-1} p_{2j}(+) + a_{k-2j-1}^{-1} p_{k-2j-1}(-)}. \quad (5)$$

Similarly, for any odd k , the odds of exposure for a control are given by

$$\theta = Q_2 / (1 - Q_2) = (k-1)^{-1} \left[\sum_{j=0}^{\frac{(k-3)}{2}} \frac{a_{2j+1}^{-1} p_{2j+1}(+) + a_{k-2j}^{-1} p_{k-2j}(-)}{a_{2j}^{-1} p_{2j}(+) + a_{k-2j-1}^{-1} p_{k-2j-1}(-)} + \sum_{j=0}^{\frac{(k-3)}{2}} \frac{a_{2(j+1)}^{-1} p_{2(j+1)}(+) + a_{k-2j-1}^{-1} p_{k-2j-1}(-)}{a_{2j+1}^{-1} p_{2j+1}(+) + a_{k-2(j+1)}^{-1} p_{k-2(j+1)}(-)} \right], \quad (6)$$

where $a_r = k! / [r!(k-r)!]$, $r = 0, 1, \dots, k$. In addition, using the definition of the odds ratio ω and the definition of the relative risk of exposure ρ_E from (4), we obtain for all $k \geq 1$

$$\omega = \frac{Q_1 / (1 - Q_1)}{Q_2 / (1 - Q_2)} = \frac{\rho_E (1 - Q_2)}{(1 - \rho_E Q_2)} = \frac{\sum_{j=0}^{k-1} (k-j) p_j(+)}{\sum_{j=1}^k j p_j(-)}. \quad (7)$$

Finally, substituting $Q_2 = \theta / (1 + \theta)$ into (7) and solving for ρ_E we obtain:

$$\rho_E = \omega (1 + \theta) / (1 + \theta \omega). \quad (8)$$

Then, for a given disease prevalence γ and for any $k \geq 1$, the relative risk of disease ρ_D is given by (2) where ω is given in (7) and ρ_E is given in (8). Therefore, ω from (7), θ from (5) or (6), ρ_E from (8), and ρ_D from (2) are closed form functions of the $2k$ discordant probabilities $p_0(+), \dots, p_{k-1}(+), p_1(-), \dots, p_k(-), \dots, p_k(\dots)$ for all $k > 1$.

Next, denote $n_j(+)$ as the number of observed matched sets in which the case is exposed (+) and exactly j of the k matched controls are exposed. Similarly, denote $n_j(-)$ as number of observed matched sets in which the case is not-exposed (-) and exactly j of the k matched controls are exposed. Let n denote the total number of observed matched sets for any k .

Therefore $n = \sum_{j=0}^k [n_j(+) + n_j(-)]$. Then replace $p_j(+)$ by $n_j(+)/n$ when $j = 0, \dots, k-1$ and $p_j(-)$ by $n_j(-)/n$ when $j = 1, \dots, k$ in (5), (6), (7), and (8). The resulting value provides a consistent estimator of ρ_D in (2).

We summarize the preceding derivation for $k > 1$ with these observations. First, the estimator of ω , created by replacing $p_j(+)$ by $n_j(+)$ and $p_j(-)$ by $n_j(-)$ in (7) is not a new formulation. The estimator was previously shown in [6] to represent the Mantel-Haenszel estimator of the odds ratio for any $k \geq 1$. Second, to the best of our knowledge, the closed form function of the odds of exposure for a control (θ) in (5) or (6) is a new formulation. Third, ω in (7) and in θ in (5) or (6) are then used in (8) to construct a new formulation of ρ_E . Fourth, for a given prevalence of disease γ , the relative risk of disease ρ_D in (2) is derived in Appendix 1 using Bayes' Theorem. All of these derivations are developed without making the rare disease assumption. Then, a consistent estimator of the relative risk of

disease p_D is constructed by replacing $p_j(+)$, $p_j(-)$ by $n_j(+)/n$, $n_j(-)/n$, respectively, in ω and θ .

3.3. Derivation of the relative risk of disease p_D for $k = 1$

We now derive the relative risk of disease p_D for $k = 1$ in terms of the two discordant probabilities $p_0(+)$, $p_1(-)$. From (4) for $k = 1$ we have

$$\begin{aligned} p_0(+) &= Q_1(1 - Q_2) \\ p_1(-) &= Q_2(1 - Q_1). \end{aligned} \quad (9)$$

By solving the two equations in (9) simultaneously for Q_1 , Q_2 (probability of exposure given disease and the probability of exposure given no disease, respectively) in terms of $p_0(+)$, $p_1(-)$ we obtain

$$\begin{aligned} Q_1 &= 2p_0(+) / \left\{ 1 + [p_0(+) - p_1(-)] + \sqrt{1 - [p_0(+) - p_1(-)]^2 - 4p_1(-)} \right\} \\ Q_2 &= \left\{ 1 - [p_0(+) - p_1(-)] - \sqrt{1 - [p_0(+) - p_1(-)]^2 - 4p_1(-)} \right\} / 2. \end{aligned} \quad (10)$$

Substituting (10) into the relative risk of exposure $\rho_E = Q_1/Q_2$ produces

$$\begin{aligned} \rho_E &= 4p_0(+) / (ab) \\ a &= 1 + [p_0(+) - p_1(-)] + \sqrt{c} \\ b &= 1 - [p_0(+) - p_1(-)] - \sqrt{c} \\ c &= \{1 - [p_0(+) - p_1(-)]\}^2 - 4p_1(-). \end{aligned} \quad (11)$$

The odds ratio is given by

$$\omega = p_0(+) / p_1(-). \quad (12)$$

Finally, for a known prevalence of disease, substituting ρ_E from (11) and ω from (12) into (2) provides the relative risk of disease p_D as a closed form function of $p_0(+)$, $p_1(-)$.

