Statistical Approaches for Imperfect Data in Public Health Studies

Hui Xie, PhD
huixie@uic.edu

Associate Professor
Division of Epidemiology & Biostatistics
Center For Biostatistical Development
School of Public Health
University of Illinois, Chicago
How Imperfect/Missing Data Arise

- Experimental Studies
  - Accidents in experiments
    - e.g. Flood in Agriculture
How Missing Data Arise

- Experimental Studies
  - Accidents in experiments
    - *e.g.* Flood in Agriculture.
    - Irrelevant as I am careful and no flood in Chicago. Great!
How Missing Data Arise

• Experimental Studies
  • Accidents in experiments
    • *e.g.* Flood in Agriculture.
    • Irrelevant as I am careful and no flood in Chicago. Great!

• Patient noncompliance in Clinical Trials:
  • Dropout, missing visits (unavoidable)
Example 1: SWOG 9039 QoL Data

- Southwest Oncology Group (SWOG) study 9039 is a QOL substudy to a RCT of hormone treatment in prostate cancer patients (Eisenberger et al. 1998, Xie and Heitjan 2009).
Example 1: SWOG 9039 QoL Data

- Southwest Oncology Group (SWOG) study 9039 is a QOL substudy to a RCT of hormone treatment in prostate cancer patients (Eisenberger et al. 1998, Xie and Heitjan 2009).
- Flutamide (n=370) vs. Placebo (n=367).
Example 1: SWOG 9039 QoL Data

- Southwest Oncology Group (SWOG) study 9039 is a QOL substudy to a RCT of hormone treatment in prostate cancer patients (Eisenberger et al. 1998, Xie and Heitjan 2009).
- Flutamide (n=370) vs. Placebo (n=367).
- Outcome: SF36 Emotional Functioning (EF).
Example 1: SWOG 9039 QoL Data

- Southwest Oncology Group (SWOG) study 9039 is a QOL substudy to a RCT of hormone treatment in prostate cancer patients (Eisenberger et al. 1998, Xie and Heitjan 2009).
- Flutamide (n=370) vs. Placebo (n=367).
- Outcome: SF36 Emotional Functioning (EF).
- Covariates: baseline performance score and disease severity.
Example 1: SWOG 9039 QoL Data

- Southwest Oncology Group (SWOG) study 9039 is a QOL substudy to a RCT of hormone treatment in prostate cancer patients (Eisenberger et al. 1998, Xie and Heitjan 2009).
- Flutamide (n=370) vs. Placebo (n=367).
- Outcome: SF36 Emotional Functioning (EF).
- Covariates: baseline performance score and disease severity.
- Longitudinal Study: EF scale measured at 0, 1, 3, and 6 month.
Example 1: SWOG 9039 QoL Data

- Southwest Oncology Group (SWOG) study 9039 is a QOL substudy to a RCT of hormone treatment in prostate cancer patients (Eisenberger et al. 1998, Xie and Heitjan 2009).
- Flutamide (n=370) vs. Placebo (n=367).
- Outcome: SF36 Emotional Functioning (EF).
- Covariates: baseline performance score and disease severity.
- Longitudinal Study: EF scale measured at 0, 1, 3, and 6 month.
- Drop-out began at week 1; Dropout rate: 23.7% (Placebo) and 20.5% (Flutamide).
## A sample of SWOG data

<table>
<thead>
<tr>
<th>SubID</th>
<th>Group</th>
<th>$Y_0$</th>
<th>$Y_1$</th>
<th>$Y_3$</th>
<th>$Y_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>126056</td>
<td>Flutamide</td>
<td>9.2</td>
<td>9.4</td>
<td>9.6</td>
<td>9.8</td>
</tr>
<tr>
<td>127046</td>
<td>Placebo</td>
<td>9.4</td>
<td>9.4</td>
<td>9.6</td>
<td>8.9</td>
</tr>
<tr>
<td>144518</td>
<td>Flutamide</td>
<td>9.6</td>
<td>*</td>
<td>9.2</td>
<td>*</td>
</tr>
<tr>
<td>144615</td>
<td>Placebo</td>
<td>7.5</td>
<td>7.7</td>
<td>7.7</td>
<td>*</td>
</tr>
<tr>
<td>130627</td>
<td>Flutamide</td>
<td>8.2</td>
<td>8.7</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>132116</td>
<td>Placebo</td>
<td>5.3</td>
<td>9.6</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>130591</td>
<td>Flutamide</td>
<td>8.9</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>130473</td>
<td>Placebo</td>
<td>9.2</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
## Missing Data Pattern

<table>
<thead>
<tr>
<th>Pattern</th>
<th>$Y_0$</th>
<th>$Y_1$</th>
<th>$Y_3$</th>
<th>$Y_6$</th>
<th>n</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completer</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>497</td>
<td>67.44%</td>
</tr>
<tr>
<td>Dropout</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>.</td>
<td>75</td>
<td>10.18%</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>.</td>
<td>.</td>
<td>33</td>
<td>4.48%</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>42</td>
<td>5.70%</td>
</tr>
<tr>
<td></td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>12</td>
<td>1.63%</td>
</tr>
<tr>
<td>Intermittent</td>
<td>X</td>
<td>X</td>
<td>.</td>
<td>X</td>
<td>26</td>
<td>3.53%</td>
</tr>
<tr>
<td>Missing</td>
<td>X</td>
<td>.</td>
<td>X</td>
<td>X</td>
<td>26</td>
<td>3.53%</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>.</td>
<td>X</td>
<td>.</td>
<td>8</td>
<td>1.09%</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>.</td>
<td>.</td>
<td>X</td>
<td>8</td>
<td>1.09%</td>
</tr>
<tr>
<td></td>
<td>.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>7</td>
<td>0.95%</td>
</tr>
<tr>
<td></td>
<td>.</td>
<td>X</td>
<td>.</td>
<td>.</td>
<td>2</td>
<td>0.27%</td>
</tr>
<tr>
<td></td>
<td>. .</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
<td>0.14%</td>
</tr>
</tbody>
</table>
How Missing Data Arise

