Missing covariate data in matched case-control studies: Do the usual paradigms apply?

Bryan Langholz

USC Department of Preventive Medicine

Joint work with
Mulugeta Gebregziabher
Larry Goldstein
Mark Huberman
CONTEXT AND SCOPE

• Missing covariate data in nested case-control studies.
• Epidemiologic individually matched case-control data.
• 1:1 (single case-single control).
• Single covariate that has missing values.
• Missing indicator methods.
## MISSING INDICATOR METHOD IN THE LITERATURE

<table>
<thead>
<tr>
<th>Reference</th>
<th>Context</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vach and Blettner (1991) <em>AJE</em> <strong>134</strong>:895-907.</td>
<td>Unmatched case-control studies (logistic)</td>
<td>“Making use of data on subjects who are neglected in complete case analysis by creating an additional category always results in biased estimation.”</td>
</tr>
<tr>
<td>Greenland and Finkle (1995) <em>AJE</em> <strong>142</strong>:1255-1264.</td>
<td>Unmatched case-control studies (logistic)</td>
<td>“The authors recommend that epidemiologists avoid using the missing-indicator method…”</td>
</tr>
<tr>
<td>Li, et al. (2004) <em>AJE</em> <strong>159</strong>:603-10.</td>
<td>Individually matched case-control</td>
<td>“…the missing-indicator method should be used cautiously.”</td>
</tr>
</tbody>
</table>
THREADS

● THREAD I: Some background: Missing data in matched pairs.

● THREAD II: Some theory: Missing data induced intensity and models.

● THREAD IV: Missing indicator methods: A bad rap?

● THREAD III: Is MCAR/MAR/NI classification “a square peg in a round hole?”
THREAD I: SOME BACKGROUND

• Motivation: Childhood leukemia study.

• Old work: “Retain complete pairs” method for matched pairs.

• Data analysis.
CASE-CONTROL STUDY OF MAGNETIC FIELDS AND CHILDHOOD LEUKEMIA

- Individually matched population based case-control study in Los Angeles County.
- Cases identified retrospectively from Los Angeles SEER Registry.
- Controls individually age matched with friends or random digit dial.
- 232 Matched pairs.
- Meters used to collect 24-hour magnetic field profile in child’s bedroom.
- Spot measurements.

\[0\text{London et al (1991) AJE 134:923-937}\]
<table>
<thead>
<tr>
<th>Information</th>
<th>Cases</th>
<th>Controls</th>
<th>Missing subjects</th>
<th>Complete matched pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interview data</td>
<td>232</td>
<td>232</td>
<td>0%</td>
<td>232</td>
</tr>
<tr>
<td>24-hour magnetic field measurements</td>
<td>164</td>
<td>144</td>
<td>34%</td>
<td>108</td>
</tr>
<tr>
<td>Spot magnetic field measurement in child’s bedroom</td>
<td>140</td>
<td>109</td>
<td>46%</td>
<td>71</td>
</tr>
</tbody>
</table>
REASONS FOR MISSING DATA

• Most because the family had moved “relevant” home so it was not possible to measure fields.

• Instrument problems.

• Some refusal.
COMPLETE CASE ANALYSIS

• Drop subjects with missing data.
• Analyze as matched case-control study.

Note that:

• Pair is dropped if *either* case or control is missing.

Large loss of data.
BREAK THE MATCHING

- Drop subjects with missing data.
- Break the matching.
- Analyze as unmatched case-control study, control for confounding by matching factors by modelling.

Note that:

- Can model age but
- Quantification of “friend” or “telephone number” matching not possible.

Potential residual confounding.
Compromise between complete case and break the matching:

- Retain matching for those pairs in which both subjects have covariate data.
- Break matching for subjects with data in pairs with “missing data partner.”

Avoids large loss of data (complete case analysis) and minimizes potential uncontrolled confounding (break the matching).
ANALYSIS

- Matched pairs part contributes conditional logistic likelihood part.

- Unmatched part contributes unconditional logistic likelihood part (perhaps control for confounding by modelling matching variables).

Estimation based on product of conditional and unconditional logistic likelihood parts.
SIMULATION STUDY PARAMETERS

• Three age groups - 50%, 40%, 10% youngest to oldest.

• Exposure - $X_i \sim \mathcal{N}(\mu_s, 1)$ with $\mu_s$ - stratum specific mean.

