Application of the Missing-Indicator Method in Matched Case-Control Studies with Incomplete Data

Mark Huberman and Bryan Langholz

A common practice in matched case-control studies with incomplete data is to perform two analyses in parallel: a matched analysis of the complete pairs and an unmatched analysis of all subjects carried out after breaking the matching in the complete pairs. The missing-indicator method, which has the advantage of making use of the data in the incomplete pairs while still preserving the matching in the complete pairs, is recommended as an alternative method of analysis. It is shown here that its estimate of the odds ratio is a compromise between the odds ratios estimated by a matched analysis of the complete pairs and an unmatched analysis of the incomplete pairs. The method is illustrated using data from a matched case-control study of the risk of childhood leukemia from exposure to residential electric and magnetic fields. Am J Epidemiol 1999; 150:1340-5.

In a matched case-control study of the risk of childhood leukemia from exposure to residential electric and magnetic fields, London et al. (1, 2) matched cases and controls according to sex, ethnicity, and age. Because a significant fraction (~60 percent) of the matched pairs contained either a case or a control whose exposure measurement was missing, London et al. used two different analytical methods to estimate the odds ratio of disease. In the first analysis, only complete pairs in which both the case and the control had exposure measurements were considered, and the information contained in the incomplete pairs was ignored. A matched analysis was used to estimate the odds ratio from the complete pairs. In the second analysis, the matching was broken in the complete pairs and data from the complete pairs were pooled with data from the incomplete pairs. An unmatched analysis was used to estimate the odds ratio from the pooled data. This study and other such studies (3-9) raise the following question: Is it possible to analyze data by a method which uses the information in incomplete pairs while still preserving the matching in the complete pairs?

In the missing-indicator method, an indicator is introduced for each subject which has the value 1 if the exposure covariate for that subject is missing and the value 0 if it is observed; and the exposure covariate of a subject is assigned the value 0 if it is missing. The missing-indicator method has been discussed previously (10), but not in the context of matched case-control studies. We show here that the missing-indicator method has just the above desired property: Its estimate of the odds ratio is a compromise between the odds ratio estimated by a matched analysis of the complete pairs and that estimated by an unmatched analysis of the incomplete pairs. Moreover, the missing-indicator method has the practical advantages of maintaining the data structure, using standard software, and avoiding the inconvenience of parallel analyses.

METHODS

Dichotomous exposure

The data structure for a matched case-control study with missing dichotomous exposure information is shown in table 1. Here $n_{k,h}$ is the number of case-control pairs in which the case has exposure level $k$ and the control has exposure level $h$, where the possible values of the exposure are either exposed (1), unexposed (0), or missing (M). In the missing-indicator method, the odds ratio, i.e., the odds of disease for exposed persons compared with the unexposed, is estimated by analyzing the data in table 1 as a matched
case-control study with multiple exposure levels. The conditional likelihood \( L \) of the data is then (11, p. 182)

\[
L = \frac{\psi_i}{1 + \psi_i} \begin{pmatrix} n_{10} & n_{01} & n_{00} \\ n_{11} & n_{01} & n_{00} \end{pmatrix} \left( \frac{1}{1 + \psi_1} \right)^{n_{10}} \left( \frac{1}{1 + \psi_M} \right)^{n_{00}}
\]

\[
\times \left( \frac{\psi_M}{\psi_M + \psi_1} \right)^{n_{11}} \left( \frac{\psi_M}{\psi_M + \psi_1} \right)^{n_{01}},
\]

(1)

where \( \psi_i \) and \( \psi_M \) are the odds ratios for exposure levels 1 and \( M \), respectively, compared with exposure level 0. It is instructive to express this likelihood as a product of two factors, \( L = L_1 L_2 \), where

\[
L_1 = \left( \frac{\psi_i}{1 + \psi_i} \right)^{n_{10}} \left( \frac{1}{1 + \psi_1} \right)^{n_{01}}
\]