- Experimental Studies
  - Patient noncompliance in Clinical Trials:
    - Dropout, missing visits
    - Self-select treatment after randomization
Example 2: The Multiple Sclerosis Trial

- An RCT of immunosuppressive therapy in the treatment of multiple sclerosis (Xie and Heitjan 2004).
Example 2: The Multiple Sclerosis Trial

- An RCT of immunosuppressive therapy in the treatment of multiple sclerosis (Xie and Heitjan 2004).
- Outcome: AD25 (Antibody-Dependent cellular cytotoxicity at E:T ratio 25:1).
Example 2: The Multiple Sclerosis Trial

- An RCT of immunosuppressive therapy in the treatment of multiple sclerosis (Xie and Heitjan 2004).
- Outcome: AD25 (Antibody-Dependent cellular cytotoxicity at E:T ratio 25:1).
- Two randomization groups: 0=double placebo (n=14); 1=azathioprine plus methylprednisolone (n=11).
Example 2: The Multiple Sclerosis Trial

- An RCT of immunosuppressive therapy in the treatment of multiple sclerosis (Xie and Heitjan 2004).
- Outcome: AD25 (Antibody-Dependent cellular cytotoxicity at E:T ratio 25:1).
- Two randomization groups: 0=double placebo (n=14); 1=azathioprine plus methylprednisolone (n=11).
- An effective therapy will reduce AD25 level.
Example 2: The Multiple Sclerosis Trial

- An RCT of immunosuppressive therapy in the treatment of multiple sclerosis (Xie and Heitjan 2004).
- Outcome: AD25 (Antibody-Dependent cellular cytotoxicity at E:T ratio 25:1).
- Two randomization groups: 0=double placebo (n=14); 1=azathioprine plus methylprednisolone (n=11).
- An effective therapy will reduce AD25 level.
- Noncompliance led to crossover. Treatment Group (3/11, 27%) crossover to Placebo group.
Table 1: AD25 Level and Compliance at 3 years

<table>
<thead>
<tr>
<th>Assigned to 0</th>
<th>Assigned to 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>21</td>
</tr>
<tr>
<td>18</td>
<td>25</td>
</tr>
<tr>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td>31</td>
<td>33 (after switching to group 0)</td>
</tr>
<tr>
<td>31</td>
<td>38 (after switching to group 0)</td>
</tr>
<tr>
<td>33</td>
<td>55 (after switching to group 0)</td>
</tr>
<tr>
<td>34</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td></td>
</tr>
</tbody>
</table>
How Missing Data Arise

• Experimental Studies
  • Patient noncompliance in Clinical Trials:
    • Dropout, missing visits
    • Self-select treatment after randomization
  • See *The prevention and treatment of missing data in clinical trials (2010).*
    —Committee on National Statistics for National Research Council
How Missing Data Arise

• Experimental Studies
  • Patient noncompliance in Clinical Trials:
    • Dropout, missing visits
    • Self-select treatment after randomization
  • See *The prevention and treatment of missing data in clinical trials (2010)*.
    —Committee on National Statistics for National Research Council

• Observational Studies
  • Survey (item or unit) nonresponse (e.g., Patient satisfaction, mHealth).
  • Missing outcome and/or predictor values in organization databases.
Example 3: Hip Fracture Data

- Matched Case-control study (CXQ 2011).
- Only 237 out of 436 (54%) have complete record.
Example 4: Scanner Panel Purchase Data

<table>
<thead>
<tr>
<th>ConsumerID</th>
<th>Day</th>
<th>Price</th>
<th>Coupon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9185</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>9213</td>
<td>1.19</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>9679</td>
<td>0.99</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>9918</td>
<td>1.39</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>9930</td>
<td>.</td>
<td>1.19</td>
</tr>
</tbody>
</table>

- Real-time data capture.
- But price and coupon values for unpurchased products are missing in database. **Missing values systematically different from observed ones.**
- Can cause substantial attenuation bias in price and promotion sensitivity demand estimates (QX 2011)
How Missing Data Arise

- **Experimental Studies**
  - Patient noncompliance in Clinical Trials:
    - Dropout, missing visits
    - Self-select treatment after randomization
  - See *The prevention and treatment of missing data in clinical trials (2010).*
    —Committee on National Statistics for National Research Council

- **Observational Studies**
  - Survey (item or unit) nonresponse (e.g., Patient satisfaction).
  - Missing outcome and/or predictor values in organization databases.
  - See *Methodological Standards in the Prevention and Handling of Missing Data (2014).*
    —Patient-Centered Outcomes Research Institute (PCORI)
How Missing Data Arise

- **Experimental Studies**
  - Patient noncompliance in Clinical Trials:
    - Dropout, missing visits
    - Self-select treatment after randomization

- **Observational Studies**
  - Survey (item or unit) nonresponse.
  - Missing outcome and/or predictor values in organization databases.