3.4. Relative risk of disease estimators, test statistics, and confidence intervals for $k = 1, 2$

For $k = 1$, replace $p_0(+)$ and $p_1(-)$ by $n_0(+)/n$ and $n_1(-)/n$ in (11), (12), and (2). Then a consistent estimator of the relative risk of disease p_D takes the exact form

$$\widehat{\rho}_{D1} = \frac{\gamma \widehat{\rho}_{E1} + (1 - \gamma) \widehat{\omega}_1}{\gamma \widehat{\rho}_{E1} + (1 - \gamma)} \quad (13)$$

$$\widehat{\rho}_{E1} = 4n_0(+) / (\widehat{nab}),$$

$$\widehat{\omega}_1 = n_0(+) / n_1(-),$$

$$\widehat{a}=1+[n_0(+)-n_1(-)]/n+\sqrt{\widehat{c}},$$

$$\widehat{b}=1-(n_0(+)-n_1(-))/n-\sqrt{\widehat{c}},$$

$$\widehat{c}=\{1-[n_0(+)-n_1(-)]/n\}^2-4n_1(-)/n.$$

The subscripts D, E, 1 denote disease, exposure, and $k = 1$, respectively.

For $k = 2$, replace $p_0(+)$, $p_1(-)$, $p_1(-)$ $p_2(-)$ by $n_0(+)/n$, $n_1(+)/n$, $n_1(-)/n$, $n_2(-)/n$ in (2), (5), (7), and (8). Then a consistent estimator of the relative risk of disease ρ_D takes the exact form

$$\widehat{\rho}_{D2} = \frac{\gamma \widehat{\rho}_{E2} + (1 - \gamma) \widehat{\omega}_2}{\gamma \widehat{\rho}_{E2} + (1 - \gamma)} \quad (14)$$

$$\widehat{\rho}_{E2} = \frac{[2n_0(+) + n_1(+)] [2n_0(+) + n_1(+) + n_1(-) + 2n_2(-)]}{[2n_0(+) + n_1(+)] [n_1(+) + 2n_2(-)] + [2n_0(+) + n_1(-)] [n_1(-) + 2n_2(-)]},$$

$$\widehat{\omega}_2 = \frac{2n_0(+) + n_1(+)}{n_1(-) + 2n_2(-)}.$$

The subscripts D, E, 2 denote disease, exposure, and $k = 2$, respectively. Note that the odds ratio estimates $\widehat{\omega}_1$, $\widehat{\omega}_2$ used in (13), (14) are the Mantel-Haenszel estimators of ω for $k = 1, 2$, respectively.

The statistic

$$S_k = [\log(\widehat{\rho}_{Dk})]^2 / \widehat{\text{var}}[\log(\widehat{\rho}_{Dk})] \sim \chi_1^2(\lambda),$$

where $\widehat{\text{var}}[\log(\widehat{\rho}_{Dk})] = \widehat{\Phi} \widehat{\Sigma}_{|n_d} \widehat{\Phi}^T$ with the $2k \times 2k$ estimated covariance matrix $\widehat{\Sigma}_{|n_d}$, the $1 \times 2k$ estimated matrix $\widehat{\Phi}$, and λ is the chi-square noncentrality parameter. The matrix $\widehat{\Sigma}_{|n_d}$ is defined in Appendix 2 for $k = 1$. The matrix Φ is derived in Appendix 3 when $k = 1, 2$.

Therefore, for an α level test and any $k = 1$, reject $H_0 : \log(\rho_D) = 0$ in favor of $H_1 : \log(\rho_D) > 0$ (or equivalently, reject $H_0 : \rho_D = 1$ in favor of $H_1 : \rho_D > 1$) if $S_k > \chi_{1,\alpha}^2$. The term $\chi_{1,\alpha}^2$ is the $100(1 - \alpha)$ percentile point of a central chi-square random variable with one degree of freedom. In addition, a $100(1 - \alpha)\%$ confidence interval on the relative risk of disease ρ_D for any $k = 1$ is given by

$$\exp(L_{\rho_{Dk}}), \exp(U_{\rho_{Dk}}) \quad (16)$$

where $L_{\rho Dk} = \log(\widehat{\rho}_{Dk}) - v^{\alpha/2} \sqrt{\text{var}[\log(\widehat{\rho}_{Dk})]}$, $U_{\rho Dk} = \log(\widehat{\rho}_{Dk}) - v^{\alpha} \sqrt{\text{var}[\log(\widehat{\rho}_{Dk})]}$ and v^{α} is the $100(1 - \alpha)$ percentile point of a standard normal distribution.

4. The relative risk of disease for Scenario II

4.1. Using logistic regression to model the probability of exposure for any $k \geq 1$

In case-control studies the influence of extraneous factors almost always produces the risk of bias. These factors cause the probabilities of exposure to vary across cases and controls. The standard approach is to use a logistic regression to model these varying probabilities, with a set of covariates z associated with the confounding factors.

Formally, consider the multivariable logistic regression that models the probability of exposure (E) given the covariates z :

$$P(E|x, z) = \exp(\alpha_0 + \alpha_1 x + \beta' z) / [1 + \exp(\alpha_0 + \alpha_1 x + \beta' z)], \quad (17)$$

where $x = 1$ signifies a case (= disease or D), $x = 0$ signifies a control (= no-disease or ND), α_0 is the intercept, α_1 is the regression parameter for the disease status, and β is the vector of regression parameters for the covariates z . Therefore, the relative risk of exposure is given by

$$\begin{aligned} \rho_E(z) &= \frac{P(E|D, z)}{P(E|ND, z)} = \frac{P(E|x=1, z)}{P(E|x=0, z)} \\ &= \left\{ \exp(\alpha_0 + \alpha_1 + \beta' z) / [1 + \exp(\alpha_0 + \alpha_1 + \beta' z)] \right\} / \left\{ \exp(\alpha_0 + \beta' z) / [1 + \exp(\alpha_0 + \beta' z)] \right\} \\ &= \exp(\alpha_1) [1 + \exp(\alpha_0 + \beta' z)] / [1 + \exp(\alpha_0 + \alpha_1 + \beta' z)]. \end{aligned} \quad (18)$$

By similar calculations, the odds ratio of exposure is $\omega_E(z) = \omega_D(z) = \omega = \exp(\alpha_1)$.

