Estimating Crude or Common Odds Ratios in Case-Control Studies with Informatively Missing Exposure Data

Robert H. Lyles and Andrew S. Allen

In case-control studies, the crude odds ratio derived from a $2 \times 2$ table and the common odds ratio adjusted for stratification variables are staple measures of exposure-disease association. While missing exposure data are encountered in the majority of such studies, formal attempts to deal with them are rare, and a complete-case analysis is the norm. Furthermore, the probability that exposure is missing may depend on true exposure status, so the missing-at-random assumption is often unreasonable. In this paper, the authors present an adjustment to the usual product binomial likelihood to properly account for missing data. Estimation of model parameters without restrictive assumptions requires an additional data collection effort akin to a validation study. Closed-form results are provided to facilitate point and confidence interval estimation of crude and common odds ratios after properly accounting for informatively missing data. Simulations assess performance of the likelihood-based estimates and inferences, and they display the potential for bias in complete-case analyses. An example is presented to illustrate the approach.


Received for publication January 25, 2001, and accepted for publication July 19, 2001.

Abbreviations: MAR, missing at random; MLE, maximum likelihood estimator; OR, odds ratio; SE, standard error.

From the Department of Biostatistics, The Rollins School of Public Health, Emory University, Atlanta, GA.

Correspondence to Dr. Robert H. Lyles, Department of Biostatistics, The Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322 (e-mail: rlyles@sph.emory.edu).

In case-control studies, epidemiologists generally encounter missing data because they are unable to obtain information on some subjects by using the chosen data collection instrument. However, most serious discussions of methods for dealing with missing data remain confined to the statistics literature. Exceptions include a review of methods for handling missing covariate information in regression analysis (1) and a study comparing approaches for estimating a common odds ratio when some values of a confounding variable are missing (2). Nevertheless, in crude and adjusted epidemiologic analyses, it is common practice to ignore missing data and to report results based solely on the observed (complete-case) data.

In many statistical treatments of missing data, an assumption is made that data are missing at random (MAR) (3). The implication is that the probability of a missing value does not depend on any missing data, although it may arbitrarily depend on observed data. While this assumption often helps facilitate a formal analytical treatment, it can be invalid in epidemiologic settings (1). In particular, the probability that exposure is missing may depend on whether a person was exposed, in which case missingness is nonignorable and adopting the MAR assumption can lead to bias. Therefore, generality with regard to the assumed missing data mechanism is beneficial when such generality is supported by feasible study design and analytical procedures.

In the non-MAR situation upon which we focus, parameters underlying the missing data mechanism are inestimable based on the observed data alone. Statistical papers exploring analysis subject to nonignorable missingness often deal with this type of inestimability by making assumptions about the missing data process (4–7). A drawback here is that goodness-of-fit tests to assess models for the missing data mechanism can be inconclusive (3), and sensitivity analysis might be the only recourse (8).

In this paper, we impose relatively few assumptions at the analysis stage, and our approach relates more to prior literature considering two-stage study designs (9) and surveys with multiple waves of respondents (10). Estimability of parameters relies on supplemental data similar to those obtained in validation studies to facilitate correction for misclassification, although here we assume that data can be missing but not misclassified. To emphasize the subtle distinction, we use the term reassessment study. This term implies that, of those persons for whom data are missing, a sample is chosen at random; then, every effort is made to acquire the information by using an assessment method that encourages a high response rate. For example, if exposure was first obtained by questionnaire, nonrespondents chosen for reassessment, or their relatives, might be contacted by telephone to obtain the information and/or incentives for response might be offered. Note that, while validation samples should represent a target population, reassessment samples are drawn from only those for whom data were originally missing.
For simplicity, we assume that the investigator is ultimately successful in acquiring the true data for subjects selected for reassessment, although our approach remains valid if data are MAR subsequent to reassessment (we revisit this point in the Discussion section). Since the case-control study involves sampling by disease status, we assume that only exposure information is missing. While the method described also applies to estimating relative risks and odds ratios in cohort studies with missing disease data and a fixed follow-up period for all subjects (i.e., cumulative incidence studies), obtaining near-complete reassessment information may be more problematic in that context because of losses to follow-up.

PRELIMINARIES

We focus on likelihood-based estimation and inference, but our assumptions regarding the underlying data and the missing data mechanism are natural, and model misspecification is of little concern. We use \( E = 1 \) if exposed, 0 if not to denote exposure status and \( D = 1 \) if diseased, 0 if not for disease status. For case-control designs, the natural model is a product binomial one: we have two independent samples that separately provide estimates of the underlying binomial proportions \( \hat{\pi}_1 = \Pr(E = 1|D = 1) \) and \( \hat{\pi}_0 = \Pr(E = 1|D = 0) \). Independence of the two samples essentially reduces the analysis to a pair of one-sample problems, with the estimated odds ratio being a function of the separate exposure probability estimates.

