Sensitivity analysis after multiple imputation: Application of a weighting approach to epidemiological data.

V. Bousquet, Y. Le Strat, C. Larsen, J-C. Desenclos

ISCB – 26/08/09 - Prague
Introduction

• Multiple imputation (MI) assumes the missing data are Missing At Random (MAR) meaning that the missingness mechanism does not depend on the unobserved data. If missing data are Non Missing At random (NMAR), MI may yield biased estimates.

• MAR assumption is not easily testable and there are very limited methods to assess the sensitivity of MI to the MAR assumption.
Introduction

• Recently, Carpenter et al. [1] proposed a sensitivity analysis method allowing to provide NMAR estimates after MI and therefore to test the robustness of the results obtained under the MAR hypothesis.

• The authors described and applied the method in a clinical trial context.

Objective

• Apply Carpenter’s sensitivity analysis method on observational data collected from a French surveillance system of hepatitis C viral infection.

• Provide guidelines to apply and interpret this method for epidemiological surveys.
Carpenter’s weighting method (1)

• Suppose we have a variable Y with missing data and X a vector of complete or incomplete variables.
  Let \( R_i = 1 \) if we observe \( Y_i \) and \( R_i = 0 \) otherwise.

• Suppose that the probability to observe Y depends on Y.

  \[
  \text{logit } \Pr(R_i=1) = \alpha + \beta X_i + \delta Y_i
  \]

If \( \delta = 0 \), given the fully observed data, the mechanism causing the missing data of Y does not depend on Y (MAR).

If \( \delta \neq 0 \), even taking into account the information in the observed data, the missingness mechanism still depends on the potentially missing Y (NMAR).
Carpenter’s weighting method (2)

Notations:
M databases are created by a multiple imputation method

\( n_1 \) is the number of imputed values

\( \hat{\theta}_m \) is the estimate of the parameter of interest from the database \( m \)

\( Y^m_i \) is the imputed value of \( Y \) for individual \( i \) in the database \( m \).

\[
\tilde{w}_m = \exp \left( \sum_{i=1}^{n_1} - \delta Y^m_i \right) \quad w_m = \frac{\tilde{w}_m}{\sum_{k=1}^{M} \tilde{w}_k} \quad \hat{\theta}_{NMAR} = \sum_{m=1}^{M} w_m \hat{\theta}_m
\]

\[
\tilde{V}(\hat{\theta}_{NMAR}) \approx \tilde{V}_W(\hat{\theta}_{NMAR}) + (1 + 1/M) \tilde{V}_B(\hat{\theta}_{NMAR})
\]

\[
\tilde{V}_W(\hat{\theta}_{NMAR}) = \sum_{m=1}^{M} w_m \delta^2_m \quad \tilde{V}_B(\hat{\theta}_{NMAR}) = \sum_{m=1}^{M} w_m (\hat{\theta}_m - \hat{\theta}_{NMAR})^2
\]
Data

- Data were collected from a French national surveillance system from 2001 to 2004: 26 hepatology reference centers located in university hospitals in France.

- Among 14,485 HCV+ patients, 3,153 drug users were selected to assess risk factors associated with severe liver disease (cirrhosis and hepatocellular carcinoma).
Variables included in the multivariate analysis

<table>
<thead>
<tr>
<th>Variables</th>
<th>% missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe liver disease</td>
<td>0</td>
</tr>
<tr>
<td>Sex</td>
<td>0</td>
</tr>
<tr>
<td>Age</td>
<td>0</td>
</tr>
<tr>
<td>Duration of HCV infection at referral</td>
<td>12.5</td>
</tr>
<tr>
<td>Time between HCV+ test and referral</td>
<td>11.5</td>
</tr>
<tr>
<td>History of excessive alcohol intake</td>
<td>14.6</td>
</tr>
<tr>
<td>HIV</td>
<td>16.8</td>
</tr>
<tr>
<td>HBsAg (HBV)</td>
<td>17.2</td>
</tr>
<tr>
<td>HCV genotype 3</td>
<td>29.6</td>
</tr>
</tbody>
</table>
Multiple imputation

- Initial analysis
  - Imputation by chained equations (package ice STATA 9.2)
  - 30 imputed databases
  - Imputation model = analysis model
  - Multivariate analysis performed (Complete Case and Multiple Imputation)

- Sensitivity analysis
  - Imputation by chained equations (package mice R 9.2)
  - 1000 imputed databases
Reasons to focus on alcohol and genotype 3

• We applied the sensitivity analysis on two variables for epidemiological and missingness mechanism reasons:

• History of excessive alcohol intake
  ➢ associated with a rapid progression of hepatic fibrosis
  ➢ Probably NMAR: a HCV+patient with an excessive alcohol consumption history could be tempted not to declare it.

