Odds Ratio Estimation in 1:n Incomplete Matched Case-Control Studies

Chan Jin and Stephen W. Looney
Dept. of Biostatistics, Georgia Regents University, Augusta, GA 30912

Abstract

A 1:n matched case-control design, in which each case is matched to n controls is commonly used to evaluate the association between exposure to a risk factor and a disease. The odds ratio (O.R.) is typically used to quantify such an association. Difficulties in estimating the true O.R. arise when the exposure status is unknown for at least one individual in a matched case-control grouping. In the case where the exposure status is known for all individuals in the group, the true O.R. can be estimated using conditional logistic regression, among other methods. In the case where the case-control data are independent, the O.R. is estimated using the cross-product ratio from the exposure-by-disease contingency table. In this presentation we suggest a simple method for estimating the O.R. when the sample consists of a combination of matched and unmatched observations, resulting from incomplete 1:n matching. This method uses a weighted average of traditional methods for estimating the O.R. with matched and unmatched data. We illustrate our method with a hypothetical example.

Key Words: matching, conditional logistic regression, cross-product ratio, simulation, bias, mean squared error

1. Introduction

As opposed to an experimental study (also known as an intervention study), an observational study is a type of research design in which the investigators observe or measure specific characteristics without attempting to intervene in the lives of the study subjects in any way. Observational studies are commonly found in clinical medicine and public health, where researchers design and conduct these studies in an attempt to unravel the etiology of and identify risk factors for human diseases. There are three basic types of observational study designs: cohort, case-control, and cross-sectional. All three designs can be used to determine whether there is an association between a factor or a characteristic and a disease. In a case-control study, two groups of individuals, those with the disease (cases), and those without (controls), are identified and information is collected on the respective exposure status of both groups. A cohort study also begins with identifying two groups of individuals, those exposed to a certain factor (exposed group), and those not exposed (non-exposed group). The investigator then follows up with both groups and compares the incidence of disease in the two groups. In a cross-sectional study, an investigator collects information on exposure and disease status simultaneously for each study subject.

To determine whether the disease is associated with the exposure of interest in an observational study, the odds ratio (OR) is often used as the measures of association. The odds of an event is defined as the ratio of the probability that the event will occur divided by the probability that the event will not occur. The OR is defined as the ratio of the odds that the cases were exposed to the odds that the controls were exposed. For rare diseases, the odds ratio closely approximates the relative risk (Gordis 2009).

Despite various advantages of case-control studies[1], one major concern when conducting a case-control study is that cases and controls may differ in characteristics or exposures other than the targeted exposure (Gordis 2009, Schlesselman 1982). To deal with this problem, cases and controls are sometimes matched on variables known to be risk factors for the disease (e.g., age, sex, race, socioeconomic status, occupation, blood type, hospital of admission, neighborhood). One or more controls are paired with each case based on the similarity of the matched variables.

Table 1. Canonical Form of 2x2 Table for an Unmatched Case-Control Study

	Cases	Controls	Total
Exposure			
Yes	a	b	m_1
No	c	d	m_2
Total	n_1	n_2	n

In an unmatched case-control study, the numbers of individuals with their respective disease and exposure status are listed in a 2 ×2 table, as seen in Table 1. The estimated odds of a case having been exposed is the estimated probability of a case having been exposed, a/n_1, divided by the probability of a case having never been exposed, c/n_1. After taking the ratio of the two probabilities, the estimated odds of a case having been exposed is a/c. Similarly, estimated odds of a case having been exposed is the estimated probability of a case having been exposed, b/n_2, divided by the probability of a case having never been exposed, d/n_2. After taking the ratio of the two probabilities, the estimated odds of a case having been exposed is b/d.The estimated odds ratio is calculated as follows, $\frac{a/c}{b/d} = ad/bc$. In a matched case-control study with one control per case, four types of case-control pairs with dichotomous exposure are possible: concordant pairs (pairs in which both the case and the control were exposed, pairs in which neither the case nor the control were exposed), and discordant pairs (pairs in which the case was exposed but the control was not, pairs in which the control was exposed but the case was not). In the canonical 2×2 table for a matched case-control study, the number in each cell no longer represents the number of individuals, but the number of matched pairs, as seen in Table 2.

Table 2. Canonical Form of 2x2Table for a Matched Case-Control Study

		Control		
		+	−	Total
Case	+	A	B	$A+B$
	−	C	D	$C+D$
Total		$A+C$	$B+D$	N

The concordant pairs (A and D, where cases and controls are either both exposed or both unexposed) do not contribute to how cases and controls differ regarding exposure history, and are ignored during the calculation of the estimated OR. Therefore, the estimated OR for matched pairs is the ratio of the discordant pairs (the ratio of the

[1] Suitable for rare diseases or those with long latency periods, relatively quick to conduct, relatively inexpensive, requires comparatively fewer subjects, existing records can sometimes be used, no risk to subjects, and allows for study of multiple risk factors for a disease.

number of pairs where the case was exposed and the control was not, to the number of pairs where the control was exposed and the case was not), i.e., B/C in Table 2.

Sometimes due to the small number of cases available (e.g., when the disease is rare), multiple controls may be matched to each case to increase the power of the study (Kim et al. 2011, Pike et al. 1997). In general, as the ratio of controls to cases increases beyond 4:1, the additional gain in statistical power for the test of the odds ratio may not be worth the extra time and effort (Miettinen 1969).

When each case has exactly one matched control (1:1 matching), the maximum likelihood estimate of the odds ratio is the ratio of discordant pairs, B/C (Cox 1958, Kraus 1960, McNemar 1947).

When each case has two matched controls (1:2 matching), Taube and Hedman (1969) derived a minimum chi-square estimate of the odds ratio and a chi-square test of significance for matched studies with multiple controls per case. Miettinen (1970) also gave an estimate based on conditional maximum likelihood, along with alternative procedures for calculating exact and approximate tests of significance and confidence intervals.