(2)

and

\[
L_2 = \left( \frac{\psi_M}{1 + \psi_M} \right)^{n_{11}} \left( \frac{1}{1 + \psi_M} \right)^{n_{01}}
\]

\[
\times \left( \frac{\psi_M}{\psi_M + \psi_1} \right)^{n_{11}} \left( \frac{\psi_M}{\psi_M + \psi_1} \right)^{n_{01}}.
\]

The first factor, \( L_1 \), is the likelihood contribution from the complete pairs. Taking the log of \( L_1 \), differentiating \( \log L_1 \) with respect to \( \psi_i \), and setting the derivative equal to zero, we find that \( L_1 \) is a maximum when

\[
\hat{\psi}_i = \frac{n_{10}}{n_{01}}
\]

(4)

which is the maximum likelihood estimate of the odds ratio from a matched analysis of the complete pairs (11). The second factor, \( L_2 \), is the likelihood contribution from the incomplete pairs. Taking the log of \( L_2 \), differentiating \( \log L_2 \) with respect to \( \psi_1 \) and \( \psi_M \), setting the derivatives equal to zero, and solving the coupled equations, we find that \( L_2 \) is a maximum when

\[
\hat{\psi}_1 = \left( \frac{r_{11} r_{01}}{r_{10} r_{00}} \right)
\]

(5)

which is the maximum likelihood estimate of the odds ratio from an unmatched analysis of the incomplete pairs (11). It follows that the estimate of the odds ratio, \( \hat{\psi}_1 \), which maximizes the overall likelihood, \( L = L_1 L_2 \), is a compromise between these two values.

In the following section, we show that the likelihood \( L_2 \) in equation 3 is identical to the likelihood from an unmatched analysis of the incomplete pairs. Consequently, breaking the matching results in an "unmatched" likelihood, \( L^* = L_1^* L_2 \), where \( L_2 \) is as in equation 3 and \( L_1^* \) is the contribution from the broken complete pairs. Clearly, the odds ratio estimate based on \( L \) is preferable because it is less vulnerable to the potential confounding that can arise when paired sets are broken.

**General exposure**

In the missing-indicator method, a missing indicator \( M \) is introduced for each subject, which has the value \( M = 1 \) if the exposure covariate \( Z \) for that subject is missing and the value \( M = 0 \) otherwise. Furthermore, the exposure covariate \( Z \) is assigned the value 0 if it is missing. The covariates of each subject are then \( Z \) and \( M \). In general, \( Z \) can be a vector of covariates; in this case, \( M = 1 \) and \( Z = 0 \) if any component of \( Z \) is missing. For example, in a matched case-control pair, if all but one covariate is observed for the case and all covariates are observed for the control, then the missing-indicator method discards the information in the covariates of the case but retains the information in the covariates of the control. (In contrast, a complete-case analysis would discard the information in the covariates of both the case and the control.) The conditional probability of having a matched case-control pair in which the covariates of the case are \( Z_1 \) and \( M_1 \) and the covariates of the control are \( Z_0 \) and \( M_0 \) is then (11)

\[
\text{Pr}(Z_1, M_1, Z_0, M_0) = \frac{\exp(\beta Z_1 + \gamma M_1)}{\exp(\beta Z_1 + \gamma M_1) + \exp(\beta Z_0 + \gamma M_0)},
\]

(6)

where \( \beta \) is the log odds ratio when \( Z \) changes by unity. For a matched case-control study with incomplete exposure data, the conditional likelihood is then

\[
L(\beta, \gamma) = \prod_{i=1}^{N} \frac{\exp(\beta Z_{1i} + \gamma M_{1i})}{\exp(\beta Z_{1i} + \gamma M_{1i}) + \exp(\beta Z_{0i} + \gamma M_{0i})},
\]

(7)
where $Z_i$ and $M_i$, $i = 1, \ldots, N$, are the covariates of the cases and $Z_0$ and $M_0$, $i = 1, \ldots, N$, are the covariates of the controls.