- **Diagnostic Medicine**
  - Unobserved gold-standard diagnostic outcome (how to define health outcome?)
  - E.g., Rheumatoid Arthritis
How Missing Data Arise

- Conceptual Missingness (“soft missing”)

Diagram showing four scenarios:

(a) Y<br>B
(a) Y<br>B
(a) Y<br>B
(a) Y<br>B

(b) Y<br>B
(b) Y<br>B
(b) Y<br>B
(b) Y<br>B

(c) Y<br>B
(c) Y<br>B
(c) Y<br>B
(c) Y<br>B

(d) Y<br>B
(d) Y<br>B
(d) Y<br>B
(d) Y<br>B
How Missing Data Arise

• Conceptual Missingness ("soft missing")
How Missing Data Arise

- Conceptual Missingness (“soft missing”)
- Causal Inference: Counterfactuals from comparative treatments or programs are never jointly observed.
How Missing Data Arise

• Conceptual Missingness ("soft missing")
  • Causal Inference: Counterfactuals from comparative treatments or programs are never jointly observed.
  • Combine complementary data: Medical Expenditure and disease occurrence are observed in two independent patient samples.
How Missing Data Arise

- Conceptual Missingness ("soft missing")
  - Causal Inference: Counterfactuals from comparative treatments or programs are never jointly observed.
  - Combine complementary data: Medical Expenditure and disease occurrence are observed in two independent patient samples.
  - Data privacy: Sensitive health data cannot be shared (HIPAA in US and Guidelines of CIHR in Canada).
Why Care

- Statistical Program can’t run with missing data.
  Sometimes easy to fix
Why Care

- Statistical Program can’t run with missing data. Sometimes easy to fix.
- Inefficient: greatly reduced sample size.
Why Care

- Statistical Program can’t run with missing data. Sometimes easy to fix
- Inefficient: greatly reduced sample size.
- Biased Inference: Missing data systematically different from observed data
Why Care

- Statistical Program can’t run with missing data.
  Sometimes easy to fix
- Inefficient:
  greatly reduced sample size.
- Biased Inference:
  Missing data systematically different from observed data
- Unusable:
  Completely missing key variables.
  No Complete Cases.
Why Care

• Statistical Program can’t run with missing data.
  Sometimes easy to fix

• Inefficient:
  greatly reduced sample size.

• Biased Inference:
  Missing data systematically different from observed data

• Unusable:
  Completely missing key variables.
  No Complete Cases.

• Why Methods Matter for Healthcare Policy and Outcome Research?:
  See Methodology Report for missing-data related issues from Patient-Centered Outcome Research Institute.
My Research Interest

Developing Statistical Methods for Imperfect Data (Missing Data, Censoring, Endogeneity), Analyzing Big and Intensive Data, and Bayesian Methods.

My Research Interest

Developing Statistical Methods for Imperfect Data (Missing Data, Censoring, Endogeneity), Analyzing Big and Intensive Data, and Bayesian Methods.

- Nonparametric Imputation or Direct Estimation Methods for Imperfect Data.
My Research Interest

Developing Statistical Methods for Imperfect Data (Missing Data, Censoring, Endogeneity), Analyzing Big and Intensive Data, and Bayesian Methods.

• Tractable Statistical Adjustment Methods for Nonrandom Missing Data.
• Nonparametric Imputation or Direct Estimation Methods for Imperfect Data.
• Large-scale Data Integration and Disclosure Control Methods to Increase Data Accessibility and Usability.
My Research Interest

Developing Statistical Methods for Imperfect Data (Missing Data, Censoring, Endogeneity), Analyzing Big and Intensive Data, and Bayesian Methods.

- Nonparametric Imputation or Direct Estimation Methods for Imperfect Data.
- Large-scale Data Integration and Disclosure Control Methods to Increase Data Accessibility and Usability.
- **GOAL**: Improve the reliability, validity and usability of data for evidence-based public health studies.
Overview of Methods for Imperfect Data

- Ad-hoc Method: Complete-Cases analysis
Overview of Methods for Imperfect Data

- Ad-hoc Method: Complete-Cases analysis
- Weighting Method
Overview of Methods for Imperfect Data

- Ad-hoc Method: Complete-Cases analysis
- Weighting Method
- Imputation Method
Overview of Methods for Imperfect Data

- Ad-hoc Method: Complete-Cases analysis
- Weighting Method
- Imputation Method
- Likelihood-Based Method
Overview of Methods for Imperfect Data

- Ad-hoc Method: Complete-Cases analysis
- Weighting Method
- Imputation Method
- Likelihood-Based Method
- Statistical Matching
Overview of Methods for Imperfect Data

- Ad-hoc Method: Complete-Cases analysis
- Weighting Method
- Imputation Method
- Likelihood-Based Method
- Statistical Matching
- Perturbation, MI and other Methods for disclosure control
Missing Data Mechanism

- Let $Y_O$ and $Y_M$ denote the missing and observed items in the data matrix. Let $G$ denote the indicator matrix for non-missingness with $G_{ij} = 1$ if the item is observed and $G_{ij} = 0$ if the item is missing.

\[
\begin{bmatrix}
Y_O, Y_M \\
\end{bmatrix} \rightarrow G
\]

\[
\begin{array}{cccc}
5 & 10 & * & 6.8 \\
2 & 4.5 & * & * \\
3.2 & * & * & * \\
\end{array}
\begin{array}{cccc}
1 & 1 & 0 & 1 \\
1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
\end{array}
\]
Missing Data Mechanism $Prob(G|Y_0, Y_M)$

- **MCAR**: Missing Completely at Random. 
  $Prob(G|Y_0, Y_M) = Prob(G)$: The probability of missing is a constant (like flipping a coin).
Missing Data Mechanism $Prob(G|Y_0, Y_M)$

- **MCAR: Missing Completely at Random.**
  
  $Prob(G|Y_O, Y_M) = Prob(G)$: The probability of missing is a constant (like flipping a coin).

- **MAR: Missing at Random.**
  
  $Prob(G|Y_O, Y_M) = Prob(G|Y_O)$: Given the observed data (e.g. previous outcome and baseline measurements), the probability of missing is independent of missing data.
**Missing Data Mechanism**  $\text{Prob}(G|Y_0, Y_M)$

- **MCAR**: Missing Completely at Random.  
  $\text{Prob}(G|Y_O, Y_M) = \text{Prob}(G)$: The probability of missing is a constant (like flipping a coin).

- **MAR**: Missing at Random.  
  $\text{Prob}(G|Y_O, Y_M) = \text{Prob}(G|Y_O)$: Given the observed data (e.g. previous outcome and baseline measurements), the probability of missing is independent of missing data.