When three or more controls are matched to each case, the method proposed by Mantel and Haenszel(1959) can be used. Mantel and Haenszel proposed a stratification-based method for estimating the odds ratio, which treats each matched pair in a 1:1 design as a stratum and computes the odds ratio for each matched pair before calculating a weighted average of the individual odds ratio. The Mantel-Haenszel method can easily be extended to 1:n matching. Conditional logistic regression can also be used to estimate the odds ratio for any degree of matching. These methods are described in more detail in Section 2.

The aforementioned methods operate under the assumption of complete data; that is, exposure data are available for each case and all controls matched to it. However, during data collection, for various reasons, information on exposure is sometimes lost or unavailable, which creates difficulties for statisticians and researchers when estimating the desired odds ratio. Examples of 1:1 matched studies in which such incomplete data were present include London et al. (1991) and Pike et al. (1997). London et al. (1991) ignored the matching and treated all cases and controls as if they were independent. Pike et al. (1997) ignored the incomplete pairs. For the general 1:1 matched setting, Breslow and Day (1980, p.113) state that “common practice is to eliminate from analyses including a certain variable all individuals for whom information on that variable is missing. In a matched pair design, the individual matched to an eliminated individual will also be eliminated.” If “variable” in this quotation is taken to mean exposure, then Breslow and Day are recommending that one should remove all incomplete pairs from the analysis.

Several methods have been proposed and compared for the 1:1 incomplete matching situation, (e.g., Haber and Chen 1991, Huberman and Langholz 1999a, b, Miller and Looney 2012); however, only Li et al. (2004) have compared the performance of methods for estimating the OR in the 1:n matching situation when the exposure data for the case and the matched controls are incomplete, and they considered only 1:4 matching under very limited simulation conditions. Therefore, it is of great interest to examine the robustness of previously proposed methods under various missing exposure data scenarios. In this study, we are concerned only with whether or not exposure data were available for all cases and each control matched to them. Possible “missingness” of the matching criteria themselves (age, race, sex, etc.) is of no interest to us. We assume that all possible matching was successfully carried out.

The methods designed to handle incomplete 1:1 matched data fall into one of the following categories: (1) disregard the incomplete pairs and analyze the remaining complete pairs; (2) disregard the complete pairs and analyze the incomplete pairs as independent data; (3) ignore the matching, pretend all cases and controls are independent and perform the “unmatched” analysis; (4) use the missing indicator (MI) method (Huberman and Langholz 1999a), which makes use of all data but is computationally difficult; and (5) the Miller and Looney (ML) method (2012). The choice of method (1) or (2) is dependent on which subset is larger, complete pairs or incomplete pairs. Only the MI and ML methods make use of all available data and take into account the dependence among the completely matched pairs. An advantage of discarding incomplete pairs is its ease of calculation. When the incomplete pairs are ignored, methods for analyzing the completely matched pairs are readily available in commonly used statistical software programs and relatively easy to implement. However, for most case-control studies, the cases are rare events and ignoring the incomplete pairs is counterproductive because of the reduction in sample size.

Conditional logistic regression can be used to test for conditional independence between case status and exposure status of the matched pairs, provided all of the pairs are complete. The estimate obtained from the conditional logistic model maximizes the likelihood function and is the natural log of the odds ratio. The logistic regression model was initially intended for use with prospective studies; however, case-control studies can also be analyzed using similar models, if the likelihood function in the retrospective case-control setting is the same as that for a prospective study, differing only by a constant term for the linear predictor (Moreno et al. 1996). For a detailed discussion, see Chapter 3. An advantage of using conditional logistic regression is that statistical software is readily available to perform the required maximum likelihood estimation (PROC LOGISTIC in SAS® and the survival package in R). In a 1:1 matched study with complete data, conditional logistic regression can be viewed as a special case of unconditional logistic regression (Agresti 2007, pp.249-250), and provides results that are equivalent to the Mantel-Haenszel estimator. Another advantage of using any type of logistic regression model is that the estimated regression coefficients can be used to estimate the log odds ratio, and such estimates can be readily adjusted for the effects of confounding variables.

Maximum likelihood methods have also been used to directly estimate the odds ratio for incompletely matched case-control studies (Campbell 1984, Haber and Chen 1991, Jewell 1984). However, several of the maximum likelihood methods require use of the Estimation-Maximization algorithm and are computationally intensive when evaluating the likelihood function (Campbell, 1984; Haber and Chen, 1991).

The missing indicator method is an extension of traditional conditional logistic regression and introduces a missing indicator explanatory variable, which takes on the value 1 when the matched pair is incomplete and the value 0 when the matched pair is complete (Huberman and Langholz 1999a, Li, Song, and Gray 2004). Note that, for our purpose, a case-control pair is termed “complete” if the exposure status is known for both the case and the control; the pair is “incomplete” if the exposure status is unknown for either the case or the control. The missing indicator method was found to produce slightly greater bias and lower confidence interval coverage probability than conditional logistic regression when the exposure status of the matched cases and controls was assumed to be independent (Li, Song, and Gray, 2004). One notable disadvantage of the missing indicator method is its computational intensity (Miller and Looney 2012). Furthermore, when the missing control exposure values are case-exposure-dependent (e.g., a control matched to an exposed case is more likely to be exposed), the missing indicator method cannot appropriately handle the missing exposure values in the estimation of the odds ratio (Li, Song, and Gray, 2004).

The Miller-Looney (ML) method is calculated using a simple linear combination of the odds ratio estimate based on the complete pairs and the odds ratio estimate based on the incomplete pairs. It was shown to have superior performance when compared with several previously proposed methods for dealing with incomplete 1:1 matched data (Miller and Looney 2012). It is described in more detail in Section 2.6. In this dissertation, the ML method will be extended to the 1:n matching situation and its performance will be evaluated under several incomplete data scenarios.