Combining (18) with (1) the relative risk of disease becomes a closed form function of γ , α_0 , α_1 , β , z :

$$\begin{aligned} \rho_D(z) &= \frac{\gamma \exp(\alpha_1) [1 + \exp(\alpha_0 + \beta' z)] / [1 + \exp(\alpha_0 + \alpha_1 + \beta' z)] + [1 - \gamma] \exp(\alpha_1)}{\gamma \exp(\alpha_1) [1 + \exp(\alpha_0 + \beta' z)] / [1 + \exp(\alpha_0 + \alpha_1 + \beta' z)] + [1 - \gamma]} \\ &= \frac{\exp(\alpha_1) [\gamma [1 + \exp(\alpha_0 + \beta' z)] + [1 - \gamma] [1 + \exp(\alpha_0 + \alpha_1 + \beta' z)]]}{\gamma \exp(\alpha_1) [1 + \exp(\alpha_0 + \beta' z)] + [1 - \gamma] [1 + \exp(\alpha_0 + \alpha_1 + \beta' z)]}. \end{aligned} \quad (19)$$

4.2. Estimation and testing of the relative risk of disease $\rho_D(z)$

The software routine SAS PROC LOGISTIC (SAS Institute Inc., Cary, NC, USA), or any comparable routine, can be used to estimate the $p + 1$ parameters α_1 , β and their $(p + 1) \times (p + 1)$ covariance matrix $\Sigma_{\widehat{\alpha}_1, \widehat{\beta}}$.

Using model (18), the cases and controls from the $n_d = \sum_{i=0}^{k-1} n_i (+) + \sum_{i=0}^k n_i (-)$ discordant matched sets are sufficient for estimating α_1 , β and their $(p + 1) \times (p + 1)$ covariance matrix $\Sigma_{\widehat{\alpha}_1, \widehat{\beta}}$. Therefore, only the discordant matched case controls are used in the estimation of α_1 , β . When SAS PROC LOGISTIC is used for the estimation, the $(k + 1)n_d$ discordant cases and controls are listed as separate observations. In addition, each matched set of k controls with their associated case is identified by a stratum value. The MODEL statement identifies a dichotomous exposure variable on the left side of the model with a dichotomous disease effect, and a list of the covariates z on the right side of the model. A STRATA statement is

also used with the stratum variable to account for the correlation between the case and its associated controls in each matched set.

The SAS output produces estimates $\widehat{\alpha}_1$, $\widehat{\beta}$ and their estimated covariance matrix $\widehat{\Sigma}_{\widehat{\alpha}_1, \widehat{\beta}}$. An estimate of α_0 is obtained through the relationship

$$\widehat{\alpha}_0 = \log \left[\bar{e} / (1 - \bar{e}) \right] - \widehat{\alpha}_1 / (k+1) - \widehat{\beta}' \bar{z}, \quad (20)$$

where

$$\bar{e} = \left[\sum_{i=0}^k [(i+1) n_i(+)+i n_i(-)] \right] / [(k+1) n].$$

Calculated over all $(k+1)n$ cases and controls, the term $\bar{e}$ is the observed proportion of exposures and $\bar{z}$ is the $p \times 1$ mean vector of covariates. For any $k \geq 1$ and any $p \times 1$ vector z , $\widehat{\rho}_{Dk}(z)$ is then the estimate of the relative risk of disease $\rho_D(z)$ calculated by replacing α_0 , α_1 , β in (19) by $\widehat{\alpha}_0$, $\widehat{\alpha}_1$, $\widehat{\beta}$.

The variance of the natural logarithm of the estimated relative risk $\log [\widehat{\rho}_{Dk}(z)]$ is given by

$$\text{var} \{ \log [\widehat{\rho}_{Dk}(z)] \} = (1 - 1) \Psi \Sigma_{\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\beta}} \Psi' (1 - 1)' \quad (21)$$

$$\Sigma_{\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\beta}} = \begin{bmatrix} 1 & -1/(k+1) & -\bar{z}' \\ 0 & 1 & 0 \\ 0 & 0 & I_p \end{bmatrix} \Sigma_{\log[\bar{e}/(1-\bar{e})], \widehat{\alpha}_1, \widehat{\beta}} \begin{bmatrix} 1 & 0 & 0 \\ -1/(k+1) & 1 & 0 \\ -\bar{z} & 0 & I_p \end{bmatrix}$$

$$\Sigma_{\log[\bar{e}/(1-\bar{e})], \widehat{\alpha}_1, \widehat{\beta}} = \begin{bmatrix} \text{var} \left\{ \log \left[\bar{e} / (1 - \bar{e}) \right] \right\} & 0 \\ 0 & \Sigma_{\widehat{\alpha}_1, \widehat{\beta}} \end{bmatrix}.$$

$$\text{var} \left\{ \log \left[\bar{e} / (1 - \bar{e}) \right] \right\} = \left[\bar{e} (1 - \bar{e}) (k+1) n \right]^{-2} t' \Sigma_{|n_d|} t$$

$$t = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \otimes (1, \dots, k-1, k)'$$

The symbol $\otimes$ represents a Kronecker product. The matrix $\Sigma_{|n_d|}$ is the $2k \times 2k$ covariance matrix of $n_0(+), \dots, n_{k-1}(+), n_1(-), \dots, n_k(-)$ given n_d provided in Appendix 2. For $k \geq 1$, the $2 \times (p+2)$ matrix Ψ is given by

$$\Psi = \begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{21} & \psi_{22} & \psi_{23} \end{bmatrix}. \quad (22)$$

$$\psi_{11} = \{\gamma r + [1 - \gamma] s\} / \{\gamma (1+r) + [1 - \gamma] (1+s)\}$$

$$\psi_{12} = 1 + [1 - \gamma] s / \{\gamma (1+r) + [1 - \gamma] (1+s)\}$$

$$\psi_{13} = \psi_{11} z'$$

$$\psi_{21} = s / \{\gamma (1+r) \exp(\alpha_1) + [1 - \gamma] (1+s)\}$$

$$\psi_{22} = \{\gamma (1+r) \exp(\alpha_1) + [1 - \gamma] s\} / \{\gamma (1+r) \exp(\alpha_1) + [1 - \gamma] (1+s)\}$$

$$\psi_{23} = \psi_{21} z'$$

$$r = \exp(\alpha_0 + \beta' z)$$

$$s = \exp(\alpha_1) r = \exp(\alpha_0 + \alpha_1 + \beta' z).$$

The statistic

$$S_k = \{\log[\widehat{\rho}_{Dk}(z)]\}^2 / \widehat{\text{var}}\{\log[\widehat{\rho}_{Dk}(z)]\} \sim \chi_1^2(\lambda), \quad (23)$$