- **MNAR**: Missing Not at Random.
  If neither of the above hold
Missing Data Mechanism $Prob(G|Y_0, Y_M)$

- **MCAR**: Missing Completely at Random.
  
  $Prob(G|Y_O, Y_M) = Prob(G)$: The probability of missing is a constant (like flipping a coin).

- **MAR**: Missing at Random.
  
  $Prob(G|Y_O, Y_M) = Prob(G|Y_O)$: Given the observed data (e.g. previous outcome and baseline measurements), the probability of missing is independent of missing data.

- **MNAR**: Missing Not at Random.
  
  If neither of the above hold

- **We will first use the SWOG Data to illustrate which assumption is more reasonable and compare different methods.**
Marginal Model for EF in SWOG

- The ideal data for subject $i$ $Y_i = (Y_{O,i}, Y_{M,i})$ is:

$$Y_i \sim MVN (Z_i \beta, \Sigma_i)$$

where

$$Z_{ij}^T \beta = \beta_{00} + \beta_{10} g_i + \beta_{0j} + \beta_{1j} g_i + \beta_p * x_{p0} + \beta_s * x_{s0}$$

$g_i$ is group indicator, $x_{p0}$ and $x_{s0}$ are baseline performance score and disease severity, respectively, and the covariance structure is:

$$\text{Cov}(Y_{ij}, Y_{ik}) = \begin{cases} 
\sigma^2 & j = k \\
\sigma^2 \rho_{jk} & j \neq k.
\end{cases}$$
Parameter Interpretation

- $\beta_{00}$: Intercept
- $\beta_{01}$: Change bt. 1 mnth and baseline (V1-V0)
- $\beta_{03}$: Change bt. 3 mnth and baseline (V3-V0)
- $\beta_{06}$: Change bt. 6 mnth and baseline (V6-V0)
- $\beta_{10}$: Treatment comparison at Baseline (P0-F0)
- $\beta_{11}$: Treatment comparison at 1 mnth (P1-F1)
- $\beta_{13}$: Treatment comparison at 3 mnth (P3-F3)
- $\beta_{16}$: Treatment comparison at 6 mnth (P6-F6)
- $\beta_p$: Effect of baseline performance status (Perf.)
- $\beta_s$: Effect of baseline disease severity status (Sever.)
Complete-case Analysis

- 497 completers, out of 737 in total
- Fit the multivariate normal model with SAS Proc Mixed on the completers, and here is the result:

<table>
<thead>
<tr>
<th>Time Effect</th>
<th>Treatment Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1-V0</td>
<td>V3-V0</td>
</tr>
<tr>
<td>Est.</td>
<td>0.52</td>
</tr>
<tr>
<td>S.E.</td>
<td>(0.09)</td>
</tr>
</tbody>
</table>

Nonsignificant Treatment Effects.
Complete-Case Analysis

- An implicit assumption (MCAR): Complete cases are a random subsample of all the cases
Complete-Case Analysis

- An implicit assumption (MCAR): Complete cases are a random subsample of all the cases.
- Is this assumption plausible?

Table 2: Baseline Values bt. completers and incompleters

<table>
<thead>
<tr>
<th>Baseline Variables</th>
<th>Completers</th>
<th>Incompleters</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF</td>
<td>8.3 (1.5)</td>
<td>8.2 (1.5)</td>
<td>0.34</td>
</tr>
<tr>
<td>Performance Score</td>
<td>2%</td>
<td>8%</td>
<td>0.0005</td>
</tr>
<tr>
<td>(% of low perf.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease Severity</td>
<td>76%</td>
<td>83%</td>
<td>0.03</td>
</tr>
<tr>
<td>(% of Severe)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Complete-Case Analysis

- An implicit assumption (MCAR): Complete cases are a random subsample of all the cases
- Is this assumption plausible?

Table 3: Baseline Values bt. completers and incompleters

<table>
<thead>
<tr>
<th>Baseline Variables</th>
<th>Completers</th>
<th>Incompleters</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF</td>
<td>8.3 (1.5)</td>
<td>8.2 (1.5)</td>
<td>0.34</td>
</tr>
<tr>
<td>Performance Score (%) of low perf.</td>
<td>2%</td>
<td>8%</td>
<td>0.0005</td>
</tr>
<tr>
<td>Disease Severity (%) of Severe</td>
<td>76%</td>
<td>83%</td>
<td>0.03</td>
</tr>
</tbody>
</table>

- Completers are significantly better than incompleters in terms of Performance Score and Disease Severity
Complete-Case Analysis

• Pros:
Complete-Case Analysis

- **Pros:**
  - Simple
Complete-Case Analysis

- **Pros:**
  - Simple
- **Cons:**
Complete-Case Analysis

• **Pros:**
  • Simple

• **Cons:**
  • **Inefficient:**
    → Lose much information if many incompleters that contain partial information.
    → larger standard errors
Complete-Case Analysis

- **Pros:**
  - Simple

- **Cons:**
  - **Inefficient:**
    - Lose much information if many incompleters that contain partial information.
    - Larger standard errors
  - **Validity?:** analysis replies on restrictive MCAR assumption
    - Highly unlikely to be true in general
    - Complete cases can be very different from incompleters in important ways.
    - Biased inference
**Imputation: LOCF**

- Last Observation Carry Forward (LOCF) is to impute missing data with the last observed outcome.