Consider the familiar \( 2 \times 2 \) data layout in table 1; for now, assume that the data are complete. In the case-control setting, we are concerned with \( \hat{\pi}_1 = n_{11}/n_{d1} \) and \( \hat{\pi}_0 = n_{01}/n_{d0} \). In addition to being intuitive estimators, these are maximum likelihood estimators (MLEs) under the product binomial model, and the MLE for the odds ratio (OR) follows as \( \hat{\text{OR}} = \hat{\pi}_1(1 - \hat{\pi}_0)/[\hat{\pi}_0(1 - \hat{\pi}_1)] \). To form a confidence interval, one typically computes \( \text{Var} [\ln(\text{OR})] = n_{11}^{-1} + n_{10}^{-1} + n_{01}^{-1} + n_{00}^{-1} \), which is derived by using the delta method (11) and reflects the well-known Woolf’s method (12).

If data are missing, the standard estimators \( \hat{\pi}_1 \) and \( \hat{\pi}_0 \) given above may not be the MLEs and, more importantly, may no longer be unbiased in large samples. Assume that we have the correct MLEs (\( \hat{\pi}_1 \), \( \hat{\pi}_0 \)) and their estimated variances. Then, by the delta method, \( \text{Var} [\ln(\text{OR})] = [\hat{\pi}_1(1 - \hat{\pi}_1)]^{-2} \text{Var} (\hat{\pi}_1) + [\hat{\pi}_0(1 - \hat{\pi}_0)]^{-2} \text{Var} (\hat{\pi}_0) \), where the first and second terms represent the approximate variance of the log-odds estimator for diseased and nondiseased subjects, respectively. In the absence of missing data, this formula reduces to the standard variance estimate given previously. As usual, we assume that sample sizes are large enough to ensure approximate normality for \( \ln(\text{OR}) \), after which upper and lower limits of a confidence interval may be computed in the usual way with exponentiation applied to obtain the interval for the odds ratio.

SCENARIOS

Scenario 1: MAR exposure data

We first describe a simple hypothetical situation. Suppose that exposure \( (E) \) is missing for some persons in a case-control study, and the probability that \( E \) is missing depends on disease status \( (D) \) but not on \( E \). Technically, the data are MAR. However, note that, conditional on \( D \), which is the level of analysis, missingness of \( E \) does not depend on the data in any way. Thus, the usual sample proportions \( \hat{\pi}_1 \) and \( \hat{\pi}_0 \) based on the observed data are MLEs, as is the complete-case estimate of the odds ratio.

Scenario 2: non-MAR exposure data with reassessment

Now consider another case-control study in which \( E \) is missing for some subjects. This time, let \( \Pr(E \text{ missing}) \) depend on both \( D \) and \( E \) via the four additional parameters \( p_{mde} = \Pr(E \text{ missing } | D = d, E = e) \) \((d,e = 0,1)\). We want to estimate \( \hat{\pi}_1 = \Pr(E = 1|D = 1) \) and \( \hat{\pi}_0 = \Pr(E = 1|D = 0) \), but since the same estimation procedure is used separately for the two independent samples, we focus on the one-sample problem of obtaining \( \hat{\pi}_1 \) by using exposure data from the diseased subjects. Therefore, assume we have \( n \) subjects with \( D = 1 \), and suppose that \( n_1 \) of them are found to be exposed, \( n_0 \) are found to be unexposed, and \( n_m \) of them have \( E \) missing. The log-likelihood for the observed data is

\[
l(\pi_1, p_{m11}, p_{m10}) = n_1[\ln(1 - p_{m11}) + \ln(\pi_1)] + n_0[\ln(1 - p_{m10}) + \ln(1 - \pi_1)] + n_m[\ln(p_{m11}\pi_1 + p_{m10}(1 - \pi_1)].
\]

However, the parameters \( p_{m11} \) and \( p_{m10} \) are not estimable in the original sample, and it is unrealistic to assume that their values are known. Thus, estimation requires prior information about these two probabilities, additional assumptions, or supplemental data.

If estimates of \( p_{m11} \) and \( p_{m10} \) and their standard errors are available from a prior study involving the same exposure variable, data collection method, and study population, then one could apply pseudo-likelihood methodology (13, 14). Using this technique, one would maximize equation 1 with respect to \( \pi_1 \) after inserting estimates in place of the unknown \( p_{m11} \) and \( p_{m10} \) and then adjust the standard error of the estimate of \( \pi_1 \) for uncertainty in the inserted estimates.