For a complete case-control pair in which both exposures are observed, $M_1 = 0$ and $M_0 = 0$, so the conditional probability in equation 6 reduces to

$$\frac{\exp(\beta Z_i)}{\exp(\beta Z_i) + \exp(\beta Z_0)},$$  

which is the conditional probability for a case-control pair (11). For a case-control pair in which the exposure of the case is observed but the exposure of the control is missing, $M_1 = 0$, $Z_0 = 0$, and $M_0 = 1$, so the conditional probability in equation 6 reduces to

$$\frac{\exp(\beta Z_i)}{\exp(\beta Z_i) + \exp(\gamma)} = \frac{\exp(-\gamma + \beta Z_i)}{1 + \exp(-\gamma + \beta Z_i)},$$  

which is the unconditional probability for a control (11). (The constant $\gamma$ is the negative of the usual nuisance parameter $\alpha$.) Similarly, for a case-control pair in which the exposure of the control is observed but the exposure of the case is missing, $Z_1 = 0$, $M_1 = 1$, and $M_0 = 0$, so the conditional probability in equation 6 reduces to

$$\frac{\exp(\gamma)}{\exp(\gamma) + \exp(\beta Z_0)} = \frac{1}{1 + \exp(-\gamma + \beta Z_0)},$$  

which is the unconditional probability for a control (11). (For a case-control pair in which the exposures of the case and control are both missing, $Z_1 = 0$, $M_1 = 1$, $Z_0 = 0$, and $M_0 = 1$, so the conditional probability in equation 6 reduces to a constant factor which does not depend on the regression parameters to be estimated and therefore contributes nothing to the estimation of $\beta$.) Thus, the conditional likelihood in equation 7 is a product of the conditional likelihood of the complete pairs and the unconditional likelihood of the incomplete pairs. It follows that the estimate of $\beta$ which maximizes the total likelihood in equation 7 is a compromise between the estimate of $\beta$ from a matched analysis of the complete pairs and the estimate from an unmatched analysis of the incomplete pairs.

**APPLICATION**

In London et al.'s matched case-control study of the risk of childhood leukemia due to electromagnetic radiation exposure (1), the risk of leukemia was reported for several different measures of exposure, including the arithmetic mean of 24-hour magnetic field measurements in the child's bedroom, low-power spot measurements of the magnetic field in the child's bedroom, and the Wertheimer-Leeper classification of the residential wiring configuration. Each of these exposure measurements has missing data, the proportions of missing subjects being 34 percent, 46 percent, and 12 percent, respectively (see table 2).

To illustrate the missing-indicator method described above, we have analyzed the association of leukemia with each of these measures of exposure in four different ways (see tables 3–5). In the first panel of each table, the matching is broken, and the data from both the complete pairs and the incomplete pairs are analyzed in an unmatched analysis. This method, which uses all of the data, has the disadvantage of introducing possible bias towards the null by breaking the matching in the complete pairs. The results reported by London et al. (1) were obtained using this method. In the second panel of the tables, the data in the incomplete pairs are dropped and the data in the complete pairs are analyzed in a matched analysis. This method, which preserves the matching in the complete pairs, has the disadvantage of discarding the information in the incomplete pairs. In the third panel of the tables, the data in the complete pairs are dropped and the data in the incomplete pairs are analyzed in an unmatched analysis. This is included only to show the contribution of the subjects from the incomplete pairs. Finally, the fourth panel shows the results from the missing-indicator method, which uses the information in the incomplete pairs while still preserving the matching in the complete pairs. In this panel, the estimates for the “missing” category are also included, although they would not necessarily be reported; from equations 9 and 10, the log odds ratio for “missing” relative to the baseline exposure is the negative of the usual nuisance parameter, i.e., $\gamma = -\alpha$.