where $\widehat{\text{var}}\{\log[\widehat{\rho}_{Dk}(z)]\}$ is given by (21) with $\Sigma_{|n_d}$ replaced by $\widehat{\Sigma}_{|n_d}$, $\Sigma_{\widehat{\alpha}_1, \widehat{\beta}}$ replaced by $\widehat{\Sigma}_{\widehat{\alpha}_1, \widehat{\beta}}$, and $\alpha_0, \alpha_1, \beta$ in Ψ replaced by $\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\beta}$. The term λ is the noncentrality parameter. Therefore, for an α level test, a specified $p \times 1$ vector z , and any $k = 1$, reject $H_0 : \log[\rho_{Dk}(z)] = 0$ in favor of $H_1 : \log[\rho_{Dk}(z)] > 0$ (or equivalently, reject $H_0 : \rho_D(z) = 1$ in favor of $H_1 : \rho_D(z) > 1$) if $S_k > \chi_{1, \alpha}^2$.

In addition, for a specified $p \times 1$ vector z , a $100(1 - \alpha)\%$ confidence interval on the relative risk of disease $\rho_D(z)$ for any $k = 1$ takes the form

$$\exp(L_{\rho_{Dk}(z)}), \exp(U_{\rho_{Dk}(z)}) \quad (24)$$

where $L_{\rho Dk}(z) = \log[\widehat{\rho}_{Dk}(z)] - v^{\alpha/2} \sqrt{\widehat{\text{var}}\{\log[\widehat{\rho}_{Dk}(z)]\}}$ and
 $U_{\rho Dk}(z) = \log[\widehat{\rho}_{Dk}(z)] + v^{\alpha/2} \sqrt{\widehat{\text{var}}\{\log[\widehat{\rho}_{Dk}(z)]\}}.$

For Scenario II the logistic model has historically been used to estimate the odds ratio. Under the rare disease assumption, the odds ratio estimator provides an alternative estimate of the relative risk of disease. To estimate the odds ratio, the following logistic model is used to represent the probability of disease (D):

$$P(D|u, z) = \exp(\tau_0 + \tau_1 u + \eta' z) / [1 + \exp(\tau_0 + \tau_1 u + \eta' z)]. \quad (25)$$

where $u = 0$ indicates no-exposure (NE), $u = 1$ indicates exposure (E), τ_0 is the intercept, τ_1 is the exposure parameter, and η is the $p \times 1$ vector of parameters associated with z . Using model (25) the odds ratio is then given by

$$\begin{aligned} \omega_D(z) = \omega_E(z) &= [P(D|u=1, z) / P(ND|u=1, z)] / [P(D|u=0, z) / P(ND|u=0, z)] \\ &= \frac{\{\exp(\tau_0 + \tau_1 + \eta' z) / [1 + \exp(\tau_0 + \tau_1 + \eta' z)]\} / \{1 / [1 + \exp(\tau_0 + \tau_1 + \eta' z)]\}}{\{\exp(\tau_0 + \eta' z) / [1 + \exp(\tau_0 + \eta' z)]\} / \{1 / [1 + \exp(\tau_0 + \eta' z)]\}} \\ &= \exp(\tau_1) = \omega \end{aligned} \quad (26)$$

The software routine SAS PROC LOGISTIC, or any comparable routine, can be used to estimate τ_1 and the variance of the estimated odds ratio $\widehat{\omega}_k = \exp(\widehat{\tau}_{1k})$. Unlike model (18), when using model (25) the $(k+1)n_d$ cases and controls from the discordant matched sets are not sufficient for estimating τ_1 , and their $(p+1) \times (p+1)$ covariance matrix $\Sigma_{\widehat{\tau}_1, \widehat{\eta}}$. Rather, all $(k+1)n$ matched cases and controls are used in the estimation procedure. The SAS PROC LOGISTIC program statements are constructed similarly to those described in Section 4.2 except now: (i) all $(k+1)n$ matched cases and controls are used; (ii) the MODEL statement now identifies the dichotomous disease status on the left side of the model; and (iii) the right side of the model contains a dichotomous exposure effect and the list of the covariates z .

More recently, van der Laan [7] suggested using a weighted logistic regression model to estimate the relative risk of disease. His estimate of $\rho_{Dk}(z)$ is calculated using model (25) where cases and controls are weighted by the prevalence of disease γ and $(1 - \gamma)/k$, respectively. Again, the routine SAS PROC LOGISTIC can be used to estimate $P(D|u, z)$ (and therefore $\rho_{Dk}(z)$) for any set of covariates z .

5. Simulations

5.1. Simulation design for Scenario I

Appendix 4 ($k = 1$) and Appendix 5 ($k = 2$) describe how values of $n_0(+)$, $n_1(+)$, $n_0(-)$, $n_1(-)$ and $n_0(+)$, $n_1(+)$, $n_2(+)$, $n_0(-)$, $n_1(-)$, $n_2(-)$ are simulated for combinations of n , p , γ , and ρ_D . We perform simulation studies for combinations of $n = 100, 200, 300, 500$, $p = 0.05, 0.1, 0.15, 0.2, 0.25, 0.5$, $\gamma = 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.5$, and $\rho = 0.6, 0.75, 0.8, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0$. For $k = 1$, we compare $\widehat{\rho}_{D1}$ and the Mantel-Haenszel estimator $\widehat{\omega}_{MH1}$ in terms of their abilities to estimate the relative risk of disease ρ_D . Similarly, for $k = 2$ we compare $\widehat{\rho}_{D2}$, the Mantel-Haenszel estimator $\widehat{\omega}_{MH2}$, and the Miettinen estimator $\widehat{\omega}_{D2}$. For both $k = 1$ and $k = 2$, we perform 100,000 simulations for each combination of n , p , γ , and ρ . We then summarize the simulated results through means, mean relative biases, mean square errors, and 95% probabilities of coverage on ρ_D .