More commonly, prior estimates are unavailable, and we are reluctant to impose strong assumptions about unknown parameters. Thus, we focus on a study-design-based strategy whereby we obtain additional data to facilitate estimation of \( p_{m11} \) and \( p_{m10} \) and we construct the joint log-likelihood for the original and additional data. Our proposed

<table>
<thead>
<tr>
<th>Disease status</th>
<th>Disease ((D = 1))</th>
<th>No disease ((D = 0))</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed ((E = 1))</td>
<td>(n_{11})</td>
<td>(n_{01})</td>
<td>(n_{e1})</td>
</tr>
<tr>
<td>Unexposed ((E = 0))</td>
<td>(n_{10})</td>
<td>(n_{00})</td>
<td>(n_{e0})</td>
</tr>
<tr>
<td>Total</td>
<td>(n_{d1})</td>
<td>(n_{d0})</td>
<td>---</td>
</tr>
</tbody>
</table>

design requires randomly choosing a sample for reassessment from among the \( n_m \) subjects for whom exposure data are missing by applying a fixed selection probability \( p_r \) to each subject. On average, this process generates a reassessment sample of size \( n_s = n_m \times p_r \). An acceptable alternative approach is to select the value \( n_r \) and assume that \( p_r = n_s/n_r \).

This design gives rise to five distinct types of likelihood contributions, specifically, 1) \( n_s \) subjects for whom \( E = 1 \); 2) \( n_m \) subjects for whom \( E = 0 \); 3) \( n_s \) subjects for whom \( E \) is originally missing but subsequently verified as \( E = 1 \) by reassessment; 4) \( n_m \) subjects for whom \( E \) is missing but who are reassessed and found to have \( E = 0 \); and 5) \( n_m \) subjects for whom \( E \) is missing and who are not selected for reassessment. To construct the log-likelihood, we determine each type of contribution; for example, type 1 subjects contribute \( \Pr(E \text{ observed and } E = 1 | D = 1 \text{ observed}) = \Pr(E \text{ observed } | E = 1 | D = 1) = \Pr(E \text{ observed } | D = 1) = (1 - p_m11) \times \pi_1 \). Similarly, those of types 2–5 contribute

\[
\begin{align*}
l(p_m11, p_m10) &= \sum_{e=0}^{1} \left[ n_e \ln(p_m e) + n_{me} \ln(p_m e_m) \right] \\
&+ n_{mr} \ln(p_m r)
\end{align*}
\]

Closed-form expressions are available for the corresponding MLEs under equation 2. In particular, we obtain \( \hat{p}_{m11} = n_1(n_1n_m + n_{r1})^{-1} \) and \( \hat{p}_{m10} = n_0(n_0n_m + n_{r0})^{-1} \). The MLE for \( \pi_1 \) is a solution to a quadratic equation (details omitted) and reduces to the intuitive form

\[
\hat{\pi}_1 = \{n_1 + n_m[n_{r1} / (n_1 + n_{r0})]\}/n.
\]

Thus, the MLE effectively replaces the unknown number of originally missing values for which \( E = 1 \) by a logical estimate based on the reassessment study, namely, \( n_m[n_{r1} / (n_{r1} + n_{r0})] \). Such a “fill-in” estimator is familiar in problems involving contingency tables with missing data (2, 3).

Equation 3 is equivalent to an estimator presented previously (10) for estimating prevalence in surveys with two waves of respondents. In contrast to equation 3, note that the complete-case estimate for \( \pi_1 \) is simply \( n_1 / (n_1 + n_0) \).

In the case-control study, we also maximize the equivalent of equation 2 separately by using exposure data for nondiseased subjects to obtain \( \hat{\pi}_0, \hat{p}_{m01}, \text{and} \hat{p}_{m00} \). The MLE of the odds ratio follows directly once we have \( \hat{\pi}_1 \) and \( \hat{\pi}_0 \). For standard error estimates of \( \hat{\pi}_1 \) and \( \hat{\pi}_0 \) to facilitate an approximate confidence interval as described previously, we provide details for computing the expected information matrix in Appendix A.

Table 2 describes simulations illustrating potential for bias and invalid inference in complete-case analysis, as contrasted with the maximum likelihood approach. Fifteen settings were considered to cover a range of values for the crude odds ratio and for the missingness parameters \( [p_{mde} (d,e = 0, 1)] \). For each setting, 5,000 data sets were generated, each with 250 cases and 250 controls, missing data according to the missingness probabilities, and 25 percent reassessment samples on average (i.e., \( p_r = 0.25 \) regardless of disease). Means for \( \hat{\pi}_1 \) and \( \hat{\pi}_0 \) means and standard deviations of the odds ratio estimates, and mean widths and