The purpose of this study is to compare and contrast several of the above-mentioned methods in 1:n matched case-control studies with incomplete matched data using a large-scale simulation study. These results will enable investigators to choose the most appropriate statistical method(s) when analyzing data from such case-control studies.

2. Methods

2.1 Inference for the Odds Ratio Based on Unmatched Data

Let ψ denote the true odds ratio. Let a,b,c,d denote the cell entries in Table 1. The sampling distribution of the usual estimate of the odds ratio is highly skewed, unless the sample size is extremely large. For example, when $\psi = 1$, $\hat{\psi}$ cannot be much smaller than ψ (since $\hat{\psi} \ge 0$), but it could be much larger with substantial probability. To deal with the skewness, statistical inference for the odds ratio often uses its natural logarithm, $ln\,\hat{\psi}$. The sample log odds ratio, $ln\,\hat{\psi}$, has a less skewed sampling distribution that is approximately normal. Its approximating normal distribution has a mean of $ln\ \psi$ and an approximate standard error of $SE = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$ (Woolf 1955). It is obvious that as the cell counts increase, the SE decreases.

Because the sampling distribution for $ln\,\hat{\psi}$ is closer to normality than that of $\hat{\psi}$, it is preferable to construct a confidence interval for $ln\ \psi$ and then back-transform to obtain confidence limits for ψ. A large-sample confidence interval for $ln\ \psi$ is given by $ln\,\hat{\psi} \pm z_{\alpha/2}(SE)$, where $z_{\alpha/2}$ denotes the upper α-percentage point of the standard normal. The confidence interval for ψ is consequently $\left(e^{ln\hat{\psi}-z_{\alpha/2}(SE)}, e^{ln\hat{\psi}+z_{\alpha/2}(SE)}\right)$.

If either b or c is equal to zero, the sample odds ratio ad/bc is undefined. If either a or d is zero, the sample odds ratio is zero. In any of these situations, we will use the "slightly amended" estimator $\hat{\psi} = \frac{(a+0.5)(d+0.5)}{(b+0.5)(c+0.5)}$, in which ½ is added to each cell count (Agresti 2007, pp.31-32). Inference for ψ is then performed in the usual way using cell entries of $a+\frac{1}{2}, b+\frac{1}{2}, c+\frac{1}{2}, d+\frac{1}{2}$. We will use the amended estimator any time any of the cell counts are zero.

2.2 Mantel-Haenszel Method

Mantel and Haenszel (1959) described a stratification-based method which estimates a summary odds ratio from a series of 2x2 tables. Suppose that cases and controls are stratified based on one or more variables into k subgroups or strata. Let the observation in the i'th stratum be written as in Table 3.

Table 3. Exposure Status Among Cases and Controls Using Mantel-Haenszel Method in the ith Stratum

	Cases	Controls	Total
Exposure			
Yes(+)	a_i	b_i	m_{1i}
No(−)	c_i	d_i	m_{2i}
Total	n_{1i}	n_{2i}	n_i

The Mantel-Haenszel (M-H) estimator of the odds ratio, adjusted for the effect of the stratification variable, is calculated as $\hat{\psi}_{mh} = \dfrac{\sum_{i=1}^{k} a_i d_i / n_i}{\sum_{i=1}^{k} b_i c_i / n_i}$. An approximate test of the hypothesis of no association ($H_0 : \psi = 1$) can be carried out as follows. For the *i*'th subgroup, the mean and variance of a_i under H_0 is given by $E(a_i) = \dfrac{n_{1i} m_{1i}}{n_i}$ and $V(a_i) = \dfrac{n_{1i} n_{2i} m_{1i} m_{2i}}{n_i^2 (n_i - 1)}$.

The M-H test of $H_0 : \psi = 1$ against the two-sided alternative $H_1 : \psi \neq 1$ is carried out using the test statistic $\chi^2_{mh} = \dfrac{\left[| \sum a_i - \sum E(a_i) | - \frac{1}{2} \right]^2}{\sum V(a_i)}$. The $\frac{1}{2}$ correction for continuity is introduced so that the *p*-value based on χ^2_{mh} more closely approximates the value based on the exact conditional test (Li, Simon, and Gart 1979). The statistic χ^2_{mh} has an approximate chi-square distribution with one degree of freedom under H_0. For a one-sided test, the approximate unit normal deviate $z = \pm \sqrt{\chi^2_{mh}}$ may be used.

Miettinen (1974, 1976) proposed a test-based method for obtaining approximate confidence limits for the true odds ratio using the M-H method. An approximate 100(1-α)% confidence interval for ψ is given by $\exp\left[\left(1 \pm \dfrac{z_{\alpha/2}}{\sqrt{\hat{\psi}^2_{mh}}} \right) \ln \hat{\psi}_{mh} \right]$.

The M-H estimate $\hat{\psi}_{mh}$ can be determined as a weighted average of the stratum-specific odds ratios, assuming that none of the b_i or c_i equals zero. In the *i*'th stratum, the odds ratio is estimated as $\hat{\psi}_i = ad / bc$. Using the weights given by $w_i = \dfrac{b_i c_i}{n_i}$, $\hat{\psi}_{mh}$ can also be written as $\hat{\psi}_{mh} = \dfrac{\sum w_i \hat{\psi}_i}{\sum w_i}$. Since within each subgroup, the ranges of the stratification variables are restricted by design, cases and controls do not differ by very much on the stratification variables. Therefore, $\hat{\psi}_i$ is relatively free from potential confounding bias. As a weighted average, $\hat{\psi}_{mh}$ is also relatively free of confounding bias.