In tables 3–5, it can be seen that the odds ratio obtained by the missing-indicator method is a compro-

---

**TABLE 2. Numbers of cases and controls entered into a study of leukemia risk and electromagnetic field exposure,* by type of exposure measurement available, Los Angeles County, California, 1980–1987**

<table>
<thead>
<tr>
<th>No. of cases</th>
<th>No. of controls</th>
<th>No. of matched pairs</th>
<th>Missing subjects (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects with interview data</td>
<td>232</td>
<td>232</td>
<td>232</td>
</tr>
<tr>
<td>Subjects with 24-hour magnetic field measurements</td>
<td>164</td>
<td>144</td>
<td>108</td>
</tr>
<tr>
<td>Subjects with a spot magnetic field measurement in the child's bedroom</td>
<td>140</td>
<td>109</td>
<td>71</td>
</tr>
<tr>
<td>Subjects with a wiring map</td>
<td>208</td>
<td>201</td>
<td>180</td>
</tr>
</tbody>
</table>

* Data were obtained from London et al. (1).
### TABLE 3. Matched and unmatched analyses of leukemia risk in relation to the arithmetic mean of 24-hour magnetic field measurements in the child's bedroom, Los Angeles County, California, 1980–1987

<table>
<thead>
<tr>
<th>Exposure category (percentile)</th>
<th>No. of cases</th>
<th>No. of controls</th>
<th>Unmatched (all pairs)</th>
<th>Matched (complete pairs)</th>
<th>Unmatched (incomplete pairs)</th>
<th>Missing indicator (all pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>OR† SE† 95% CI†</td>
<td>OR  SE  95% CI</td>
<td>OR  SE  95% CI</td>
<td>OR  SE  95% CI</td>
</tr>
<tr>
<td>0–49</td>
<td>85</td>
<td>69</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>50–74</td>
<td>35</td>
<td>42</td>
<td>0.68 0.28 0.39 1.17</td>
<td>0.65 0.38 0.31 1.37</td>
<td>0.51 0.54 0.18 1.44</td>
<td>0.61 0.31 0.34 1.12</td>
</tr>
<tr>
<td>75–89</td>
<td>24</td>
<td>22</td>
<td>0.89 0.34 0.46 1.71</td>
<td>1.10 0.48 0.43 2.80</td>
<td>0.52 0.57 0.17 1.56</td>
<td>0.81 0.36 0.40 1.66</td>
</tr>
<tr>
<td>90–100</td>
<td>20</td>
<td>11</td>
<td>1.48 0.41 0.66 3.28</td>
<td>1.62 0.46 0.66 3.97</td>
<td>0.93 0.78 0.20 4.31</td>
<td>1.41 0.40 0.65 3.05</td>
</tr>
<tr>
<td>Missing data</td>
<td>68</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p for trend</td>
<td></td>
<td></td>
<td>0.94 (continuous data)</td>
<td>0.82 (continuous data)</td>
<td>0.53 (continuous data)</td>
<td>0.96 (continuous data)</td>
</tr>
</tbody>
</table>

* Two cases and one control with highly influential observations were included in the analysis of categorical data but not in the analysis of continuous data.
† OR, odds ratio; SE, standard error of the log odds ratio; CI, confidence interval.

### TABLE 4. Matched and unmatched analyses of leukemia risk in relation to low-power magnetic field measurements in the child's bedroom, Los Angeles County, California, 1980–1987

<table>
<thead>
<tr>
<th>Exposure category (percentile)</th>
<th>No. of cases</th>
<th>No. of controls</th>
<th>Unmatched (all pairs)</th>
<th>Matched (complete pairs)</th>
<th>Unmatched (incomplete pairs)</th>
<th>Missing indicator (all pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>OR*  SE*  95% CI*</td>
<td>OR  SE  95% CI</td>
<td>OR  SE  95% CI</td>
<td>OR  SE  95% CI</td>
</tr>
<tr>
<td>0–49</td>
<td>67</td>
<td>56</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>50–74</td>
<td>34</td>
<td>28</td>
<td>1.01 0.31 0.55 1.87</td>
<td>0.97 0.46 0.39 2.38</td>
<td>1.04 0.49 0.40 2.74</td>
<td>1.00 0.34 0.52 1.93</td>
</tr>
<tr>
<td>75–89</td>
<td>23</td>
<td>14</td>
<td>1.37 0.38 0.65 2.91</td>
<td>1.24 0.50 0.47 3.29</td>
<td>1.69 0.65 0.47 6.01</td>
<td>1.39 0.39 0.65 3.00</td>
</tr>
<tr>
<td>90–100</td>
<td>16</td>
<td>11</td>
<td>1.22 0.43 0.52 2.83</td>
<td>1.09 0.61 0.33 3.64</td>
<td>1.23 0.62 0.36 4.13</td>
<td>1.16 0.43 0.50 2.72</td>
</tr>
<tr>
<td>Missing data</td>
<td>92</td>
<td>123</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p for trend</td>
<td></td>
<td></td>
<td>0.61 (continuous data)</td>
<td>0.51 (continuous data)</td>
<td>0.87 (continuous data)</td>
<td>0.72 (continuous data)</td>
</tr>
</tbody>
</table>