• $\mu_s = \mu_{age} + U(-.5, .5)$, $\mu_{age}$ = age = 0, 1, or 2.

• Age - quantifiable matching factor, the $U(-.5, .5)$ for unquantifiable.

• Independent “risk sets” of size 50 were generated for 400 cases.
• Disease status - prob based on rate ratios, multiplicative in age and exposure with 
  \( rr_{age} = 2, rr_X = \sqrt{2}, \beta_X = .35. \)

• A single individual was randomly sampled from the 49 in the risk set to serve as the control in the matched pair set.

• 25% missing \( X \) completely at random: 56% pairs had complete data.

• Results based on 500 trials.
## SIMULATION RESULTS

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean parameter</th>
<th>Empirical s.e.</th>
<th>Estimated s.e.</th>
<th>$\sqrt{MSE}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No missing</td>
<td>0.35</td>
<td>0.080</td>
<td>0.075</td>
<td>0.075</td>
</tr>
<tr>
<td>Complete case</td>
<td>0.35</td>
<td>0.098</td>
<td>0.101</td>
<td>0.101</td>
</tr>
<tr>
<td>Break the matching:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pooled over age</td>
<td>0.19</td>
<td>0.052</td>
<td>0.061</td>
<td>0.171</td>
</tr>
<tr>
<td>Age stratified</td>
<td>0.26</td>
<td>0.065</td>
<td>0.073</td>
<td>0.114</td>
</tr>
<tr>
<td>Retain matching in complete pairs:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pooled over age</td>
<td>0.29</td>
<td>0.078</td>
<td>0.077</td>
<td>0.099</td>
</tr>
<tr>
<td>Age-stratified</td>
<td>0.32</td>
<td>0.082</td>
<td>0.083</td>
<td>0.087</td>
</tr>
<tr>
<td>Age-trend</td>
<td>0.31</td>
<td>0.081</td>
<td>0.082</td>
<td>0.088</td>
</tr>
</tbody>
</table>
“RETAIN THE MATCHING” METHOD

- Efficiency is as least as good as complete-case analysis.

- Bias is never as bad as breaking the matching.

Good compromise to the two most commonly used alternatives.
Missing indicator analysis is exactly the “retain the matching in complete pairs” analysis.

\(^0\)Huberman and Langholz (1999) *AJE* 150:1340-1345.
<table>
<thead>
<tr>
<th>C-C Set</th>
<th>Case(1) control(0)</th>
<th>Age group</th>
<th>Exposure (0 if missing)</th>
<th>Z</th>
<th>Missing indicator</th>
<th>Age=1</th>
<th>Age=2</th>
<th>Age=3</th>
<th>Age-trend Mtrend</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>1</td>
<td>3</td>
<td>2.99688</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>38</td>
<td>0</td>
<td>3</td>
<td>0.00000</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>39</td>
<td>1</td>
<td>3</td>
<td>2.91090</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>39</td>
<td>0</td>
<td>3</td>
<td>0.00000</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>40</td>
<td>1</td>
<td>2</td>
<td>0.92814</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>0</td>
<td>2</td>
<td>0.00000</td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>41</td>
<td>1</td>
<td>1</td>
<td>0.62488</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>41</td>
<td>0</td>
<td>1</td>
<td>0.90149</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>42</td>
<td>1</td>
<td>1</td>
<td>0.51196</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>42</td>
<td>0</td>
<td>1</td>
<td>-0.76174</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>1</td>
<td>3</td>
<td>2.48709</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>0</td>
<td>3</td>
<td>0.00000</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>44</td>
<td>1</td>
<td>3</td>
<td>2.65434</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>44</td>
<td>0</td>
<td>3</td>
<td>0.00000</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>45</td>
<td>1</td>
<td>2</td>
<td>0.00000</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>0</td>
<td>2</td>
<td>0.00000</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>46</td>
<td>1</td>
<td>1</td>
<td>0.00000</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>46</td>
<td>0</td>
<td>1</td>
<td>0.00000</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
THREAD II: SOME THEORY

• Develop a theory for *predictable* missingness (doesn’t depend on case-control status).
  – Consider complete data cohort intensity (rates) under proportional hazards.
  – Use the double expectation formula to derive general form of missing data “induced” intensity.
  – Extend the proportional hazards model to encompass the possible form(s) of the induced intensity.
  – Propose a semi-parametric version of this model.
  – Fitting the model.
  – Complete case analysis.