2.3 1:n Matched M-H Method

Let $n_j(+)$ denote the number of matched sets where the case is exposed and exactly j of the controls are exposed. Let $n_j(-)$ denote the number of matched sets where the case is not exposed and exactly j of the controls are exposed. By extending the method used for finding the M-H estimate with two matched controls per case, the estimate with c matched controls per case is given by $\hat{\psi}^2_{mh} = \dfrac{\sum_{j=0}^{c}(c-j)n_j(+)}{\sum_{j=0}^{c} jn_j(-)}$ (Miettinen 1970). Let $m_j = n_{j-1}(+) + n_j(-)$ denote the number of completely matched sets with exactly j exposed persons, where $n_{-1}(+) \equiv 0$. Then the test statistic based on $\hat{\psi}^2_{mh}$ can be expressed as follows (Pike and Morrow 1970): $\chi^2_{mh} = \dfrac{\left(|T_1 - T_2| - \frac{1}{2}\right)^2}{T_3}$, where $T_1 = \sum a_i = \sum_{j=0}^{c-1} n_j(+)$, $T_2 = \sum E(a_i) = \sum_{j=0}^{c} \dfrac{jm_j}{c+1}$, and $T_3 = \sum V(a_i) = \sum_{j=0}^{c} \dfrac{j(c+1-j)m_j}{(c+1)^2}$. The test statistic χ^2_{mh} is approximately distributed as chi-square with one degree of freedom under $H_0 : \psi = 1$. An approximate 100(1-α)% confidence interval for ψ is given by $\exp\left[\left(1 \pm \dfrac{z_{\alpha/2}}{\sqrt{\hat{\psi}^2_{mh}}}\right)\ln\hat{\psi}_{mh}\right]$.

2.4 Conditional Logistic Regression

In the 1:1 matched setting, suppose there are q matched pairs, $h = 1, 2, ..., q$, and θ_{hi} is the probability of the i'th subject in the h'th pair with the event (i = 1, 2). Suppose that $\mathbf{z_{hi}}$ represents the set of explanatory variables for the i'th subject in the h'th matched pair. In traditional conditional logistic regression, multiple explanatory variables may be present (e.g., the exposure variable plus any confounders). However, for the purpose of this dissertation, we only consider one explanatory variable, i.e., exposure status.

The likelihood for the vector of explanatory variables having values $\mathbf{z_{h1}}$ given that subject *h1* is the case (e) and having values $\mathbf{z_{h2}}$ given that subject *h2* is the control $(\bar{e})$ is $\Pr(\mathbf{z_{h1}} | e)\Pr(\mathbf{z_{h2}} | \bar{e})$.

The sum of this likelihood $\Pr(\mathbf{z_{h1}} | e)\Pr(\mathbf{z_{h2}} | \bar{e})$ and that for its counterpart, the likelihood for the vector of explanatory variables having values $\mathbf{z_{h1}}$ given the control and being $\mathbf{z_{h2}}$ given the case, is given by $\Pr(\mathbf{z_{h1}} | e)\Pr(\mathbf{z_{h2}} | \bar{e}) + \Pr(\mathbf{z_{h1}} | \bar{e})\Pr(\mathbf{z_{h2}} | e)$ and therefore the conditional likelihood for a particular matched pair having the observed pairing of explanatory variables $\mathbf{z_{h1}}$ with the case e and the explanatory variables $\mathbf{z_{h2}}$ with the control $\bar{e}$ is $\dfrac{\Pr(\mathbf{z_{h1}} | e)\Pr(\mathbf{z_{h2}} | \bar{e})}{\Pr(\mathbf{z_{h1}} | e)\Pr(\mathbf{z_{h2}} | \bar{e}) + \Pr(\mathbf{z_{h1}} | \bar{e})\Pr(\mathbf{z_{h2}} | e)}..$

Applying Bayes' Theorem to each of the six terms in the above expression, the conditional likelihood becomes $\dfrac{\Pr(e | \mathbf{z_{h1}})\Pr(\bar{e} | \mathbf{z_{h2}})}{\Pr(e | \mathbf{z_{h1}})\Pr(\bar{e} | \mathbf{z_{h2}}) + \Pr(\bar{e} | \mathbf{z_{h1}})\Pr(e | \mathbf{z_{h2}})}..$

If we assume a logistic model for θ_{hi}, the probability of the exposure status of the *i*'th subject in the *h*'th matched pair being either *e* or $\overline{e}$, then appropriate substitutions can be made into the conditional likelihood:
$\theta_{\mathbf{h}i} = \dfrac{\exp(\boldsymbol{\gamma}'\mathbf{z_{hi}})}{1+\exp(\boldsymbol{\gamma}'\mathbf{z_{hi}})}$, where the $z_{\mathbf{h}ik}$ are values of the $k = 1,2,...,t$ explanatory variables for the *i*'th subject in the *h*'th matched pair, and the γ_k are the corresponding coefficients of the $\mathbf{z}_{\mathrm{k}}$.

Substituting θ_{hi} for $\Pr(e\,|\,\mathbf{z_{h1}})$ and $(1-\theta_{hi})$ for $\Pr(\overline{e}\,|\,\mathbf{z_{h1}})$ gives $\dfrac{\exp(\boldsymbol{\gamma}'\mathbf{z}_{\mathbf{h}1})}{\exp(\boldsymbol{\gamma}'\mathbf{z}_{\mathbf{h}1})+\exp(\boldsymbol{\gamma}'\mathbf{z}_{\mathbf{h}2})}$, , which is the same as $\dfrac{\exp\left[\boldsymbol{\gamma}'(\mathbf{z}_{\mathbf{h}1}-\mathbf{z}_{\mathbf{h}2})\right]}{1+\exp\left[\boldsymbol{\gamma}'(\mathbf{z}_{\mathbf{h}1}-\mathbf{z}_{\mathbf{h}2})\right]}$. The conditional likelihood for the entire data is then $\prod_{h=1}^{q}\dfrac{\exp\left[\boldsymbol{\gamma}'(\mathbf{z}_{\mathbf{h}1}-\mathbf{z}_{\mathbf{h}2})\right]}{1+\exp\left[\boldsymbol{\gamma}'(\mathbf{z}_{\mathbf{h}1}-\mathbf{z}_{\mathbf{h}2})\right]}$. For this conditional likelihood, matched pairs with $\mathbf{z}_{\mathbf{h1}k} = \mathbf{z}_{\mathbf{h2}k}$ for all *k* are uninformative, and so the concordant matched pairs can be excluded from the analysis.