- **Analysis using LOCF:**

<table>
<thead>
<tr>
<th></th>
<th>Time Effect</th>
<th>Treatment Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V1-V0</td>
<td>V3-V0</td>
</tr>
<tr>
<td>Completer Analysis</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>LOCF</td>
<td>0.38</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Nonsignificant Treatment Effects.
LOCF

- Pros:
LOCF

- Pros:
  - Simple
LOCF

- **Pros:**
  - Simple
  - Provide conservative inference in some important cases
LOCDF

• Pros:
  • Simple
  • Provide conservative inference in some important cases

• Cons:
LOCF

- **Pros:**
  - Simple
  - Provide conservative inference in some important cases

- **Cons:**
  - Too optimistic about sampling error
    Standard error is under-estimated due to ignoring the uncertainty of imputation
LOCF

- **Pros:**
  - Simple
  - Provide conservative inference in some important cases

- **Cons:**
  - Too optimistic about sampling error
    Standard error is under-estimated due to ignoring the uncertainty of imputation
  - Biased inferences in general.
LOCF

- **Pros:**
  - Simple
  - Provide conservative inference in some important cases

- **Cons:**
  - Too optimistic about sampling error
    Standard error is under-estimated due to ignoring the uncertainty of imputation
  - Biased inferences in general.
  - Does not use the observed data efficiently.

2 4 6 8 *

It is more plausible to impute by 10 instead of by 8 (LOCF).
Multiple Imputation

- Replace each missing value with $n$ plausible values
Multiple Imputation

- Replace each missing value with $n$ plausible values
- The resultant $n$ complete dataset are analyzed by standard procedures
Multiple Imputation

- Replace each missing value with $n$ plausible values
- The resultant $n$ complete dataset are analyzed by standard procedures
- Results from these $n$ analyses are then combined appropriately
Multiple Imputation

• Replace each missing value with $n$ plausible values
• The resultant $n$ complete dataset are analyzed by standard procedures
• Results from these $n$ analyses are then combined appropriately
• See (Rubin 1976, 1987)
Multiple Imputation

• Replace each missing value with $n$ plausible values
• The resultant $n$ complete dataset are analyzed by standard procedures
• Results from these $n$ analyses are then combined appropriately
• See (Rubin 1976, 1987)
• Available softwares:
  SAS Proc MI and MIANALYZE
  Solas 3.0
  SPlus libraries, NORM, CAT, MIX, PAN
  (Schafer 1997)
**SAS MI code for SWOG Data**

- **Impute missing values by random draws from**
  \[ f(Y_M | Y_O, G), \text{ which can be achieved by a simulation method (e.g. MCMC).} \]

- **Specify an imputation model:**
  Usually it is a probabilistic model for the ideal complete data e.g.
  \[ Y_M, Y_O \sim \text{MVN}(\mu(X), \Sigma(X)), \text{ where} \]
  \( X \) denotes observed covariates.

- **SAS code:**
  ```sas
  proc mi data=qolwide out=qolwide5 nimpute=5;
  by group perf sever;
  mcmc chain=multiple initial=em(itprint);
  var y0 y1 y3 y6;
  run;
  ```
SAS Code for Combining Inferences

• The resultant \( n \) complete dataset are analyzed by standard procedures:

```sas
data qol5;
    set qolwide5;
    y=y0; time=0; output;
    y=y1; time=1; output;
    y=y3; time=3; output;
    y=y6; time=6; output;
proc sort data=qol5; by _imputation_;
proc mixed data=qol5;
    class group;
    model y=t1 t3 t6 group t1*group t3*group t6*group perf sever/solution covb;
    repeated /subject=sub type=un;
    ods output SolutionF=mixparms CovB=mixcovb;
    by _imputation_; 
```
Combining Results for P6-F6

- Results from these $n$ analyses are then combined appropriately

<table>
<thead>
<tr>
<th>Imputation</th>
<th>P6-F6</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.16</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td>3</td>
<td>0.23</td>
<td>0.11</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>0.11</td>
</tr>
<tr>
<td>5</td>
<td>0.18</td>
<td>0.11</td>
</tr>
<tr>
<td>Average</td>
<td>0.18</td>
<td>0.12</td>
</tr>
</tbody>
</table>

- Combining Formulas:

$$\bar{Q} = \frac{1}{n} \sum_{i=1}^{n} \hat{Q}_i$$

$$V_{total} = V_{within} + (1 + \frac{1}{n})V_{between}$$
SAS Code for Combining Inferences

- SAS Proc MIANALYZE can be used to combine inferences

```sas
proc mianalyze parms=mixparms
covb(effectvar=rowcol)=mixcovb;
class group;
  modeleffects Intercept t1 t3 t6 group t1*group t3*group t6*group perf sever;
run;
```
## Multiple Imputation

<table>
<thead>
<tr>
<th>Time Effect</th>
<th>Treatment Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1-V0</td>
<td>V3-V0</td>
</tr>
<tr>
<td>Completer Analysis</td>
<td></td>
</tr>
<tr>
<td>Est.</td>
<td>0.52</td>
</tr>
<tr>
<td>S.E.</td>
<td>(0.09)</td>
</tr>
<tr>
<td>LOCF</td>
<td></td>
</tr>
<tr>
<td>Est.</td>
<td>0.38</td>
</tr>
<tr>
<td>S.E.</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td></td>
</tr>
<tr>
<td>Est.</td>
<td>0.46</td>
</tr>
<tr>
<td>S.E.</td>
<td>(0.08)</td>
</tr>
</tbody>
</table>

Significance Treatment Effect at Time 1 under MI.
Multiple Imputation

- **Pros:**
  - Broader Validity: valid under MAR.
  - Flexible: can handle both missingness in responses and covariates
  - Efficient
Multiple Imputation

• Pros:
  • Broader Validity: valid under MAR.
  • Flexible: can handle both missingness in responses and covariates
  • Efficient

• Cons:
  • Generally one needs to specify an imputation model \( f(Y_M | Y_O, R) \).
    This can be challenging with many missing variables (some continuous, others discrete)
  • Involves simulation, may not be as efficient as other method such as likelihood-based inference
Our New MI Approach

We propose a novel MI framework to overcome the above limitations (CXQ 2011).

- No parametric distributional assumptions.
Our New MI Approach

We propose a novel MI framework to overcome the above limitations (CXQ 2011).