5.2. Simulation design for Scenario II

Appendix 6 ($k = 1$) and Appendix 7 ($k = 2$) describe how values of $n_0(+)$, $n_1(+)$, $n_0(-)$, $n_1(-)$ and $n_0(+)$, $n_1(+)$, $n_2(+)$, $n_0(-)$, $n_1(-)$, $n_2(-)$ are simulated for combinations of n , p , γ , $\rho_D(z)$, β , and z . In these studies simulation runs are performed for combinations of $n = 200, 500$, $p = 0.05, 0.1, 0.2, 0.25$, $\gamma = 0.05, 0.1, 0.2, 0.3$, $\rho_{Dk}(z) = 0.6, 0.8, 1.0, 1.5, 2.0$, $\beta = 0.5, 1.0$, and $z = 0.5, 1.0$.

The van der Laan [7] relative risk of disease estimator is calculated as a weighted average over all observed case vectors z_i for $i = 1, \dots, n$ and all observed control vectors z_j for $j = 1, \dots, n_k$. Our relative risk of disease estimator is calculated at a single specified vector z . Therefore, for comparison purposes, the van der Laan procedure is modified to estimate the relative risk of disease at a single specified vector z ([20], Section 3.2.5). Our estimator $\widehat{\rho}_{Dk}(z)$ is then compared to the van der Laan relative risk of disease estimator and to the odds ratio estimator $\widehat{\omega}_k = \exp(\widehat{\tau}_1 k)$ from (26) for the combinations of n , p , γ , $\rho_D(z)$, β , and z listed in the preceding paragraph for $k = 1, 2$.

For both values of k , 10,000 simulations are run for each combination of n , p , γ , $\rho_D(z)$, β , and z . We then summarize the simulated results through means, mean relative biases, mean square errors, and 95% probabilities of coverage on $\rho_D(z)$.

5.3. Simulation results

For Scenario I, Table I provides the means, mean relative biases, mean square errors, and 95% coverage probabilities on ρ_D when $k = 2$. Only combinations of $p = 0.25$, $E(n_d) = np = 50$, $\gamma = 0.05, 0.1, 0.30$, and $\rho_D = 0.6, 1.0, 1.5, 2.0$ are presented as they are representative of the patterns produced by the other combinations.

In Table I both our estimator $\widehat{\rho}_{D2}$ and the Mantel–Haenszel, Miettinen estimators $\widehat{\omega}_{MH2}$, $\widehat{\omega}_{M2}$ provide similar relative biases (relative to ρ_D) and mean square errors when the prevalence of disease γ is small (< 0.05) and the relative risk of disease ρ is small (< 1.0). However, as γ and ρ increase, relative biases and mean square errors of $\widehat{\rho}_{D2}$ become much smaller in relation to the relative biases and mean square errors of $\widehat{\omega}_{MH2}$, $\widehat{\omega}_{M2}$. For example, when $\gamma = 0.3$, $\rho_D = 2.0$ the relative bias from $\widehat{\omega}_{MH2}$, $\widehat{\omega}_{M2}$ are both about 72 ($\approx 0.717/0.01$) times larger than the relative bias from $\widehat{\rho}_{D2}$. For the same combination of γ , ρ_D , the mean square error from $\widehat{\omega}_{MH2}$, $\widehat{\omega}_{M2}$ are both about 37 ($\approx 3.3/0.09$) times larger than the mean square error from $\widehat{\rho}_{D2}$.

The 95% probabilities of coverage on $\widehat{\rho}_D$ are between 0.950 and 0.954 for all combinations of γ , p except when $\gamma = 0.3$, $p = 2.0$ where the coverage probability is 0.959. In contrast, the probabilities of coverage on $\widehat{\omega}_{MH2}$ and $\widehat{\omega}_{M2}$ are close to 0.95 except when $\gamma = 0.1$ and $\rho_D = 2$. For example, when $\gamma = 0.3$, $p = 2.0$, the probability of coverage is 0.640. In conclusion, when estimating the relative risk of disease, our estimator $\widehat{\rho}_{Mk}$ performs substantially better than either the Mantel–Haenzel or the Miettinen estimator whenever the disease prevalence, $\gamma > 0.05$ and the relative risk of disease, $\rho_D > 1$.

For Scenario II, Table II provides the means, mean relative bias, mean square errors, and 95% coverage probabilities when $k = 2$. Only combinations of $p = 0.25$, $E(n_d) = np = 50$, $\gamma = 0.05, 0.1, 0.3$, $\rho_D(z) = 0.6, 1.0, 1.5, 2.0$, $z = 0.5$, and $\beta = 1$ are presented as they are representative of the patterns produced by the other combinations. Comparisons of our estimator of the relative risk of disease $\widehat{\rho}_{D2}(z=0.5)$, the logistic regression odds ratio estimator $\widehat{\omega}_k = \exp(\widehat{\tau}_1 k)$, and van der Laan's estimator at $z = 0.5$ are made from Table II. When $\rho_D(z) = 0.6, 1$ and $\gamma = 0.05, 0.1$ all estimators produced means close to the parameters,

have similar relative biases and mean square errors, and have 95% probabilities of coverage on $p_D(z)$. Our estimator $\hat{\rho}_{Dk}$ ($z=0.05$), however, performs substantially better than the logistic regression odds ratio estimator $\hat{\omega}_k = \exp(\hat{\tau}_{1k})$ when $p_D(z) > 1$ or $\gamma > 0.1$. In addition, our estimator and the van der Laan weighted logistic estimator perform equally well.

5.4. Sensitivity results

Because the prevalence of disease γ is assumed to be known, it is instructive to determine how sensitive our estimator (and van der Laan's estimator) are to misspecification of γ . The two estimators are presented in Table III for the combinations of $p = 0.25$, $E(n_d) = np = 50$, $\gamma = 0.05, 0.1, 0.3$, $p_D(z) = 0.6, 1.0, 1.5, 2.0$, $z = 0.5$, and $\beta = 1$, when the assumed prevalence of disease is 0.85γ and 1.15γ . First note that the Moser–Halabi and the van der Laan estimators perform almost identically for all combinations. Finally, the results in Table III are close to the results in Table II for the same combinations of $E(n_d)$, γ , $p_D(z)$. That is, the Moser–Halabi and van der Laan estimators of the relative risk of disease are relatively insensitive to misspecifications of the prevalence of disease γ .