To extend this likelihood to the 1:n matched setting, the conditional likelihood is $\prod_{h=1}^{q}\left[1+\sum_{i=1}^{n}\exp\left[\boldsymbol{\gamma}'(\mathbf{z}_{\mathbf{h}i}-\mathbf{z}_{\mathbf{h}0})\right]\right]^{-1}$, where $i = 1,2,...,m$ indexes the controls and i = 0 corresponds to the case. The maximization of this likelihood can be performed with the “clogit” function the “survival” package of R (R Core Team 2015, Therneau and Grambsch 2000, Therneau 2015).

2.5 Fleiss’ Method

Let *r* denote the number of controls matched to a particular case, which may vary from as low as 1 (matched pairs) to as high as *n* (1:*n* matching, i.e., 1 case matched with *n* controls). The analysis first stratifies all cases according to the value of *r*, as follows:

Table 4. Frequency of Exposure Using Fleiss’ Notation

Status of case	Number of Controls exposed			
	0	1	…	r
Exposed	$Z_{10}^{(r)}$	$Z_{11}^{(r)}$	…	$Z_{1r}^{(r)}$
Unexposed	$Z_{00}^{(r)}$	$Z_{01}^{(r)}$	…	$Z_{0r}^{(r)}$

Thus, $Z_{1j}^{(r)}$ represents the number of matched sets with *r* controls where both the case and exactly *j* of the controls were exposed. Similarly, $Z_{0j}^{(r)}$ represents the number of matched sets with *r* controls where the case was unexposed and exactly *j* of the controls were exposed.

Define $A^{(r)} = \dfrac{1}{r+1}\sum_{j=0}^{r}(r-j)Z_{1j}^{(r)}$ and $B^{(r)} = \dfrac{1}{r+1}\sum_{j=0}^{r} jZ_{0j}^{(r)}$. Fleiss' estimator of the odds ratio is given by $\hat{\psi}_F = \dfrac{\sum_{r=1}^{n} A^{(r)}}{\sum_{r=1}^{n} B^{(r)}}$. The large sample variance estimate is $\widehat{V\left(\ln\hat{\psi}_F\right)} = \dfrac{\sum_{r=1}^{n} C^{(r)}}{\left[\sum_{r=1}^{n} A^{(r)}\right]^2}$, where $C^{(r)} = \dfrac{1}{\left(r+1\right)^2}\left[\sum_{j=0}^{r}(r-j)^2 Z_{1j}^{(r)} + \hat{\psi}_{mh}^2 \sum_{j=0}^{r} j^2 Z_{0j}^{(r)}\right]$.

2.6 Missing Indicator Method

The Missing Indicator method introduces an indicator variable for all matched data that is set to 1 if there is a missing exposure value and 0 otherwise. Missing exposure values are then replaced with 0 in the data. Both exposure and the missing indicator variable are entered into a conditional logistic regression model.

2.7 Proposed Method

Let β = *ln* ψ. Then $\hat{\beta}_p = \sum_{i=0}^{n} w_i \hat{\beta}_i$, where w_0 = weight assigned to the estimator based on any unmatched case and controls, $\hat{\beta}_0$ = log of the "usual" *ad*/*bc* odds ratio estimator based on any unmatched case and controls, w_i = weight attached to the matched estimator based on all complete 1:*i* matches, and $\hat{\beta}_i$ = log odds ratio estimator based on all complete 1:*i* matches using conditional logistic regression, $i = 1,\ldots,n$.

Each weight w_i is given by the reciprocal of the estimated variance of that estimator divided by the sum of the reciprocals of the estimated variances of all estimators. If $\hat{\theta}_1,\ldots,\hat{\theta}_n$ are independent, unbiased estimators of an unknown parameter θ, then the minimum variance unbiased linear combination of $\hat{\theta}_1,\ldots,\hat{\theta}_n$ is given by $\sum_{i=0}^{n} w_i \hat{\theta}_i$, , where $w_i = \dfrac{V(\hat{\theta}_i)^{-1}}{\sum_{i=1}^{n} V(\hat{\theta}_i)^{-1}}$, $i = 1,\ldots,n$ (Hodges and Lehmann 1970, pp. 288, 306-308). The variance of the proposed estimator is given by $V\left(\hat{\beta}_p\right) = \left[\sum_{i=0}^{n} V\left(\hat{\beta}_i\right)^{-1}\right]^{-1}$. .

For any of the estimators of β = *ln* ψ, an approximate 100(1-α)% confidence interval (CI) for β is found using $\hat{\beta} \pm z_{\alpha/2} SE(\hat{\beta})$ and then the endpoints are exponentiated to find a 100(1-α)% CI for ψ.

3. Motivating Example

To illustrate the different methods of estimating the odds ratio with incompletely matched 1:n data, we adapt an example from Breslow and Day (1980, p. 178) that is based on a study by Mack et al. (1976) that examined the association between estrogen (exposure, defined as any nonzero dose) and endometrial cancer. Data for a total of 59 5-

tuples were collected, where each 5-tuple consisted of four controls matched to a case of endometrial cancer. In our hypothetical example, data for 26 sets of 5-tuples are complete, i.e. the exposure status for the case and the four matched controls are known. Data for 33 5-tuples are incomplete. Four of these 5-tuples have missing exposure data for one of the four controls. For the remaining 29 5-tuples, the exposure status is known for either the case or one of the controls, but not for any of the remaining 4 subjects in the 5-tuple.