• Adapted (depends on case-control status) missingness.
KEY ASSUMPTION FOR THIS PART

• I assume that missingness is *predictable* - does not depend on case-control status.
Key point (Ørnulf’s talk):

- If the true intensity is a member of the specified proportional hazards model, then the partial likelihood(s) provide valid estimation of $\beta_0$.

- Works for both Cox (for full cohort) and conditional logistic (for individually matched case-control data) regression.
Let \((N_i, Y_i, Z_i)\) counting, censoring, and covariate processes \((i.i.d.)\).

- \(Z\) - corresponding filtration.

- Then the “complete data” \(Z\)-intensity is

\[
\lambda_i(t; Z) dt = E[dN_i(t) | Z_{t-}] 
\]

• Semi-parametric model for the rate of disease as a function of factors (e.g., radiation, genes).

\[ \lambda(t, y, z; \alpha(\cdot), \beta) = y \alpha(t) r(z; \beta). \]

  - \( \alpha \) - rate of disease in “unexposed” as a function of time (nuisance function).
  - \( r \) - “rate ratio” as a function of \( z \).
  - \( \beta \) - rate ratio parameters to be estimated.

• Assume that there exist \( \lambda_0, \beta_0 \) such that

\[ \lambda_i(t; \mathcal{Z}) = Y_i(t) \lambda_0(t) r(Z_i(t); \beta_0). \]
DATA MODEL FOR (PREDICTABLE) MISSING DATA

- $Z_i(t)$ - covariate with no missing data.
- $X_i(t)$ - single covariate with missing data.
- $M_i(t)$ - missing indicator, independent and left continuous.
- Representation for missing covariate data $(N_i, Y_i, (1 - M_i)X_i, M_i, Z_i)$.
- $\mathcal{M}$-filtration generated by $
\{N_i(t), Y_i(t), [1 - M_i(t)]X_i(t), M_i(t), Z_i(t)\}$.  

Note: $M_i$ is $\mathcal{M}$-predictable.
The “Innovation Theorem” says that the $\mathcal{M}$ (missing data)-intensity

$$\lambda_i(t; \mathcal{M}) dt = E[dN_i(t)|\mathcal{M}_{t-}] = E[\lambda_i(t; \mathcal{Z})|\mathcal{M}_{t-}] dt$$

(Double expectation formula for processes)

Proportional hazards:

$$\lambda_i(t; \mathcal{M}) = Y_i(t)\lambda_0(t)E[r(X_i, Z_i(t); \beta_0)|\mathcal{M}_{t-}]$$
MISSING DATA INDUCED INTENSITY

Assumptions:

- Distribution of $X$ only depends on summaries of other factors at $t$.
- Recall i.i.d.

Define:

- $F_X(\cdot; z, t; \beta_0, \alpha_0)$ - distribution function for $X$ conditional on $Y(t) = 1$, $M(t) = 1$, and $Z(t) = z$. 
MISSING DATA INDUCED INTENSITY

\[ \lambda_i(t; \mathcal{M}) = Y_i(t) \lambda_0(t) r(X_i(t), Z_i(t); \beta_0)^{1-M_i(t)} \]

\[ \times \eta[t, Z_i(t); \beta_0, \lambda_0, F_X(\cdot; t, Z_i(t); \beta_0, \lambda_0)]^{M_i(t)} \]

where

\[ \eta[t, z; \beta_0, \lambda_0(\cdot), F_X(\cdot; t, z; \beta_0, \lambda_0)] = \]

\[ E[r(X, z; \beta_0)|Y(t) = 1, M(t) = 1] \]

\[ = \int r(x, z; \beta_0)dF_X(x; t, z; \beta_0, \lambda_0). \]
MISSING DATA INDUCED MODEL

• We will need to expand the complete data model with a component to accommodate $\eta$.

• In (very general) the model would look like:

$$\lambda(t, y, (x, z), m; \alpha, \beta) =$$

$$y \alpha(t) \ r(x, z; \beta)^{1-m} \ \eta[t, z; \beta, \alpha, F_\psi(\cdot; t, z; \beta, \alpha)]^m$$

where

$$\eta(t, z; \beta, \alpha) = \int r(x, z; \beta) \ dF_\psi(x; t, z; \beta, \alpha)$$

and $F_\psi$ is a family of distribution functions.
Modelling the missingness is modelling $\eta$, the “expected rate ratio.”