### Table 2. Simulation results for case-control study analysis of the crude odds ratio

<table>
<thead>
<tr>
<th>Setting</th>
<th>( \pi_1 )</th>
<th>( \pi_0 )</th>
<th>OR ( \hat{\pi}_0 )</th>
<th>Mean ( \hat{\pi}_1 )</th>
<th>Mean ( \hat{\pi}_0 )</th>
<th>Mean OR (SD)</th>
<th>Mean 95% CI width (coverage)</th>
<th>Mean 95% CI width (coverage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.3</td>
<td>0.4</td>
<td>0.64</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>1.00</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>3</td>
<td>0.3</td>
<td>0.3</td>
<td>1.56</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>4</td>
<td>0.3</td>
<td>0.3</td>
<td>1.56</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>0.3</td>
<td>1.56</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>6</td>
<td>0.3</td>
<td>0.3</td>
<td>1.56</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>7</td>
<td>0.3</td>
<td>0.3</td>
<td>1.56</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>8</td>
<td>0.3</td>
<td>0.3</td>
<td>1.56</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>9</td>
<td>0.3</td>
<td>0.3</td>
<td>1.56</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
<td>0.3</td>
<td>1.56</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>11</td>
<td>0.3</td>
<td>0.3</td>
<td>1.56</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>12</td>
<td>0.3</td>
<td>0.3</td>
<td>1.56</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>13</td>
<td>0.3</td>
<td>0.3</td>
<td>1.56</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>14</td>
<td>0.3</td>
<td>0.3</td>
<td>1.56</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>15</td>
<td>0.3</td>
<td>0.3</td>
<td>1.56</td>
<td>2.83</td>
<td>3.20</td>
<td>0.36</td>
<td>0.30</td>
<td>0.40</td>
</tr>
</tbody>
</table>

\* \( \pi_1 = \Pr(E = 1 | D = 1) \) and \( \pi_0 = \Pr(E = 1 | D = 0) \), where \( E \) and \( D \) denote exposure status and disease status, respectively.

\( p_{mde} = \Pr(E \text{ missing } | D = d, E = e) (d,e = 0, 1). \)

\( OR, \) odds ratio; SD, standard deviation; CI, confidence interval.
coverages of 95 percent confidence intervals are displayed under the complete-case and maximum likelihood strategies. Note that the “complete-case” approach evaluated reflects what would be done in practice, that is, it ignores the reassessment data.

With significant disparity among the four missingness probabilities, Table 2 illustrates both severe overestimation (settings 1–3) and severe underestimation (settings 13–15) of the odds ratio when complete-case analysis is used, together with abysmal confidence interval performance (coverages near 10 percent). With less disparity in the probabilities (settings 4–6 and 10–12), complete-case analysis remains subject to (less extreme) bias and subnominal confidence interval coverage. Should all missingness probabilities be equal (settings 7–9), there are no problems with complete-case analysis. The maximum likelihood analysis displays minimal bias and nearly nominal confidence interval coverage throughout, and it outperforms the complete-case approach except in the special cases (settings 7–9).

Scenario 3: estimating a common odds ratio across strata with non-MAR exposure data

Here we extend the previous considerations to allow adjustment for confounding variables via stratification (i.e., a Mantel-Haenszel-type approach). Specifically, suppose there are S strata defined by such variables and we want to estimate a common odds ratio subject to missing exposure data. In the event of MAR exposure data (i.e., if scenario 1 applies within all strata), then the usual Mantel-Haenszel common odds ratio estimate and its confidence interval (12) are valid. In the non-MAR case, missingness probabilities may vary by disease and exposure status, as before, as well as across strata. We adjust our prior notation to accommodate these features by using the subscript s to identify the stratum. The five parameters specific to the sth stratum are

\[
\pi_{sd} = \text{Pr}(E \text{ missing} \mid D = d, E = e, \text{stratum} = s) \quad \text{and} \quad \pi_{s} = \text{Pr}(E = 1 \mid D = 1, \text{stratum} = s) \quad (d,e = 0,1; s = 1, \ldots, S),
\]

while the parameter of main interest is \( \psi \) (the common log odds ratio). Since the investigator may vary reassessment probabilities across strata, we denote these as \( p_{rs} \) (s = 1, . . ., S).