The data can be summarized in the following tables:

Table 5. No Missing Exposure Status
(Complete 5-tuples)

Status of case	Number of controls exposed					Total
	0	1	2	3	4	
Exposed	0	0	11	9	0	20
Unexposed	1	0	3	1	1	6
Total	1	0	14	10	1	26

So, for example, there were 11 5-tuples in which the case was exposed and 2 of the 4 matched controls were exposed. The remaining two controls were not exposed.

Table 6. The Exposure Status is Missing for One of the Controls

Status of case	Number of controls exposed				Total
	0	1	2	3	
Exposed	1	3	0	0	4
Unexposed	0	0	0	0	0
Total	1	3	0	0	4

So, for example, there are three 4-tuples in which the case was exposed and one of the three matched controls was exposed. The other two matched controls were not exposed.

Table 7. Only the Exposure Status of the Case or One of the Controls is Known

Status of case	Number of controls exposed		
	0	1	Unknown
Exposed	-	-	2
Unexposed	-	-	6
Unknown	4	17	-

In other words, 2 of the cases were exposed and 6 were unexposed. Of the 21 controls, 17 were exposed and 4 were not.

3.1 Matching Ignored

The matching ignored estimator ignores the fact that matching was used and assumes that cases and controls are all independent of each other. Woolf's method is used to find an approximate confidence interval (CI) for ψ. To apply this method, all 59 5-tuples are broken down into individual cases and controls and reconstructed as the following canonical 2x2 table:

Table 8. Canonical 2x2 Table for Motivating Example

	Cases	Controls	Total
Exposure			
Yes (+)	26	82	108
No (−)	12	55	67
Total	38	137	175

The matching ignored estimator $\hat{\psi}_n$ is calculated as (26·55)/(12·82) = 1.453. Then, $\hat{\beta}_N = \ln \hat{\psi}_N = \ln(1.453) = 0.374$. The approximate standard error of $\hat{\beta}_N$ is given by $\sqrt{\frac{1}{26}+\frac{1}{55}+\frac{1}{12}+\frac{1}{82}} = 0.390$. The 95% confidence interval for $\hat{\beta}_N$ is then 0.374 ± 1.96(0.390) = (-0.391, 1.138). Therefore, $\hat{\psi}_N = 1.453,$, with 95% confidence interval $(e^{-0.391}, e^{1.138}) = (0.677, 3.122)$. Note that the same results would have been obtained had logistic regression or Mantel-Haenszel been used, treating the cases and controls as independent.

There are two advantages of using the matching ignored estimator: it is easy to calculate and it makes use of all available data. However, this estimator ignores the fact that matching was used, which typically increases the bias of the estimated OR.

3.2 Matched-Data Only Methods

Conditional Logistic Regression (CLR)

The CLR estimate is calculated using all cases with at least one matched control. In the example, the 29 unmatched subjects were ignored, and the 5-tuples that had complete exposure information for all 4 controls and the 5-tuples that had missing exposure data for one of the 4 matched controls were included when estimating the odds ratio. The results of this analysis are given in Table 10.

Mantel-Haenszel (M-H)

The M-H odds ratio estimate is calculated using the same data as the CLR estimate. The results are given in Table 10.

Fleiss

The Fleiss odds ratio estimate is calculated using the same data as the M-H and CLR estimates. The results are given in Table 10.

3.3 Unmatched Data Only

All matched *n*-tuples are ignored and the logistic regression method is applied to the *n*-tuples where the exposure status is known for either the case or one the controls, but not both. For the motivating example, we used logistic regression to estimate the odds ratio for the 29 unmatched subjects. These subjects were broken down into individual cases and controls and reconstructed as the following canonical 2x2 table:

Table 9. Canonical 2x2 Table for Unmatched Cases in Motivating Example

	Cases	Controls	Total
Exposure			
Yes (+)	2	17	19
No (−)	6	4	10
Total	8	21	29

The unmatched estimator $\hat{\psi}_U$ is calculated as (2·4)/(6·17) = 0.078. Then, $\hat{\beta}_U = \ln\hat{\psi}_U = \ln(0.078) = -2.546$. The approximate standard error of $\hat{\beta}_U$ is given by $\sqrt{\frac{1}{2}+\frac{1}{17}+\frac{1}{6}+\frac{1}{4}} = 0.988$. The 95% confidence interval for $\hat{\beta}_U$ is then -2.546 ± 1.96(0.988) = (-4.482, -0.609). Therefore, $\hat{\psi}_U = 0.078,$, with 95% confidence interval $(e^{-4.482}, e^{-0.609}) = (0.011, 0.544)$.

3.4 Matched and Unmatched Combined Estimator

The matched and unmatched combined estimator makes use of all data, and is a weighted average of any one of the matched estimators and the "unmatched only" estimator. In the example, this estimator uses all 175 observations. The combined estimator of the log odds ratio is given by $\hat{\beta}_a = w_m\hat{\beta}_m + w_u\hat{\beta}_u$ where $\hat{\beta}_m$ is the matched data only estimator obtained using either CLR, M-H, or Fleiss method, and $\hat{\beta}_U$ is the unmatched subjects only estimator.

3.5 M-tuples Only Estimators

The M-tuples only estimates are intermediate steps needed for the proposed estimator (Section 2.7). Either the CLR, M-H, or Fleiss method is used to obtain the matched estimate separately for all complete pairs, all complete 3-tuples, all complete 4-tuples, etc. In the example, the odds ratios based on the complete 1:3 matches (four 4-tuples, 16 observations), and complete 1:4 matches (26 5-tuples, 130 observations) were calculated. Because there were no unexposed cases in the 4-tuples, 0.5 was added to each cell when estimating and performing inference for the odds ratio, as proposed by Agresti (2007, pp. 31-32).