* OR, odds ratio; SE, standard error of the log odds ratio; CI, confidence interval.


<table>
<thead>
<tr>
<th>Exposure category</th>
<th>No. of cases</th>
<th>No. of controls</th>
<th>Unmatched (all pairs)</th>
<th>Matched (complete pairs)</th>
<th>Unmatched (incomplete pairs)</th>
<th>Missing indicator (all pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>OR† SE† 95% CI†</td>
<td>OR  SE  95% CI</td>
<td>OR  SE  95% CI</td>
<td>OR  SE  95% CI</td>
</tr>
<tr>
<td>Underground</td>
<td>11</td>
<td>11</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Very low</td>
<td>20</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ordinary low</td>
<td>64</td>
<td>87</td>
<td>0.85 0.30 0.48 1.52</td>
<td>0.95 0.31 0.51 1.76</td>
<td>0.55 1.02 0.08 3.94</td>
<td>0.90 0.30 0.50 1.62</td>
</tr>
<tr>
<td>Ordinary high</td>
<td>72</td>
<td>54</td>
<td>1.55 0.30 0.85 2.81</td>
<td>1.71 0.34 0.87 3.34</td>
<td>1.47 1.06 0.18 11.7</td>
<td>1.70 0.33 0.90 3.22</td>
</tr>
<tr>
<td>Very high</td>
<td>41</td>
<td>24</td>
<td>1.98 0.36 0.99 3.97</td>
<td>2.31 0.40 1.06 5.06</td>
<td>1.11 1.17 0.11 11.0</td>
<td>2.15 0.38 1.03 4.50</td>
</tr>
<tr>
<td>Missing data</td>
<td>24</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p for trend</td>
<td></td>
<td></td>
<td>0.006 (four categories)</td>
<td>0.005 (four categories)</td>
<td>0.50 (four categories)</td>
<td>0.003 (four categories)</td>
</tr>
</tbody>
</table>

* The numbers of cases and controls in each exposure category differ somewhat from those used by London et al. (1).
† OR, odds ratio; SE, standard error of the log odds ratio; CI, confidence interval.
‡ Reference category (underground and very low combined).

misme between the odds ratios obtained from a matched analysis of the complete pairs and an unmatched analysis of the incomplete pairs. More specifically, the log odds ratio from the missing-indicator method is a weighted average of the log odds ratios from a matched analysis of the complete pairs and an

*Am J Epidemiol* Vol. 150, No. 12, 1999
unmatched analysis of the incomplete pairs, the weighting factors being approximately equal to the proportion of the total number of observed subjects in the complete pairs and the incomplete pairs, respectively. It can also be seen in tables 3–5 that the standard error of the log odds ratio is smaller in the missing-indicator method than in either a matched analysis of the complete pairs or an unmatched analysis of the incomplete pairs, and is approximately the same as in an unmatched analysis of all of the pairs. This is to be expected, since the missing-indicator method uses the information in both the complete pairs and the incomplete pairs. Finally, in table 5, where the proportion of missing data is small, there is evidence for the possible bias introduced by breaking the matching in the complete pairs: The odds ratios obtained when the matching is broken differ from those obtained by a matched analysis of the complete pairs more than those obtained by the missing-indicator method (which preserves the matching in the complete pairs).