For stratum s, 10 types of observations are possible. Specifically, we have \( n_{sde} \) subjects for whom \( E \) is originally observed, \( D = d, \) and \( E = e \) (d,e = 0,1). Likewise, there are \( n_{sde} \) subjects for whom \( D = d \) and \( E \) is originally missing but who are reassessed and found to have \( E = e \) (d,e = 0,1). Finally, let \( n_{msd} \) represent the number of subjects with \( D = d \) and \( E \) missing who are not selected for reassessment (d = 0,1). The first four types of subjects make likelihood contributions of the forms

\[
\begin{align*}
(1 - p_{smd}) \times \pi_{sd} & \quad \text{and} \quad p_{sd} = (1 - p_{smd}) \times (1 - \pi_{sd}) \quad (d = 0,1), \\
\pi_{0d} = \text{Pr}(E = 1 \mid D = 0, \text{stratum} = s) & = \pi_{s1} \times (1 + (1 - \pi_{s1}) \times \exp(\psi)) \\
(1 - \pi_{0d}) & \quad \text{and} \quad p_{rd0} = (1 - \pi_{0d}) \times (1 - \pi_{rd1}) \times \exp(\psi).
\end{align*}
\]

The next four types of subjects make contributions of the forms

\[
\begin{align*}
p_{rds} = p_{rs} \times p_{smd} \times \pi_{sd} & \quad \text{and} \quad p_{rds} = p_{rs} \times p_{smd} \times (1 - \pi_{sd}) \quad (d = 0,1), \\
(1 - \pi_{rd0}) & \quad \text{and} \quad p_{rds} = (1 - \pi_{rs}) \times (1 - \pi_{rd1}) \times \exp(\psi).
\end{align*}
\]

The likelihood specific to the sth stratum is a product of terms corresponding to these 10 types and is analogous to that given in equation 2 for crude analysis. The overall log-likelihood is a sum of stratum-specific log-likelihoods and (neglecting combinatoric terms) can be written as

\[
l(\psi, \pi, p_{rs}) = \sum_{s=1}^{S} \left\{ \sum_{d=0}^{1} [n_{msd} \times \ln(p_{msd}^*) + \sum_{e=0}^{1} n_{sde} \times \ln(p_{sde}) + n_{sde} \times \ln(p_{sde}) \bigg] \right\}, \tag{4}
\]

where \( \pi \) and \( p_{rs} \) are \((S \times 1)\) and \((4S \times 1)\) vectors, respectively, containing the exposure and missingness probabilities \( \pi_{s} \) and \( p_{smd} \) (s = 1, . . ., S; d,e = 0,1).

An advantage of the likelihood approach is flexibility to consider more parsimonious models (e.g., one could make and test the hypothesis that all or some missingness parameters are equal across strata). However, while a program implemented by using SAS IML software (15) to numerically maximize equation 4 and obtain standard errors is available from the authors, we can no longer readily derive closed-form expressions for the MLEs and the expected information matrix. An attractive alternative that makes use of the closed-form results is the following asymptotically unbiased, weighted-average estimator:

\[
\hat{\psi} = \frac{\sum_{s=1}^{S} \psi_{s} \hat{\nu}_{s}}{\sum_{s=1}^{S} 1/\hat{\nu}_{s}}, \tag{5}
\]

where \( \psi_{s} \) and \( \hat{\nu}_{s} \) represent the sth stratum-specific MLE for the log-odds ratio and its estimated variance, respectively. A valid estimator for \( \text{Var}(\hat{\psi}) \) is

\[
\left( \sum_{s=1}^{S} 1/\hat{\nu}_{s} \right)^{-1},
\]

thus, one can conduct point and confidence interval estimation of the common odds ratio with informatively missing exposure data by applying the methodology for the crude analysis within separate strata and combining the results.

Table 3 summarizes a brief simulation study to evaluate the common odds ratio estimator in equation 5. For comparison, we also evaluated the MLE based on equation 4 and the standard Mantel-Haenszel odds ratio estimator based on complete-case analysis (12). Missingness probabilities were varied across three strata as indicated, and the following settings are illustrated: 1) exposure missing completely at random in each stratum, 2) moderate association between missingness and exposure status, and 3) extreme association between missingness and exposure status. Each setting was simulated 5,000 times with a true common odds ratio of 1 and with 40 percent reassessment probabilities and 500 subjects in each stratum; other conditions were as specified in the footnote to Table 3. Note that the weighted estimator (equation 5) performed just as well as the MLE; in simula-
TABLE 3. Simulation results for case-control study analysis of the common odds ratio