3.6 Missing Indicator Estimator

The Missing Indicator (MI) estimate uses all available data. The results of this analysis are given in Table 10.

3.7 Proposed Estimator

The proposed estimator is a weighted average of all available matched-data-only estimators and the "unmatched only" estimator. Each weight is given by the reciprocal of the estimated variance of that estimator divided by the sum of the reciprocals of the estimated variances of all estimators (see Section 2.7). In the example, we used CLR to estimate the odds ratio for the complete 5-tuples. We used Agresti's method as described in Section 2.1 for the 4-tuples, and logistic regression for the unmatched data. This estimator makes use of all the data, unlike any of the other estimators, with the exception of the unmatched estimator and the missing indicator estimator.

3.8 Summary of Results

Table 10. Comparisons of Estimation Results for Motivating Example

Method	**Sample Size**	$\ln \widehat{OR}$	$SE(\ln \widehat{OR})$	$\widehat{OR}$	**95% CI**
Matching Ignored	175	0.374	0.390	1.453	0.676-3.121
Matched Only (CLR)	146	1.150	0.509	3.158	1.165-8.564
Matched Only (CMH)	146	1.179	0.521	3.250	1.171-9.023
Matched Only (Fleiss)	146	1.179	0.517	3.250	1.180-8.948
Unmatched Subjects Only (LR)	29	-2.546	0.988	0.078	0.011-0.543
Matched(CLR)+Unmatched (“Combined” Estimator)	175	0.374	0.453	1.454	0.599-3.530
5-tuples Only (CLR)	130	0.796	0.515	2.216	0.807-6.084
5-tuples Only (CMH)	130	0.869	0.546	2.385	0.818-6.956
4-tuples Only (Agresti)	16	1.898	0.997	6.670	0.945-47.100
Missing Indicator method	175	0.352	0.385	1.422	0.669-3.023
4-tuples(Agresti)+5-tuples (CLR)+Unmatched (Proposed Estimator)	175	0.396	0.415	1.486	0.658-3.354

Four estimation methods used all 175 observations: the "matching ignored" estimator, the “combined” estimator, the Missing Indicator method, and our proposed method. All other methods are either intermediate results (5-tuples only, 4-tuples only), or do not include the entire data set (matched only, unmatched subjects only). Out of the four estimators that used all the data, the missing indicator estimator yielded the smallest standard error and the narrowest confidence interval, followed closely by the proposed method. Notice the confidence interval of the "matching ignored" estimator lies completely inside the CI of the proposed estimator. However, using the "matching ignored" estimator is not recommended because the dependence among the matched cases and controls is not taken into account in the analysis. We hope to show in a simulation study that the proposed estimator is the preferred estimator under most conditions.

References

1. Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (2nd ed.), Hoboken, NJ: Wiley.
2. Bradley, J. V. (1978). “Robustness?”, *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
3. Breslow, N. E., and Day, N. E. (1980). *Statistical Methods in Cancer Research: Vol 1. The Analysis of Case-Control Data*. Lyon: IARC.
4. Broders, A. C. (1920). “Squamous-Cell Epithelioma of the Lip, A Study of Five Hundred and Thirty-Seven Cases”, *Journal of the American Medical Association*, 74, 656-664. doi: 10.1001/jama.1920.02620100016007.
5. Campbell, G. (1984). “Testing Equality of Proportions with Incomplete Correlated Data”, *Journal of Statistical Planning and Inference*, 10, 311-321. doi: 10.1016/0378-3758(84)90056-9.
6. Cole, P. (1979). “The Evolving Case-Control Study (with Comment by ED Acheson and Discussion)”, *Journal of Chronic Diseases*, 32, 15-34.

7. Cornfield, J. (1951). “A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix”, *Journal of the National Cancer Institute*, 11, 1269-1275.

8. Cox, D. R. (1958). “Two Further Applications of a Model for Binary Regression”, *Biometrika*, 45, 562-565.

9. Cox, D. R. (1970). *Analysis of Binary Data*. London: Methuen.

10. Farewell, V. T. (1979). “Some Results on the Estimation of Logistic Models Based on Retrospective Data”, *Biometrika*, 66, 27-32. doi: 10.1093/Biomet/66.1.27.

11. Gordis, L. (2009). *Epidemiology* (4th ed.), Philadelphia: Elsevier.

12. Guy, W. A. (1843). “Contributions to a Knowledge of the Influence of Employments upon Health”, *Journal of the Statistical Society of London*, 6, 197-211. doi: 10.2307/2337789.

13. Haber, M., and Chen, C. C. H. (1991). “Estimation of Odds Ratios from Matched Case-Control Studies with Incomplete Data”, *Biometrical Journal*, 33, 673-682. doi: 10.1002/bimj.4710330606.

14. Hodges, J. L., and Lehmann, E. L. (1970). *Basic Concepts of Probability and Statistics* (2nd ed.), Oakland, CA: Holden-Day.

15. Huberman, M., and Langholz, B. (1999a). “Application of the Missing-Indicator Method in Matched Case-Control Studies with Incomplete Data”, *American Journal of Epidemiology*, 150, 1340-1345.

16. Huberman, M., and Langholz, B. (1999b). “Re: ‘Combined Analysis of Matched and Unmatched Case-Control Studies: Comparison of Risk Estimates from Different Studies’”, *American Journal of Epidemiology*, 150, 219-210.

17. Jewell, N. P. (1984). “Small-Sample Bias of Point Estimators of the Odds Ratio from Matched Sets”, *Biometrics*, 40, 421-435. doi: 10.2307/2531395.