As a final comment, from Equation (1) as the prevalence of disease γ tends to zero the relative risk of disease tends to the odds ratio. Therefore, using increasingly smaller prevalence of disease values in our estimator is ultimately equivalent to making the rare disease assumption and estimating the odds ratio, as in the logistic regression estimator. However, these two estimators of the odds ratio are still not exactly the same because the former predicts the probability of exposure given disease and the latter predicts the probability of disease given exposure. For comparison purposes therefore simulations are run where our estimator is calculated with the prevalence of disease set to zero and therefore smaller than the true prevalence of disease values (0.05, 0.10, 0.30). For all combinations tested the two estimators are virtually identical in mean value, mean relative bias, mean square error, and coverage probability. In contrast, if our estimator is calculated with the prevalence of disease 50% higher than the true prevalence then the coverage probabilities of our estimator were excessively low.

In summary, our estimator performs as well or better than the logistic regression estimator if the prevalence of disease is known, is set within 15% of the true prevalence, or is set to zero. By extension therefore, our estimator is preferred provided the prevalence of disease is set between zero and 15% above the true prevalence.

6. Examples

6.1. Scenario I example

We use the data from the Garber *et al.* [24] case-control study to illustrate our estimator. In this real-life example, there were $n = 124$ matched triplets and $n_d = 32$ discordant triplets with $n_0(+) = 21$, $n_1(+) = 2$, $n_1(-) = 8$, $n_2(-) = 1$. On the basis of previous trials, the incidence of TE among breast cancer cases that were treated with tamoxifen ranges from 2–8%. We assume the known prevalence of TE is 8%, although we believe that this is a conservative assumption. From (2), (5), (7), (8), and Appendix 3, we obtain the estimated relative risk of disease $\hat{\rho}_{D2} = 3.604$ and the estimated variance of $\log(\hat{\rho}_{D2})$ equals 0.103. Therefore, from (16), a 95% CI on the relative risk of disease is (1.924–6.753). If we assume a prevalence of disease $\gamma = 0.02$ the estimated relative risk of disease $\hat{\rho}_{D2} = 4.172$ (95% CI = 1.977–8.807), at $\gamma = 0.05$ we obtain $\hat{\rho}_{D2} = 3.869$ (95% CI = 1.953–7.664), and at $\gamma = 0.10$ we have $\hat{\rho}_{D2} = 3.445$ (95% CI = 1.901–6.242).

The Mantel–Haenszel estimator of the odds ratio $\hat{\omega}_{MH2} = 4.4$ and the estimated variance of $\log(\hat{\omega}_{MH2})$ equals 0.154. The 95% CI on the odds ratio ω is (2.037–9.504). The Miettinen

estimator of the odds ratio $\widehat{\omega}_{M2}=4.66$ and the estimated variance of $\log(\widehat{\omega}_{M2})$ equals 0.157. The 95% CI on the odds ratio ω is (2.141 – 10.139).

6.2. Scenario II example

For our estimator we fit the logistic regression model (17) where factor V Leiden (exposure) is the outcome. A stratum variable is used to identify each of the $96 = 3n_d$ discordant matched cases and controls to account for the correlation between the cases and controls. Two covariates are considered: family history of TE and smoking history. Again, we assume the prevalence of disease $\gamma = 0.08$.

From all $372 = 3n$ individual cases and controls, the observed means of the two covariates $\bar{z} = (0.145, 0.094)$ and the overall exposure rate $\bar{e} = 0.094$. From the model the estimated parameter values are: $\widehat{\alpha}_1 = 1.641$ for case status, $\widehat{\beta}_1 = -0.1647$ for family history of TE, and $\widehat{\beta}_2 = -0.8677$ for smoking. Using (20) we obtain $\widehat{\alpha}_0 = -2.708$. The estimated variance–covariance matrix is

$$\widehat{\Sigma}_{\widehat{\alpha}_1, \widehat{\beta}_1, \widehat{\beta}_2} = \begin{bmatrix} 0.17874 & -0.04289 & -0.13344 \\ -0.04289 & 0.49746 & -0.04176 \\ -0.13344 & -0.04176 & 1.23039 \end{bmatrix}. \quad (27)$$

From (19), the estimated relative risk of disease, $\widehat{\rho}_{D2}(\bar{z}) = 4.071$. From (21), the estimated variance of $\log[\widehat{\rho}_{D2}(\bar{z})]$ equals 0.104. Therefore, from (24), a 95% CI on the relative risk of disease at $\bar{z}$ is (2.161–7.669). If we assume $\gamma = 0.02$, the estimated relative risk of disease is $\widehat{\rho}_{D2}(\bar{z}) = 4.861$ (95% CI = 2.248 – 10.512), at $\gamma = 0.05$ we obtain $\widehat{\rho}_{D2}(\bar{z}) = 4.433$ (95% CI = 2.208 – 8.899), and at $\gamma = 0.10$ we have $\widehat{\rho}_{D2}(\bar{z}) = 3.859$ (95% CI = 2.126–7.003). In addition, for $\gamma = 0.08$ the van der Laan relative risk of disease estimate is 3.908 (95% CI = 2.187 – 6.982).

We also fit a logistic regression model (25) where case status is the outcome. For all $372 = 3n$ individual cases and controls, a stratum variable is used and the two covariates (family history of TE and smoking history) are specified in the model. The estimated odds ratio $\widehat{\omega}_2$ based on this conditional logistic regression is 4.834 and the estimated variance of $\log(\widehat{\omega}_2)$ equals 0.164. The 95% CI on the odds ratio ω is (2.184 – 10.700).

7. Discussion

In this paper, we develop consistent estimators of the relative risk of disease ρ_D in matched case-control studies with a set of k controls matched to each case when the prevalence of disease γ is known. These estimators are applicable when the probability of exposure is constant within each of the case and control groups, or, when the probability of exposure varies within each group. Furthermore, the estimators are formulated without making the rare disease assumption.

When the probability of exposure is constant within each of the case and control groups, our estimator $\widehat{\rho}_D$ is provided as a closed form function. Through simulation, our estimator is shown to be an improvement over the Mantel-Haenszel and Miettinen odds ratio estimators when estimating the relative risk of disease.