- No parametric distributional assumptions.
- Flexible to model a mixture of discrete and continuous variables.
Our New MI Approach

We propose a novel MI framework to overcome the above limitations (CXQ 2011).

- No parametric distributional assumptions.
- Flexible to model a mixture of discrete and continuous variables.
- Easily handle the bounded or semi-continuous variables, which can be a problem for other imputation approaches.
Our New MI Approach

We propose a novel MI framework to overcome the above limitations (CXQ 2011).

- No parametric distributional assumptions.
- Flexible to model a mixture of discrete and continuous variables.
- Easily handle the bounded or semi-continuous variables, which can be a problem for other imputation approaches.
- Simultaneously address both the issue of inflexibility of the joint normal model and the issue of potential inconsistency of sequential imputation models.
An Example Code

• An example R Code

```r
bone.jor<- 
major(data=complete(bone.mice,nimpute), miss=miss, std=TRUE, digits=digits, nimpute = nimpute, nintv=nintv, burnin=burnin, predictorMatrix = predictorMatrix, visitSequence=visitSequence, isPassive=c(rep(0,11),1,1,1,0), passiveOperator=c(rep(0,11),1,1,1,0), nstep=nstep )
```

• For more details

http://tigger.uic.edu/~huixie/Research/Methods.html
### Table 4: Analysis of the imputed hip fracture data

<table>
<thead>
<tr>
<th>Variable</th>
<th>CC</th>
<th>MICE</th>
<th>CGM</th>
<th>IMPA</th>
<th>IMPB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethoh</td>
<td>1.41(0.40)</td>
<td>1.13(0.29)</td>
<td>1.15(0.30)</td>
<td>1.27(0.31)</td>
<td>1.31(0.30)</td>
</tr>
<tr>
<td>Smoke</td>
<td>-9.21(5.69)</td>
<td>-5.32(4.34)</td>
<td>-3.05(4.52)</td>
<td>-2.97(4.54)</td>
<td>-3.14(4.63)</td>
</tr>
<tr>
<td>Dementia</td>
<td>2.80(0.79)</td>
<td>1.69(0.47)</td>
<td>1.54(0.47)</td>
<td>1.60(0.48)</td>
<td>1.63(0.47)</td>
</tr>
<tr>
<td>Antiseiz</td>
<td>4.12(1.29)</td>
<td>2.45(0.62)</td>
<td>2.51(0.63)</td>
<td>2.67(0.66)</td>
<td>2.76(0.65)</td>
</tr>
<tr>
<td>LevoT4</td>
<td>3.15(1.34)</td>
<td>0.41(0.65)</td>
<td>1.03(0.63)</td>
<td>1.00(0.66)</td>
<td>0.89(0.62)</td>
</tr>
<tr>
<td>AntiChol</td>
<td>5.08(4.15)</td>
<td>-0.72(1.99)</td>
<td>-1.26(2.34)</td>
<td>-2.87(2.29)</td>
<td>-3.32(2.20)</td>
</tr>
<tr>
<td>Albumin</td>
<td>5.90(4.04)</td>
<td>-3.07(3.40)</td>
<td>2.53(2.97)</td>
<td>2.80(3.02)</td>
<td>2.60(3.40)</td>
</tr>
<tr>
<td>BMI</td>
<td>-0.12(0.04)</td>
<td>-0.12(0.03)</td>
<td>-0.11(0.03)</td>
<td>-0.11(0.03)</td>
<td>-0.10(0.03)</td>
</tr>
<tr>
<td>log(HGB)</td>
<td>4.60(5.99)</td>
<td>-7.56(4.80)</td>
<td>1.02(4.35)</td>
<td>1.46(4.43)</td>
<td>1.26(4.76)</td>
</tr>
<tr>
<td>smoke*loghgb</td>
<td>4.05(2.28)</td>
<td>2.40(1.74)</td>
<td>1.40(1.79)</td>
<td>1.82(1.80)</td>
<td>1.64(1.84)</td>
</tr>
<tr>
<td>AntiChol*albumin</td>
<td>-2.36(1.40)</td>
<td>0.02(0.55)</td>
<td>0.07(0.62)</td>
<td>0.36(0.65)</td>
<td>0.50(0.63)</td>
</tr>
<tr>
<td>Albumin*loghgb</td>
<td>-2.67(1.67)</td>
<td>0.95(1.35)</td>
<td>-1.43(1.19)</td>
<td>-1.58(1.22)</td>
<td>-1.52(1.36)</td>
</tr>
</tbody>
</table>
Findings from Systematic Analyses

- When no interaction exists, all MI methods: MI using SOR, the Joint normal and Sequential imputation method (MICE) perform reasonably well and better than CC.
- JN and sequential imputation method can perform poorly in accommodating interactions.
- SOR provides a robust and flexible alternative that performs better than existing MI softwares.
A Likelihood-Based Analysis

- Work with the likelihood of the observed data $Y_O, G$.

\[ f(Y_O, G) = f(Y_O) f(G|Y_O) \]

\[ f(G|Y_O) = \int f(G|Y_O, Y_M) f(Y_M|Y_O) dY_M \]

- Assuming MAR and parameter distinctness, it is valid to use the observed data $f(Y_O)$, ignoring modeling the missing data mechanism $f(G|Y_O)$. This is called *ignorability*.
A Likelihood-Based Analysis

• For now, assume ignorability holds so we don’t need to model missing data mechanism. Then....

• Mixed-Effects Model Analysis

• Available Softwares:
  SAS Proc MIXED, NLMIXED
  MIXOR (Hedeker & Gibbons 1994)
  Splus function gls() and R lme4 package.