<table>
<thead>
<tr>
<th>Setting†</th>
<th>Stratum‡</th>
<th>Among diseased</th>
<th>Among nondiseased</th>
<th>Mean OR (SD)**</th>
<th>Mean 95% CI width (coverage)</th>
<th>Maximum likelihood¶</th>
<th>Weighted average§</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(p_{m11})</td>
<td>(p_{m10})</td>
<td>(p_{m01})</td>
<td>(p_{m00})</td>
<td>Mean OR (SD)</td>
<td>Mean 95% CI width (coverage)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>1.012 (0.127)</td>
<td>0.535 (95.9%)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>1.166 (0.156)</td>
<td>0.643 (83.1%)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>1.348 (0.182)</td>
<td>0.815 (52.4%)</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>1.012 (0.127)</td>
<td>0.535 (95.9%)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.20</td>
<td>0.30</td>
<td>0.30</td>
<td>0.22</td>
<td>1.166 (0.156)</td>
<td>0.643 (83.1%)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.24</td>
<td>0.26</td>
<td>0.26</td>
<td>0.24</td>
<td>1.348 (0.182)</td>
<td>0.815 (52.4%)</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.10</td>
<td>0.30</td>
<td>0.50</td>
<td>0.20</td>
<td>1.006 (0.138)</td>
<td>0.535 (95.4%)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.20</td>
<td>0.30</td>
<td>0.40</td>
<td>0.10</td>
<td>1.348 (0.182)</td>
<td>0.815 (52.4%)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.30</td>
<td>0.30</td>
<td>0.20</td>
<td>0.40</td>
<td>1.348 (0.182)</td>
<td>0.815 (52.4%)</td>
</tr>
</tbody>
</table>

* Settings 1, 2, and 3, respectively, represent no, moderate, and extreme association between missingness and true exposure within each stratum.
† Stratum-specific sample sizes (no. of cases, no. of controls) were (300, 200), (200, 300), and (250, 250) for strata 1, 2, and 3, respectively.
‡ \(p_{m00}\) = \(Pr(E\text{ missing }| D = d, E = e)\) (d,e = 0,1) for each stratum as indicated.
§ Standard Mantel-Haenszel common odds ratio (OR) estimator and confidence interval (CI) based on complete-case data.
¶ Based on numerical maximization of equation 4 (refer to the text).
# Based on the estimator in equation 5 (refer to the text).
** SD, standard deviation.

Comments and special cases

Before committing to additional sampling, it would help to obtain some sense of the magnitude of bias to be expected with complete-case analysis or what conditions would ensure no bias. Regarding the latter, we offer the following about the crude (single-stratum) analysis: 1) If the missingness probabilities are equal for both diseased and nondiseased subjects (i.e., \(p_{m11} = p_{m10}\) and \(p_{m01} = p_{m00}\)), then the complete-case approach is valid and produces the MLE of the odds ratio (this reverts to scenario 1, and a special case was illustrated in cases 7–9 (table 2)); 2) if missingness is related to exposure but not to disease status (i.e., \(p_{m11} = p_{m01}\) and \(p_{m10} = p_{m00}\) but \(p_{m11} \neq p_{m01}\) and \(p_{m10} \neq p_{m00}\)), then the complete-case estimators \(\hat{\tau}_1\) and \(\hat{\tau}_2\) are biased but the complete-case odds ratio estimator is valid. The validity of the complete-case odds ratio when missingness is unrelated to disease also holds in the case of continuous exposure.

Aside from the preceding special cases, note that complete-case analysis yields biased estimators and confidence interval coverage approaching zero percent as sample size increases. The magnitude of bias in complete-case exposure probability (i.e., \(\hat{\tau}_1\) or \(\hat{\tau}_2\)) estimates depends primarily on two factors: 1) the amount of missing exposure data and 2) the extent of association between missingness and the true exposure classification (the latter determined by the values of \(p_{mde}\) (d,e = 0,1)). In a spirit similar to that of prior authors (e.g., Vach and Blettner (2)), one can show that, in large samples, the complete-case odds ratio estimator approaches the quantity \(OR \times \frac{\{(1 - p_{m11}) \times (1 - p_{m00})\} / \{(1 - p_{m10}) \times (1 - p_{m01})\}}{\}. This expression provides a preliminary guide to assess potential for bias based on various assumed values of the missingness probabilities.

EXAMPLE

For illustration, we use observed cell counts from an example by Rosner (12). The data stem from a case-control study of association between breast cancer and age at first birth among mothers from six countries (16). To exemplify the methods, we augment the observed data with missing and reassessment study data in an arbitrary but reasonable way.

Data from 13,465 women in the breast cancer study are presented in table 4. Suppose hypothetically that there were 300 additional cases for whom exposure (age at first birth) was missing. Although this information would be unavailable for analysis, let us further assume that 72 (24 percent) of these cases were in fact aged 30 (table 4). To exemplify the methods, we augment the observed data with missing and reassessment study data in an arbitrary but reasonable way.