When the probability of exposure varies within each group, the components of our estimator $\widehat{\rho}_D(z)$ are provided either in closed form or through maximum likelihood estimations. We have demonstrated by simulation that our estimator $\widehat{\rho}_D(z)$ and van der Laan's weighted logistic regression estimator produce similar results. In addition, both of these estimators are an improvement over the logistic regression odds ratio estimator. Our proposed estimator of the relative risk of disease assumes that the prevalence of disease is known. This may be a realistic assumption when national registries are available. Furthermore, if the prevalence of disease is unknown, there are existing methodologies that will estimate prevalence of disease [25,26].

Sensitivity analyses are run in both the simulations and in our example using a range of values for the prevalence of disease. The results of the sensitivity analyses show that our estimator is robust to misspecification of the prevalence of disease.

Finally, if the prevalence of disease is unknown we recommend using a prevalence value in our estimator that is between zero and 15% above the true prevalence.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by National Institutes of Health Grants CA33601 and CA155296_1AJ. We thank the reviewers for their constructive comments. We also thank John Taylor for his editorial review.

References

1. Langholz B, Borgan O. Counter-matching: A stratified nested case-control sampling method. *Biometrika*. 1995; 82:69–79.
2. Langholz B, Borgan O. Estimation of absolute risk from nested case-control data. *Biometrics*. 1997; 53:767–774. [PubMed: 9192463]
3. Borgan O, Goldstein L, Langholz B. Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *The Annals of Statistics*. 1995; 23:1749–1778.
4. Xiang A, Langholz B. Robust variance estimation for rate ratio parameter estimates from individually matched case-control data. *Biometrika*. 2003; 90:741–746.
5. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*. 1959; 22:719–748. [PubMed: 13655060]
6. Miettinen O. Estimation of Relative Risk from Individually Matched Series. *Biometrics*. 1970; 26:75–86. [PubMed: 5461791]
7. van der Laan MJ. Estimation based on case-control designs with known prevalence probability. *International Journal of Biostatistics*. 2008; 4(Article 17)
8. Pike MC, Morrow RH. Statistical analysis of patient control studies in epidemiology. Factor under investigation an all or none variable. *British Journal of Preventative Social Medicine*. 1970; 24:42–44.
9. Ejigou A, McHugh R. Estimation of relative risk from matched pairs in epidemiologic research. *Biometrics*. 1977; 33:552–556. [PubMed: 911974]
10. Ejigou A, McHugh R. Relative risk estimation under multiple matching. *Biometrika*. 1981; 68:85–91.
11. Greenhouse SW. Jerome Cornfield's contributions to epidemiology. *Biometrics Supplement*. 1982; 38:33–45.
12. Cornfield J. A method of estimating comparative rates from clinical data: Applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*. 1951; 11:1269–1275.
13. Anderson JA. Separate sample logistic discrimination. *Biometrika*. 1972; 59:19–35.

14. Prentice RL, Pike R. Logistic disease incidence models and case-control Studies. *Biometrika*. 1979; 66:403–411.
15. Breslow NE, Powers W. Are there two logistic regressions for retrospective studies? *Biometrics*. 1978; 34:100–105. [PubMed: 630033]
16. Breslow NE. Statistics in epidemiology: the case-control study. *Journal of the American Statistical Society*. 1996; 91:14–28.
17. Breslow NE, Day NE, Halvorsen KT, Prentice RL, Sabai C. Estimation of multiple relative risk functions in matched case control studies. *American Journal of Epidemiology*. 1978; 108:299–307. [PubMed: 727199]
18. Holford TR, White C, Kelsey JL. Multivariate analysis for case-control Studies. *American Journal of Epidemiology*. 1978; 107:245–255. [PubMed: 629262]
19. Breslow, NE.; Day, NE. Statistical methods in cancer research: Volume 1, The analysis of case-control studies. International Agency for Research on Cancer; Lyon: 1980.
20. Rose S, van der Laan MJ. Simple optimal weighting of cases and controls in case-control studies. *International Journal of Biostatistics*. 2008; (Article 19)
21. Zhang J, Yu KF. What's a relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association*. 1998; 280:1690–1691. [PubMed: 9832001]
22. McNutt LA, Wu C, Xue X, et al. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology*. 2003; 157:940–943. [PubMed: 12746247]
23. Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *American Journal of Epidemiology*. 2004; 160:301–305. [PubMed: 15286014]
24. Garber JE, Halabi S, Tolane S, et al. Factor V Leiden mutation and the risk of thromboembolic events in women receiving adjuvant tamoxifen for breast cancer. Results from CALGB 9872. *Journal of the National Cancer Institute*. 2010; 102:942–949. [PubMed: 20554945]
25. Gail MH, Kessler L, Midthune D, et al. Two approaches for estimating disease prevalence from population-based registries of incidence and total mortality. *Biometrics*. 1999; 55:1137–1144. [PubMed: 11315059]
26. Coldman AJ, McBride ML, Braun T. Calculating the prevalence of cancer. *Statistics in Medicine*. 1992; 11:1579–1589. [PubMed: 1439362]

Comparison of Moser–Halabi relative risk of disease estimate $\widehat{\rho}_{Dk}$ to Mantel–Haenszel $\widehat{\omega}_{MHk}$ and Miettinen $\widehat{\omega}_{Mk}$ odds ratio estimates in their abilities to predict the relative risk of disease when $k = 2$, $p = 0.25$.

Table I

$E(n_d)$	γ	p	Moser–Halabi relative risk of disease estimate				Mantel–Haenszel odds ratio estimate				Miettinen odds ratio estimate			
			M	RB	MSE	COV	M	RB	MSE	COV	M	RB	MSE	COV
50	0.05	0.6	0.616	0.027	0.041	0.954	0.605	0.009	0.042	0.955	0.606	0.010	0.042	0.955
		1.0	1.028	0.028	0.092	0.953	1.034	0.034	0.105	0.954	1.034	0.034	0.104	0.954
		1.5	1.551	0.034	0.187	0.953	1.604	0.070	0.245	0.954	1.603	0.069	0.244	0.953
		2.0	2.075	0.038	0.334	0.952	2.206	0.103	0.504	0.950	2.203	0.102	0.498	0.951
	0.10	0.6	0.613	0.022	0.038	0.955	0.591	−0.015	0.041	0.957	0.592	−0.014	0.041	0.957
		1.0	1.023	0.023	0.080	0.952	1.035	0.035	0.105	0.954	1.035	0.035	0.104	0.954
		1.5	1.541	0.027	0.157	0.951	1.653	0.102	0.277	0.949	1.652	0.101	0.274	0.948
		2.0	2.057	0.028	0.261	0.952	2.341	0.170	0.637	0.938	2.338	0.169	0.630	0.937
	0.3	0.6	0.602	0.004	0.028	0.953	0.522	−0.130	0.040	0.941	0.523	−0.129	0.040	0.941
		1.0	1.006	0.006	0.045	0.950	1.034	0.034	0.104	0.954	1.034	0.034	0.103	0.954
		1.5	1.513	0.008	0.067	0.951	1.968	0.312	0.580	0.882	1.966	0.311	0.575	0.880
		2.0	2.020	0.010	0.093	0.959	3.439	0.719	3.348	0.640	3.435	0.717	3.319	0.638