18. Johnson, N. L., and Kotz, S. (1969). *Distribution in Statistics*. Boston: Houghton Mifflin.

19. Kim, F. M., Hayes, C., Williams, P. L., Whitford, G. M., Joshipura, K. J., Hoover, R. N., Douglass, C. W., and the National Osteosarcoma Etiology Group (2011). “An Assessment of Bone Fluoride and Osteosarcoma”, *Journal of Dental Research*, 90, 1171-1176. doi: 10.1177/0022034511418828.

20. Kraus, A. S. (1960). “Comparison of a Group with a Disease and a Control Group from the Same Families, in the Search for Possible Etiologic Factors”, *American Journal of Public Health and the Nation’s Health*, 50, 303-311. doi: 10.2105/AJPH.50.3_Pt_1.303.

21. Lane-Claypon, J. E. (1926). *A Further Report on Cancer of the Breast*. London: HMSO.

22. Last, J. M. (1995). *A Dictionary of Epidemiology* (3rd ed.), New York: Oxford University Press.

23. Li, S. H., Simon, R. M., and Gart, J. J. (1979). “Small Sample Properties of the Mantel-Haenszel Test”, *Biometrika*, 66, 181-183.

24. Li, X. B., Song, X. Y., and Gray, R. H. (2004). “Comparison of the Missing-Indicator Method and Conditional Logistic Regression in 1:m Matched Case-Control Studies with Missing Exposure Values”, *American Journal of Epidemiology*, 159, 603-610. doi: 10.1093/aje/kwh075.

25. Liang, K. Y., and Zeger, S. L. (1988). “On the Use of Concordant Pairs in Matched Case Control Studies”, *Biometrics*, 44, 1145-1156. doi: 10.2307/2531742.

26. London, S. J., Thomas, D. C., Bowman, J. D., Sobel, E., Cheng, T. C., and Peters, J. M. (1991). “Exposure to Residential Electric and Magnetic Fields and Risk of Childhood Leukemia”, *American Journal of Epidemiology*, 134, 923-937.

27. Louis, P. C. A. (1844). *Research on Phthisis: Anatomical, Pathological and Therapeutical*. London: Sydenham Society.

28. Mack, T. M., Pike, M. C., Henderson, B. E., Pfeffer, R. I., Gerkins, V. R., Arthur, M., and Brown, S. E. (1976). “Estrogens and Endometrial Cancer in a Retirement Community”, *New England Journal of Medicine*, 294, 1262-1267.

29. Mantel, N., and Haenszel, W. (1959). “Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease”, *Journal of the National Cancer Institute*, 22, 719-748.

30. McNemar, Q. (1947). “Note on the Sampling Error of the Difference between Correlated Proportions or Percentages”, *Psychometrika*, 12, 153-157. doi: 10.1007/bf02295996.

31. Miettinen, O. S. (1969). “Individual Matching with Multiple Controls in the Case of All-or-None Responses”, *Biometrics*, 25, 339-355. doi: 10.2307/2528794.

32. Miettinen, O. S. (1970). “Estimation of Relative Risk from Individually Matched Series”, *Biometrics*, 26, 75-86. doi: 10.2307/2529046.

33. Miettinen, O. S. (1974). “Simple Interval Estimation of Risk Ratio”, *American Journal of Epidemiology*, 100, 515-516.

34. Miettinen, O. S. (1976). “Estimability and Estimation in Case-Referent Studies”, *American Journal of Epidemiology*, 103, 226-235.

35. Miller, K. M., and Looney, S. W. (2012). “A Simple Method for Estimating the Odds Ratio in Matched Case-Control Studies with Incomplete Paired Data”, *Statistics in Medicine*, 31, 3299-3312. doi: 10.1002/sim.5355.

36. Moreno, V., Martin, M. L., Bosch, F. X., deSanjose, S., Torres, F., and Munoz, N. (1996). “Combined Analysis of Matched and Unmatched Case-Control Studies: Comparison of Risk Estimates from Different Studies”, *American Journal of Epidemiology*, 143, 293-300.

37. Pike, M. C., and Morrow, R. H. (1970). “Statistical Analysis of Patient-Control Studies in Epidemiology. Factor Under Investigation: An All-or-None Variable”, *British Journal of Preventive and Social Medicine*, 24, 42-44.

38. Pike, M. C., Peters, R. K., Cozens, W., Probst-Hensch, N. M., Felix, J. C., Wan, P. C., and Mack, T. M. (1997). “Estrogen-Progestin Replacement Therapy and Endometrial Cancer”, *Journal of National Cancer Institute*, 89, 1110-1116. doi: 10.1093/jnci/89.15.1110.

39. Prentice, R. L., and Pyke, R. (1979). “Logistic Disease Incidence Models and Case-Control Studies”, *Biometrika*, 66, 403-411. doi: 10.1093/Biomet/66.3.403.

40. R: A Language and Environment for Statistical Computing, Vienna, Austria.

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

42. Taube, A., and Hedman, B. (1969). “On the Consequences of Matching in Retrospective Studies with Special Regard to the Calculation of Relative Risks”, *Acta Societatis Medicorum Upsaliensis*, 74, 1-16.

43. Therneau, T. M., and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*, New York: Springer.

44. Wacholder, S., McLaughlin, J. K., Silverman, D. T., and Mandel, J. S. (1992). “Selection of Controls in Case-Control Studies. I. Principles”, American Journal of Epidemiology, 135, 1019-1028.

45. Wacholder, S., Silverman, D. T., McLaughlin, J. K., and Mandel, J. S. (1992). “Selection of Controls in Case-Control Studies. II. Types of Controls”, *American Journal of Epidemiology*, 135, 1029-1041.

46. Woolf, B., (1955). “On Estimating the Relation Between Blood Group and Disease”, *Annals of Human Genetics*, 19, 251-253.