$\eta(t, z; \beta, \alpha, F_\psi(x, z, t; \beta, \alpha))$ depends $\beta$, $\alpha$ and $F_\psi$.

Missing data methods for failure time data can be seen as different ways of structuring $\eta$.

Prentice (1982)\textsuperscript{1} assumes a distributional form on $X|Z \sim N(\mu_Z, \sigma^2)$ and take expectation as function of $\beta$.

A SEMI-PARAMETRIC APPROACH

Model that encompasses the “missing data” induced intensity under proportional hazards:

\[ \lambda(t, y, (x, z), m; \alpha, \beta) = y \alpha(t) r(x, z; \beta)^{1-m} \eta(t, z; \gamma)^m. \]

where \( \eta(t, z; \gamma) \) corresponds to the “expected rate ratio” \( E[r(X, Z; \beta)|Y = 1, M = 1, Z = z)]. \)

- Since the model is still proportional hazards, standard partial likelihood methods are valid for both cohort and case-control data.
- Need to capture the variation in expected rate ratio \( \eta \) with \( t \) and (known) modelled covariates \( Z \).
SOME OBSERVATIONS

• “Non-parametric” with respect to distribution of $X$ (just like complete data model).

• Probably don’t need to model $\eta$ too well as it is a nuisance parameter in the model.
“Missing indicator” method.
COMPLETE CASE ANALYSIS

Theorem: If \( \eta(t, z; \gamma) = \gamma(t, z) \) ("unstructured") then, the partial likelihood estimator is the complete case analysis.

(Pseudo) Proof:

\[
y \alpha(t) r(x, z; \beta)^{1-m} \gamma(t, z)^m = \begin{cases} 
y \alpha(t) r(x, z; \beta) & \text{if } m = 0 \\
y \tilde{\gamma}(t, z) & \text{if } m = 1
\end{cases}
\]

where \( \tilde{\gamma}(t, z) = \alpha(t)\gamma(t, z) \) captures the disease rate in missing.

- "Missing data" stratified model.
- PL contribution from subjects with missing data is constant, so "dropped" from analysis.
- Complete case analysis.
CONSEQUENCES OF THE THEOREM

- Complete case analysis is “infinite parameter” missing indicator-$t-Z$ interaction model.

- Complete case analysis is the non-parametric-limiting-case of missing indicator modelling.

- Spectrum:
  - Single parameter - most bias, least variance.
  - Complete case - least bias, most variance.

Model missingness with “enough but not too many” parameters.
SIMULATION STUDY: PREDICTABLE MISSINGNESS

- **X** - Dichotomous, can be missing (RR=2).
- **Z** - Three levels, no missing (RR=4/level).
- 1:1 matched data (as before), matching factor is not a confounder.
- **X**-missingness dependence (predictable): none, **Z**, **X**, **Z** − **X**.

**Missing data methods:**
- Complete case.
- **Z**-specific missing indicators (right model).

**Results based on 500 trials.**
SIMULATION RESULTS: $X$ MISSING IN 50%:

BIAS

<table>
<thead>
<tr>
<th>Missing $X$ depend.</th>
<th>$E\hat{\beta}_X$ Complete case</th>
<th>$E\hat{\beta}_X$ Missing indicators</th>
<th>$E\hat{\beta}_Z$ Complete case</th>
<th>$E\hat{\beta}_Z$ Missing indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>0.72</td>
<td>0.71</td>
<td>1.44</td>
<td>1.41</td>
</tr>
<tr>
<td>$Z$</td>
<td>0.72</td>
<td>0.69</td>
<td>1.44</td>
<td>1.42</td>
</tr>
<tr>
<td>$X$</td>
<td>0.73</td>
<td>0.71</td>
<td>1.44</td>
<td>1.42</td>
</tr>
<tr>
<td>$Z - X$</td>
<td>0.71</td>
<td>0.70</td>
<td>1.43</td>
<td>1.42</td>
</tr>
</tbody>
</table>

\(^1\)No missing: $E\hat{\beta}_X = 0.69$, $E\hat{\beta}_Z = 1.40$
SIMULATION RESULTS: $X$ MISSING IN 50% VARIANCE