Table 4. Data from a case-control study of the association between breast cancer and age at first birth*

<table>
<thead>
<tr>
<th>Age at first birth (years)</th>
<th>Disease status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case ((D + 1))</td>
<td>Control ((D - 0))</td>
</tr>
<tr>
<td>(\geq 30) ((E = 1))</td>
<td>683</td>
</tr>
<tr>
<td>(&lt; 30) ((E = 0))</td>
<td>2,537</td>
</tr>
<tr>
<td>Total</td>
<td>3,220</td>
</tr>
</tbody>
</table>

† D, disease status; E, exposure status.
Missing Data in the Case-Control Study

CONCLUSIONS

We have described a likelihood-based approach that permits valid estimation and inference for crude and common odds ratios in case-control studies subject to nonignorably missing exposure data. As with more familiar procedures for correcting misclassification, the approach requires an additional data collection effort for a subsample of participants. This additional effort, which we called a reassessment study, permits identification of parameters associated with the missing data mechanism and facilitates estimation without imposing assumptions or restrictions. The approach for estimating the common odds ratio allows missingness probabilities to vary (as might be expected) by stratification variables.

The practical benefit of the design considered is that an exceptional (here, meaning likely to obtain complete information) data collection effort need be applied for only some of the participants. Thus, valid estimates are derived without incurring the full cost of applying that same effort across the board. As mentioned earlier, we assumed that the investigator is successful in obtaining the exposure information for subjects selected for reassessment; however, this goal may well be overly optimistic. Missing information that cannot be recovered for some such subjects might reasonably be accommodated by considering those persons as making type 5 likelihood contributions (scenario 2), so they would not contribute to estimating the MLE via equation 3. The implicit assumption is that, unlike for the original data collection effort, the probability of the redoubled effort producing a missing value does not depend on actual exposure status (although it can depend on disease status). This assumption connects equation 3 with a prevalence estimator given previously (10). The MAR assumption at this second (as opposed to first) level of sampling will often be more reasonable (17); in any case, the impact from violation of MAR at the second level should be small unless the redoubled effort is unsuccessful for a substantial number of subjects.

Nevertheless, in the case of large-scale studies with a considerable amount of missing data that remains after the second stage, one might consider extending design and analytical considerations to incorporate one or more additional reassessment stages. As a hypothetical example, initial exposure assessment might be made by questionnaire, the first reassessment by requesting information over the telephone, and a second reassessment among those not responding to the first one by offering incentives. Further stages could be contemplated, although their usefulness decreases greatly. Appendix B illustrates construction of the likelihood based on a three-stage scenario, and a program suited for such analysis is available from the authors. Limited simulation studies in this context indicate that the original two-stage approach (i.e., with a single reassessment) can provide markedly better estimates and inferences than complete-case analysis, and it suffers relative to a three-stage design only with substantially incomplete initial reassessment and under unusually extreme non-MAR conditions after the second stage.

Use of the proposed method is most feasible in studies in which a noteworthy amount of data is missing and for which resources permit a reassessment study extensive enough to obtain reasonable precision in the estimates of \( p_{nde} \) (i.e. 0.1) within each stratum. If very little information is available for this purpose, then variance estimation as described in Appendix A can be unreliable. When a significant amount of data is missing, even a mild association between missingness and exposure status could introduce bias into complete-case results, making reassessment a potentially attractive option.

The ability to conduct an essentially assumption-free analysis facilitated by supplemental data owes partly to the fact that the 2 x 2 table is the foundation of the considerations given here. Nevertheless, the Mantel-Haenszel-like approach is the fundamental facilitator for covariate-adjusted analyses that are usually the ultimate basis for epidemiologic findings, and our method extended naturally to that setting. Adaptation of our approach to multilevel categorical exposures should be straightforward. Further considerations to incorporate supplemental data into more general models, that is, those with continuous exposure and
covaerates, effect modification, and/or missing covariate as well as exposure data, would be of interest and might be facilitated via the expectation-maximization (EM) algorithm (18). Such considerations may help reduce concerns about model misspecification that can linger when existing approaches for handling nonignorable missing data are applied in the context of regression analysis for case-control studies.

REFERENCES

APPENDIX A

Expected Information Matrix Associated with Equation 3

Here, it is useful to rewrite the log-likelihood in equation 3 as

\[ l(\pi_1, p_{m11}, p_{m10}; x) = \sum_{i=1}^{n} \{x_{i1} \ln[(1 - p_{m11})(1 - \pi_1)] + x_{i2} \ln[(1 - p_{m10})(1 - \pi_1)] + x_{i3} \ln(p_{m11}\pi_1) + x_{i4} \ln(p_{m10}(1 - \pi_1)) \}, \]

where

\[ x_{ij} = \begin{cases} 1, & \text{if } ith \text{ observation is of type } j \\ 0, & \text{otherwise} \end{cases}, \quad j = 1, \ldots, 5. \]

Note that \( n = n_1 + n_0 + n_{11} + n_{00} + n_{10} = \sum_{i=1}^{n} x_{i1} + \ldots + \sum_{i=1}^{n} x_{i5}, \) and \( \sum_{i=1}^{n} X_{ij} \sim \text{Binomial}(n, p_j), j = 1, \ldots, 5, \) with the p_j’s equal to the respective arguments of the five logarithms in the above log-likelihood.