STRATIFICATION AND TESTS OF HETEROGENEITY

In many of the studies in which matching is broken, some components of the matching are retained by a grouped stratified analysis (6–9). For example, in a case-control study of the effect of estrogen-progestin replacement therapy on a woman’s risk of developing endometrial cancer, Pike et al. (6) matched cases and controls according to age and either neighborhood or socioeconomic status. For various reasons, 7 percent of the cases and 8 percent of the controls were excluded from the final analyses. Additionally, for 4 percent of the cases, no eligible control was found. Thus, there were cases without a matched control and controls without a matched case. When the matching was broken (to avoid losing information from the incomplete pairs), cases and controls were stratified on age and socioeconomic status.

The missing-indicator method can be applied here also. For every case missing a matched control, a “virtual” matched control is created in the same stratum as the case; and for every control missing a matched case, a “virtual” matched case is created in the same stratum as the control. The exposure covariate of every virtual subject is assigned the value \( Z = 0 \). For every subject, stratum-specific missing indicators are then defined, the indicator for the \( j \)th stratum being coded as 1 for a virtual subject in the \( j \)th stratum and as 0 otherwise. The resulting conditional logistic likelihood contribution is as in equation 8 for a complete pair and as in equation 9 or 10 but with \( \gamma \) replaced by \( \gamma_j \), the stratum-specific “intercept” term, for an incomplete pair in the \( j \)th stratum. The total likelihood is thus a product of the conditional logistic regression likelihood for the matched pairs and the unconditional logistic regression likelihood for the stratified single cases and controls. As in the case of an unstratified analysis, the missing-indicator method has the advantage of making use of the information in the incomplete pairs while still preserving the matching in the complete pairs.

The application of the missing-indicator method in either a stratified analysis or an unstratified analysis combines information from complete and incomplete pairs. Some guidance for deciding whether it is reasonable to pool the data from the complete and incomplete pairs can be obtained by testing for a difference in the odds ratios between the complete pairs and the incomplete pairs. This test can be done by defining an indicator for membership in a complete pair which has the value 0 if the exposure covariate is observed for both members of a case-control pair and the value 1 otherwise, and then testing for an interaction between this indicator and the exposure covariate. If a statistically significant interaction is found, then pooling of the data for the complete and incomplete pairs is certainly questionable. When this test is applied to the data in tables 3–5, no statistically significant interaction is found (\( p > 0.75 \)), indicating that analysis of the data by the missing-indicator method is reasonable.

DISCUSSION

In a matched case-control study with missing exposure data, a complete-case analysis ignores the information in the incomplete pairs. Whenever breaking the matching and performing an unmatched analysis (either stratified or unstratified) is considered as an alternative method of analysis, the missing-indicator method is a better alternative, since it makes use of the information in the incomplete pairs while still preserving the matching in the complete pairs, in addition to avoiding the inconvenience of parallel complete- and broken-pairs analyses.

The validity of an analysis using the missing-indicator method as presented here depends on two main criteria. It is a requirement that 1) the complete-pairs analysis is valid and 2) potential confounding in the incomplete-pairs contribution is adequately controlled for. In a recent review of methods for handling missing covariates in case-control studies, it was recommended that the missing-indicator method be avoided and more sophisticated methods be preferred (12). We believe that this is too broad an assessment. These methods come with their own set of assumptions, potential biases, and analytical difficulties. In the matched case-control setting we have considered, the missing indicator method can be understood as including unmatched case-control components in the likelihood. This makes the possible biases easy to
understand as a confounding problem that may be solved by stratifying the unmatched part. This may be done using standard analytical methods and software. For matched case-control studies, we believe that the missing-indicator method is an improvement on current common practice, has a simple interpretation, and may be appropriate in many situations.

ACKNOWLEDGMENTS

This work was supported by National Cancer Institute grant CA42949.

The authors express their appreciation to Drs. Stephanie London and John Peters for allowing the use of their case-control study data and to Drs. Duncan Thomas and Malcolm Pike for their comments on the manuscript.

REFERENCES