<table>
<thead>
<tr>
<th>Missing</th>
<th>$\text{var } \hat{\beta}_X^{a}$</th>
<th>$\text{var } \hat{\beta}_Z^{a}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>Complete case: 0.138</td>
<td>Missing indicators: 0.064</td>
</tr>
<tr>
<td>$Z$</td>
<td>Normal: 0.121</td>
<td>Missing indicators: 0.062</td>
</tr>
<tr>
<td>$X$</td>
<td>Normal: 0.129</td>
<td>Missing indicators: 0.064</td>
</tr>
<tr>
<td>$Z-X$</td>
<td>Normal: 0.105</td>
<td>Missing indicators: 0.061</td>
</tr>
</tbody>
</table>

$^a$Variance well estimated by usual inverse information estimator.
ADAPTED MISSINGNESS

Definition:

- Missingness depends on case-control status.
### ADAPTED MISSINGNESS SIMULATION RESULTS: $X$ MISSING IN 50%:

<table>
<thead>
<tr>
<th>Missing $X$ depend.</th>
<th>$E\hat{\beta}_X$</th>
<th>$E\hat{\beta}_Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Complete case</td>
<td>Missing indicators</td>
</tr>
<tr>
<td>$D$</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>$D - Z$</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>$D - X$</td>
<td>-0.26</td>
<td>-0.24</td>
</tr>
</tbody>
</table>
THEORY FOR ADAPTED MISSINGNESS: SECOND STAGE SAMPLING

- $r_1$ - case-control set.
- $r_2$ - subjects in the case-control set with complete data ("2nd stage sample").
- $N_{i,r_1,r_2}$ - counts $i,r_1,r_2$ "events."
- $\pi(r_2|i,r_1)$ - probability of complete data ($r_2$) given case-control status.

Proportional hazards:

$$\lambda_{i,r_1,r_2}(t) = Y_i \lambda_0(t) r(X_i, Z_i; \beta_0) \pi(r_1|i) \pi(r_2|i, r_1)$$
CONJECTURE: PREDICTABLE VS. ADAPTED MISSINGNESS

- Predictable (missingness does not depend on case-control status) - can estimate all parameters.
- Adapted (missing depends on case-control status) - can’t estimate all parameters. ²

Conjecture:
To estimate all parameters in the model predictability is sufficient and necessary.

²D dependent missingness: can’t estimate \( \lambda_0 \).
THREAD III: HAVE MISSING INDICATOR METHODS GOTTEN A BAD RAP?
SO, IS EVERYONE ELSE WRONG ABOUT MISSING INDICATORS?

Two considerations:

- Lack of understanding that one needs to “modelling the missingness:” model the variation in expected rate ratio.
- Something special about individually matched case-control setting.
<table>
<thead>
<tr>
<th>Method</th>
<th>Mean parameter</th>
<th>Empirical s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>No missing</td>
<td>0.34</td>
<td>0.05</td>
</tr>
<tr>
<td>Complete case</td>
<td>0.34</td>
<td>0.11</td>
</tr>
<tr>
<td>Missing indicator models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single MI</td>
<td>0.24</td>
<td>0.09</td>
</tr>
<tr>
<td>$Z$-MI interaction</td>
<td>0.34</td>
<td>0.11</td>
</tr>
<tr>
<td>$Z$-MI trend</td>
<td>0.32</td>
<td>0.11</td>
</tr>
</tbody>
</table>

$^2$80% missing, MCAR
Likely Situation: We’re All “Right”

- In other settings, there is less (no?!?) efficiency gain using missing indicator?
  - Survival analysis.
  - Single stratum unmatched case-control studies.

Need to evaluate each setting.
THREAD IV: “CLASSICAL”
MCAR/MAR/NI CLASSIFICATION OF
MISSINGNESS

Piak and Sacco (2000) Regression calibration (imputation)
Piak (2004) Regression calibration (imputation)

All assume MAR.
## SIMULATION RESULTS: $X$ MISSING IN 50%: BIAS

<table>
<thead>
<tr>
<th>Missing $X$ depend.</th>
<th>Miss type</th>
<th>$E\hat{\beta}_X$</th>
<th>$E\hat{\beta}_Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Comp case</td>
<td>Miss ind</td>
</tr>
<tr>
<td></td>
<td>Predictable missingness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>MCAR</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>$Z$</td>
<td>MAR</td>
<td>0.72</td>
<td>0.69</td>
</tr>
<tr>
<td>$X$</td>
<td>NI</td>
<td>0.73</td>
<td>0.71</td>
</tr>
<tr>
<td>$Z - X$</td>
<td>NI</td>
<td>0.71</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Adapted missingness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D$</td>
<td>MAR</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>$D - Z$</td>
<td>MAR</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>$D - X$</td>
<td>NI</td>
<td>-0.26</td>
<td>-0.24</td>
</tr>
</tbody>
</table>
CLASSICAL CLASSIFICATIONS: UNBIASEDNESS OF COMPLETE CASE OR MISSING INDICATOR

- MCAR OK, but too strong (sufficient but not necessary).
- MAR can be unbiased (e.g., MAR($Z$)) or biased (e.g., MAR($D - Z$)).
- NI can be unbiased (e.g., NI($X$)) or biased (e.g., NI($D - X$)).