The estimated \((3 \times 3)\) covariance matrix \( \nabla l[\tilde{\pi}_1, \tilde{p}_{m11}, \tilde{p}_{m10}] = \tilde{\theta}^{-1}, \) where \( \tilde{\theta} \) has individual terms

\[ \frac{\partial^2 l}{\partial \pi_1^2} = n[(1 - p_{m11} + p_{m10}) + n[(1 - p_{m10} + p_{m11})^2 \frac{1 - \pi_1}{p_{m11}\pi_1 + p_{m10}(1 - \pi_1)}), \theta_{12} = \theta_{21} = E \]

\[ = n \pi_1 + np_{m11} + n \pi_1^2 (1 - p_{m11}), \theta_{13} = \theta_{31} = E \]

\[ = n \pi_1 + np_{m11} + n \pi_1^2 (1 - p_{m11}), \theta_{22} = E \]

\[ = n \pi_1 (1 - \pi_1)(1 - p_{m11} + p_{m10}(1 - \pi_1)). \]
and

$$\hat{\theta}_{33} = E \left( - \frac{\partial^2 l}{\partial \pi_{m10}} \right) = \frac{n(1 - \pi_1)}{(1 - p_{m10})} + \frac{n p_1 (1 - \pi_1)}{p_{m10}} + \frac{n (1 - \pi_1)^2 (1 - p_1)}{p_{m11} \pi_1 + p_{m10} (1 - \pi_1)},$$

and we replace the unknown parameters ($\pi_1, p_{m11}, p_{m10}$) by their respective MLEs. Recall that the selection proportion $p_r$ is treated as a fixed known constant. These calculations were used to obtain standard error estimates for $\hat{\pi}_1$ and $\hat{\pi}_0$ and subsequently (as described in the Preliminaries section) standard errors for the odds ratio in the simulations summarized in tables 2 and 3. In the unlikely event that $\hat{p}_{m11}$ or $\hat{p}_{m10}$ equals zero, the above results must be amended slightly after removing a null term from the log-likelihood.

**APPENDIX B**

**Likelihood to Incorporate a Second Reassessment Study**

Let $p_{ra}$ and $p_{rb}$ represent initial and subsequent reassessment probabilities designated by the investigator. We seek to estimate $\pi_1 = \Pr(E = 1|D = 1)$ by using data from diseased subjects together with the missingness probabilities $p_{mea} = \Pr(E$ missing$|D = 1, E = e)$ and $p_{meb} = \Pr(E$ missing$|D = 1, E = e)$ ($e = 0, 1$), where the $a$ and $b$ subscripts refer to the initial and first reassessment sampling conditions, respectively.

Assume that there are $n_e$ subjects with $E$ originally observed and $E = e$ and $n_{rea}$ subjects with $E$ missing but selected for the first reassessment and found to have $E = e$ ($e = 0, 1$). Furthermore, let there be $n_{reb}$ subjects with $E$ missing who are selected for the first reassessment but still fail to provide data but are then selected for the second reassessment and found to have $E = e$ ($e = 0, 1$). Finally, let $n_{mra}$ be the number with $E$ missing but not selected for the first reassessment, and let $n_{mrb}$ be the number with $E$ missing, initially reassessed but still failing to provide data, who are not selected for the second reassessment. Nonrespondents to the second reassessment study can be viewed as of the final type, corresponding to a “MAR at the third level” assumption (refer to the Conclusions section of the text). When the same subscripts are used, the eight possible types of observations contribute $p_1 = (1 - p_{m1a}) \times \pi_1$.

$$p_0 = (1 - p_{m0a}) \times (1 - \pi_1), p_{ra} = p_{ra} \times p_{m1a} \times (1 - p_{m1b}) \times \pi_1, p_{rb} = p_{rb} \times p_{m1a} \times (1 - p_{m1b}) \times \pi_1, p_{m1a} = p_{m1a} \times p_{m0a} \times (1 - p_{m1b}) \times (1 - \pi_1),$$

and

$$p_{m1b} = p_{m1a} \times (1 - p_{m1b}) \times [p_{m1a} \times p_{m1b} \times \pi_1 + p_{m0a} \times p_{m0b} \times (1 - \pi_1)],$$

respectively. The log-likelihood is then

$$l(\pi_1, p_{m1a}, p_{m1b}, p_{m0a}, p_{m0b}) = \sum_{e = 0}^{1} \left( n_e \ln(p_e) + n_{rea} \ln(p_{rea}) + n_{reb} \ln(p_{reb}) \right) + n_{mra} \ln(p_{mra}) + n_{mrb} \ln(p_{mrb}).$$