M, mean; RB, relative bias; MSE, mean square error; COV, probability of coverage on pD .

Table II

Comparison of the Moser–Halabi estimate of the relative risk of disease $\hat{\rho}_{\text{dk}}(\bar{z})$ to the logistic odds ratio estimate $\hat{\omega}_k = \exp(\hat{\tau}_{1k})$ and to van der Laan's estimator in their ability to predict the relative risk of disease when $k = 2$, $p = 0.25$.

$E(n_d)$	γ	p	D	Moser–Halabi logistic relative risk estimate				Logistic odds ratio estimate				van der Laan estimate			
				M	RB	MSE	COV	M	RB	MSE	COV	M	RB	MSE	COV
50	0.05	0.6	1.0	0.623	0.038	0.041	0.952	0.618	0.030	0.043	0.957	0.633	0.054	0.041	0.953
				1.007	0.007	0.089	0.953	1.017	0.017	0.100	0.954	1.010	0.010	0.084	0.952
				1.491	−0.006	0.179	0.953	1.535	0.025	0.217	0.956	1.476	−0.016	0.160	0.953
				1.959	−0.020	0.309	0.946	2.056	0.027	0.406	0.952	1.919	−0.040	0.271	0.942
	0.10	0.6	1.0	0.619	0.032	0.037	0.954	0.602	0.004	0.041	0.959	0.628	0.047	0.038	0.952
				1.008	0.008	0.082	0.947	1.024	0.024	0.105	0.948	1.012	0.012	0.078	0.947
				1.485	−0.010	0.149	0.952	1.580	0.053	0.233	0.955	1.470	−0.020	0.135	0.950
				1.952	−0.024	0.254	0.946	2.177	0.088	0.493	0.948	1.911	−0.044	0.221	0.942
	0.3	0.6	1.0	0.614	0.023	0.029	0.947	0.539	−0.101	0.039	0.945	0.622	0.037	0.029	0.944
				0.994	−0.006	0.046	0.948	1.022	0.022	0.101	0.951	0.997	−0.003	0.043	0.949
				1.460	−0.027	0.068	0.955	1.846	0.231	0.432	0.915	1.447	−0.035	0.062	0.954
				1.944	−0.028	0.096	0.954	3.098	0.549	2.216	0.759	1.899	−0.050	0.084	0.942

Sensitivity of the Moser–Halabi and van der Laan estimate of the relative risk of disease $\widehat{\rho}_{\text{Dk}}^{\dagger}(\bar{z})$ when the presumed[†] prevalence of disease is 0.85γ and 1.15γ for $k=2$, $p=0.25$, and $E(n_d)=50$.

$E(n_d)$	γ	ρ_D	Moser-Halabi estimate						Van der Laan estimate									
			Designate prevalence of disease 0.85 γ			Designate prevalence of disease 1.15 γ			Designate prevalence of disease 0.85 γ			Designate prevalence of disease 1.15 γ						
			M	RB	MSE	COV	M	RB	MSE	COV	M	RB	MSE	COV	M	RB	MSE	COV
50	0.05	0.6	0.622	0.036	0.041	0.952	0.041	0.041	0.951	0.631	0.052	0.042	0.956	0.634	0.057	0.041	0.952	
		1.0	1.007	0.007	0.090	0.953	1.006	0.006	0.087	0.953	1.011	0.011	0.086	0.953	1.010	0.010	0.083	0.952
		1.5	1.498	-0.001	0.185	0.954	1.485	-0.010	0.173	0.953	1.535	0.024	0.165	0.954	1.470	-0.020	0.155	0.953
		2.0	1.975	-0.012	0.323	0.947	1.944	-0.028	0.297	0.945	1.934	-0.033	0.281	0.944	1.905	-0.048	0.261	0.940
	0.10	0.6	0.616	0.026	0.038	0.954	0.623	0.038	0.037	0.953	0.625	0.041	0.038	0.953	0.632	0.053	0.037	0.950
		1.0	1.009	0.009	0.085	0.947	1.007	0.007	0.079	0.948	1.013	0.013	0.081	0.947	1.010	0.010	0.075	0.947
		1.5	1.499	0.000	0.159	0.952	1.472	-0.019	0.140	0.951	1.484	-0.011	0.144	0.951	1.458	-0.028	0.127	0.951
		2.0	1.985	-0.008	0.377	0.948	1.921	-0.040	0.235	0.944	1.942	-0.029	0.237	0.946	1.882	-0.060	0.208	0.938
	0.30	0.6	0.600	0.000	0.030	0.951	0.628	0.047	0.028	0.939	0.609	0.015	0.029	0.952	0.637	0.061	0.028	0.932
		1.0	0.997	-0.003	0.052	0.949	0.992	-0.008	0.040	0.948	0.999	-0.001	0.049	0.948	0.994	-0.006	0.038	0.949
		1.5	1.506	0.004	0.084	0.954	1.417	-0.055	0.059	0.951	1.492	-0.005	0.075	0.951	1.406	-0.062	0.056	0.950
		2.0	2.062	0.031	0.133	0.944	1.838	-0.081	0.094	0.920	2.010	0.005	0.102	0.946	1.800	-0.100	0.094	0.900

[†] Presumed implies that while the true prevalence of disease equals γ the prevalence of disease used in calculations of the estimators is either 0.85γ or 1.15γ .