Not very useful for predicting validity of methods.
IS MCAR/MAR REASONABLE?

Even if asked prior to disease occurrence, missing the following are likely to be dependent on the actual value (NI):

- Did you smoke during pregnancy?
- What is your current income?
- Radiation dose as abstracted from medical records.

Better not to make a MCAR/MAR assumption.
MULTIPLE IMPUTATION APPLIED TO MATCHED PAIRS SIMULATION STUDY

- $X$ and $Z$ covariates, $X$ with missing.
- Used multiple imputation (SAS PROCs MI, PHREG, and MIANALYZE).
- Imputation model has case-control status and $Z$.
- Only looked at performance for estimating $\beta_X = .69$. 
## Missing Indicator and Multiple Imputation

<table>
<thead>
<tr>
<th>Missing depend.</th>
<th>Miss type</th>
<th>Missing indicators</th>
<th>Multiple imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>MCAR</td>
<td>$\hat{\beta}_X$ 0.71, var$\hat{\beta}_X$ 0.26</td>
<td>$\hat{\beta}_X$ 0.70, var$\hat{\beta}_X$ 0.27</td>
</tr>
<tr>
<td>$Z$</td>
<td>MAR</td>
<td>$\hat{\beta}_X$ 0.71, var$\hat{\beta}_X$ 0.27</td>
<td>$\hat{\beta}_X$ 0.71, var$\hat{\beta}_X$ 0.25</td>
</tr>
<tr>
<td>$X$</td>
<td>NI</td>
<td>$\hat{\beta}_X$ 0.70, var$\hat{\beta}_X$ 0.27</td>
<td>$\hat{\beta}_X$ 0.70, var$\hat{\beta}_X$ 0.28</td>
</tr>
<tr>
<td>$Z - X$</td>
<td>NI</td>
<td>$\hat{\beta}_X$ 0.70, var$\hat{\beta}_X$ 0.26</td>
<td>$\hat{\beta}_X$ 0.69, var$\hat{\beta}_X$ 0.28</td>
</tr>
<tr>
<td>$D$</td>
<td>MAR</td>
<td>$\hat{\beta}_X$ 0.70, var$\hat{\beta}_X$ 0.27</td>
<td>$\hat{\beta}_X$ 0.72, var$\hat{\beta}_X$ 0.30</td>
</tr>
<tr>
<td>$D - Z$</td>
<td>MAR</td>
<td>$\hat{\beta}_X$ 0.74, var$\hat{\beta}_X$ 0.27</td>
<td>$\hat{\beta}_X$ 0.70, var$\hat{\beta}_X$ 0.28</td>
</tr>
<tr>
<td>$D - X$</td>
<td>NI</td>
<td>$\hat{\beta}_X$ 1.06, var$\hat{\beta}_X$ –</td>
<td>$\hat{\beta}_X$ 2.04, var$\hat{\beta}_X$ –</td>
</tr>
</tbody>
</table>
Comparison:

- Performance, including precision, very similar to missing indicators.

- Both methods require “modelling missingness.”

Classical classification and multiple imputation:

- MCAR/MAR/NI *not relevant.*
MISSING INDICATOR FOR INDIVIDUALLY MATCHED CASE-CONTROL STUDIES

- Standard conditional logistic software using standard modelling techniques.

- Predictability is key assumption for valid estimation.

- Need to model the variation in the expected rate ratio (missing indicator-covariate interactions).
GENERAL COMMENTS

- Classical classification of missingness was not very useful in assessing estimator behavior nor guiding the method of analysis: Some research to be done here?

- Predictable vs. adapted classifies well and is closely associated with the concept of information bias.

- Adding MAR or MCAR assumptions about missing type may yield estimators with better efficiency.

- MAR and MCAR should be evaluated carefully.