• Splus Code:
  \texttt{gls(y \sim t_1 + t_3 + t_6 + t_1 * group + t_3 * group + t_6 * group + group + perf + sever, data = q.dat, correlation = corSymm(form = \sim 1 | sub), method = "ML")}
<table>
<thead>
<tr>
<th></th>
<th>Time Effect</th>
<th>Treatment Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V1-V0</td>
<td>V3-V0</td>
</tr>
<tr>
<td>Completer Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Est.</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>S.E.</td>
<td>(0.09)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>LOCF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Est.</td>
<td>0.38</td>
<td>0.40</td>
</tr>
<tr>
<td>S.E.</td>
<td>(0.07)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Est.</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>S.E.</td>
<td>(0.08)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>Likelihood MAR Inference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Est.</td>
<td>0.45</td>
<td>0.46</td>
</tr>
<tr>
<td>S.E.</td>
<td>(0.08)</td>
<td>(0.08)</td>
</tr>
</tbody>
</table>
Likelihood-Based Inference

• Pros:
  • Valid: inference is valid under MAR if done appropriately,
    Can assess the sensitivity of inference if data is MNAR
  • Efficient
  • No need to simulate for imputation

• Cons:
  • If covariates are missing, needs to model the missing covariates too.
  • This can be challenging with many variables (some continuous, others discrete) subject to missing
Our New Direct Estimation Approach

To overcome these limitations, we propose a novel direct estimation approach for regression analysis (QX 2011).

• Allow for arbitrary patterns of missingness in covariates
• Allow for a mixture of discrete and continuous covariates.
• Allow for interrelationship among covariates
• No covariate distributional assumptions
• Unlike ad-hoc approaches, predicted missing covariate values are consistent with statistical/econometrical regression models.
Improve Model Estimation

Table 5: Discrete Choice Model Estimation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SI Model</th>
<th>MVN Model</th>
<th>DF Model I</th>
<th>DF Model II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (A)</td>
<td>1.8(0.28)</td>
<td>3.5 (0.44)</td>
<td>3.7 (0.45)</td>
<td>3.0 (0.36)</td>
</tr>
<tr>
<td>Intercept (B)</td>
<td>1.6(0.20)</td>
<td>3.1 (0.36)</td>
<td>3.3 (0.36)</td>
<td>2.8 (0.31)</td>
</tr>
<tr>
<td>Price</td>
<td>-3.4(0.50)</td>
<td>-6.1(0.76)</td>
<td>-6.6(0.86)</td>
<td>-5.4(0.66)</td>
</tr>
<tr>
<td>Coupon</td>
<td>53.6 (3.32)</td>
<td>2.4 (0.56)</td>
<td>4.4 (1.28)</td>
<td>3.5 (1.24)</td>
</tr>
</tbody>
</table>

- Naive single imputation (SI) has large bias in parameter estimates in the discrete choice demand model.
- Smaller Bias remains in parametric direct estimation method (MVN).
- The bias affects policy decision making, such as optimal price setting.
## Improve Demand Simulation

### Table 6: The Impact of Price Cut on Market Shares

<table>
<thead>
<tr>
<th>Prod</th>
<th>SI Before</th>
<th>SI After</th>
<th>MVN Before</th>
<th>MVN After</th>
<th>DF Model I Before</th>
<th>DF Model I After</th>
<th>DF Model II Before</th>
<th>DF Model II After</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>35.5%</td>
<td>27.2%</td>
<td>33.8%</td>
<td>22.5%</td>
<td>33.2%</td>
<td>21.3%</td>
<td>33.0%</td>
<td>23.4%</td>
</tr>
<tr>
<td>B</td>
<td>41.2%</td>
<td>55.1%</td>
<td>40.1%</td>
<td>63.9%</td>
<td>40.3%</td>
<td>65.1%</td>
<td>40.3%</td>
<td>61.5%</td>
</tr>
<tr>
<td>C</td>
<td>23.2%</td>
<td>17.7%</td>
<td>26.1%</td>
<td>13.6%</td>
<td>26.5%</td>
<td>13.6%</td>
<td>26.7%</td>
<td>15.1%</td>
</tr>
</tbody>
</table>

Note: Assuming 20% price cut of product B.
Figure 1: Comparison of Optimal Prices.

Substantial differences on the optimal price suggested by the different methods: the suggested optimal price cuts are -22%, -8%, -3%, -14% for SI, MVN, DF I, and DF II respectively.
Our New Statistical Matching Approach

- Conceptual Missingness ("soft missing")

\[
\begin{array}{ccc}
Y_C & Y_A & Y_B \\
(a) & & \\
(b) & & \\
(c) & & \\
(d) & & \\
\end{array}
\]
Our New Statistical Matching Approach

We propose a novel statistical matching method (QX 2013) to combine complementary datasets from independent samples to relate nonoverlapping variables.

- Allow for Matching data from similar units when matching data from same unit is impossible.
- Model-based nonparametric statistical matching procedures
- Overcome important limitations of current nonparametric matching procedures
- Performs better than model-based parametric statistical matching procedures.
Table 7: Fusion Results with the monetary value of counterfeit purchases over the past year ($Y_B^*$).

<table>
<thead>
<tr>
<th>$Y_A$</th>
<th>Param.</th>
<th>Hot-deck</th>
<th>FORM</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. It is good for health.</td>
<td>-21.9 (10.3)**</td>
<td>-28.6 (11.3)**</td>
<td>-35.9 (11.6)**</td>
</tr>
<tr>
<td>8. It is convenient to buy.</td>
<td>2.1 (6.3)</td>
<td>14.3 (6.8)**</td>
<td>3.5 (9.3)</td>
</tr>
<tr>
<td>10. I need for work and social interaction.</td>
<td>-7.3 (5.2)</td>
<td>-14.3 (5.7)**</td>
<td>-16.9 (7.4)**</td>
</tr>
<tr>
<td>11. Shop often in mall.</td>
<td>-4.1 (8.2)</td>
<td>-22.9 (8.5)**</td>
<td>-2.1 (14.0)</td>
</tr>
<tr>
<td>14. Shop often in licensed store.</td>
<td>-15.9 (5.5)**</td>
<td>-24.3 (3.8)**</td>
<td>-23.3 (6.8)**</td>
</tr>
<tr>
<td>16. Shop often on Internet.</td>
<td>14.9 (5.6)**</td>
<td>27.1 (5.3)**</td>
<td>22.9 (6.1)**</td>
</tr>
<tr>
<td>19. Interested in receiving catalog.</td>
<td>6.8 (6.5)</td>
<td>15.7 (8.0)*</td>
<td>14.7 (8.3)*</td>
</tr>
</tbody>
</table>
**Improve Prediction Accuracy**

Table 8: Comparing Different Fusion Methods on Individual Prediction of $\ln(Y_B^*)$.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>Improve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric</td>
<td>0.38</td>
<td>0%</td>
</tr>
<tr>
<td>Hot-deck</td>
<td>0.57</td>
<td>-50%</td>
</tr>
<tr>
<td>FORM</td>
<td>0.30</td>
<td>22%</td>
</tr>
</tbody>
</table>
Weighting Method

- Marginal models with GEE’s are often used for non-normal outcome
- GEE method is not likelihood-based analysis, invalid under MAR
- Weighted GEE provide valid inference under MAR
- Weight requires modeling the missing-data mechanism
- Not widely implemented yet
- If interested, see Robins et al. 1995.
Impact of MNAR

• The above analyses are potentially biased if missing is **Missing Not at Random (MNAR):** selection on unobservable; nonignorability.

• Need a tool to measure the effect of nonignorable missingness on standard analyses!

• Quantitatively measure sensitivity of MLE to nonignorability at the MAR model.

• Collect data on missing mechanism and do nonignorable modeling only if sensitivity is noticeable.
Example: SWOG QoL Data

- **Dropout model:**

  \[
  G_{ij} \mid s_{ij}, y_{ij}, G_{i,j-1} = 1 \sim \text{Bernoulli} \left( h(\gamma_{0,j} + \gamma_{0,i}y_{i,j-1} + \gamma_{1}y_{ij}) \right),
  \]

  \[j = 1, 2, 3\]

- Dropout depends on weeks, the current and immediately previous \( Y \) value.

- \( \gamma_1 = 0 \): MAR, otherwise MNAR.

- Sensitivity analysis: vary \( \gamma_1 \) check the change of estimates.

- Need to fit complicated nonignorable model: Can take hours or days of computational time.

- Not scalable to large datasets.
Our Sensitivity Analysis Approach

- Our approach

\[
\hat{\beta}(\gamma_1) - \hat{\beta}(0) \approx ISNI
\]

- No need to fit any nonignorable models: reduce computation time to just a few seconds.
- Scalable to large datasets.
- Accurate for moderate nonignorability when a rich set of missingness predictors exist (e.g., in longitudinal data).
Our Sensitivity Analysis Approach


\[ \text{ISNI} = -\nabla^2 L_{\beta,\beta}^{-1} \nabla^2 L_{\beta,\gamma_1}, \]

where

\[
\nabla^2 L_{\beta,\beta} = \sum_{i=1}^{n} \frac{\partial^2 \ln f_i(y_i^{(k_i)})}{\partial \beta \partial \beta^T} \bigg|_{\hat{\beta}(0)}
\]

\[
\nabla^2 L_{\beta,\gamma_1} = -\sum_{i:k_i < m} \frac{h_{i,k_i+1}'}{1 - h_{i,k_i+1}} \frac{\partial E_{\beta}^{Y_{i,k_i+1}|Y_i^{(k_i)}}}{\partial \beta} \bigg|_{\hat{\beta}(0)}
\]
### ISNI in Milk Data (Xie and Heitjan 2009)

<table>
<thead>
<tr>
<th></th>
<th>Time Effect Estimate</th>
<th>Treatment Effect Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_{01}$</td>
<td>$\beta_{02}$</td>
</tr>
<tr>
<td>MAR Est.</td>
<td>0.45</td>
<td>0.47</td>
</tr>
<tr>
<td>SE</td>
<td>0.07</td>
<td>0.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Logit</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = 0$</td>
<td>ISNI</td>
<td>-0.08</td>
</tr>
<tr>
<td>logL=-529.25</td>
<td>$C$</td>
<td>1.28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Probit*</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = 0.4$</td>
<td>ISNI</td>
<td>-0.09</td>
</tr>
<tr>
<td>logL=-529.01</td>
<td>$C$</td>
<td>1.17</td>
</tr>
</tbody>
</table>

- **$C$**: minimum nonignorable missingness needed to have important sensitivity. $C = 1$ suggested as cutoff value.
- Sensitivity to nonignorable dropout increases from baseline to month 6
- Treatment Comparisons are not sensitive
Our Sensitivity Analysis Approach


• Drive new sensitivity indices applicable for mobile Health and big health data.

• An NIH R01 grant award (1R01CA178061-01A1 Xie) on further methodological development and software creation.
Summary

Consider using missing data methods if you want

- More powerful treatment, program and policy impact evaluation.
Summary

Consider using missing data methods if you want

• More powerful treatment, program and policy impact evaluation.

• More effective identification of risk factors and causal relationships.
Summary

Consider using missing data methods if you want

- More powerful treatment, program and policy impact evaluation.
- More effective identification of risk factors and causal relationships.
- To improve policy simulation.
Summary

Consider using missing data methods if you want

- More powerful treatment, program and policy impact evaluation.
- More effective identification of risk factors and causal relationships.
- To improve policy simulation.
- To improve usability and accessibility of large medical databases.
Summary

Consider using missing data methods if you want

- More powerful treatment, program and policy impact evaluation.
- More effective identification of risk factors and causal relationships.
- To improve policy simulation.
- To improve usability and accessibility of large medical databases.
- To quantify and improve the reliability of information extraction and decision making.