• Genotype 3
  ➢ Recently reported to be related to the pathogenicity of the virus
  ➢ Genotype 3 is probably MCAR or MAR: this variable is reported by the investigators independently to the characteristics of the patients or the HCV genotyping.
Step 1: Explore the missingness mechanism (1)

• Create a missing indicator:
  For example: \( R_{\text{alcohol}} = 0 \) if alcohol is missing and 1 otherwise

• Fit a logistic regression model to explain the missing indicator including:
  √ all covariates retained after univariate analysis (CC)
  √ the variable of interest = severe liver disease

\[
\text{logit } \Pr(R_{\text{alcohol}}=0 \mid \text{covariates}) = \alpha + \beta_0 \text{case} + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 \text{durhc} + \beta_4 \text{delay} + \beta_5 \text{hiv} + \beta_6 \text{aghbs} + \beta_7 \text{geno3}
\]
Step 1: Explore the missingness mechanism (2)

| Ralc   | Coef.  | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|--------|--------|-----------|-------|------|----------------------|
| case   | 0.3747385 | 0.3919389 | 0.96  | 0.339 | -3.934476 1.142925   |
| sex    | 0.2870801 | 0.1986675 | 1.45  | 0.148 | -1.023011 0.6764613 |
| age    | 0.0885285 | 0.2111654 | 0.42  | 0.675 | -0.325348 0.5024051 |
| durhc  | -0.1018725 | 0.2077937 | -0.49 | 0.624 | -0.5091406 0.3053957 |
| delay  | -0.4009601 | 0.1848716 | -2.17 | 0.030 | -0.7633018 -0.038618 |
| hiv    | 0.3786807 | 0.4153187 | 0.91  | 0.362 | -0.4353289 1.19269  |
| HBsAg | -0.8808649 | 0.4439548 | -1.98 | 0.047 | -1.751 0.0107294    |
| geno3  | -0.1638119 | 0.1819875 | -0.90 | 0.368 | -0.5205008 0.1928769 |
| _cons  | 2.221081  | 0.2460697 | 9.03  | 0.000 | 1.738793 2.703369    |

- The parameter of the variable of interest is not significant
  - the missingness mechanism of alcohol does not depend on the variable of interest => unbiased CC estimate.

- Two possible conclusions:
  - $R_{\text{alcohol}}$ is MAR depending on covariates
  - $R_{\text{alcohol}}$ is NMAR depending on covariates and $R_{\text{alcohol}}$
Step 2: Propose a plausible range of $\delta$ (1)

It is important to limit the range of $\delta$ because given $M$, when $\delta$ increases, the NMAR estimate is only based on one imputed database.

$\delta = 0.5$ is retained as a sensible value for alcohol.
Step 2: Propose a plausible range of $\delta$ (2)

A plausible range of $\delta$ could also be proposed from the positive regression coefficients giving a range between 0.09 and 0.39.
Step 3: Graphical diagnostic (Alcohol)

M = 1000 databases

Number of imputations used to estimate $\hat{OR}_{NMAR}$

Estimates of $\hat{OR}_{MAR}$ obtained from multiple imputation

$\delta = 0.5$

Normalised weights $w_i$

estimate under MAR
Results (1)

\[ \text{Variation} = \frac{\hat{\text{OR}}_{\text{NMAR}} - \hat{\text{OR}}_{\text{MAR}}}{\hat{\text{OR}}_{\text{MAR}}} \]

History of alcohol intake

HCV Genotype 3
## Results (2)

Robustness of the results to the non respect of the MAR hypothesis (alcohol) to a high percentage of missing values (genotype 3)

<table>
<thead>
<tr>
<th>Variables</th>
<th>% missing data</th>
<th>δ</th>
<th>OR$_{\text{MAR}}$</th>
<th>OR$_{\text{NMAR}}$</th>
<th>Variation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol</td>
<td>14.6</td>
<td>0.5</td>
<td>2.98 [2.17 ; 4.09]</td>
<td>2.87 [2.10 ; 3.93]</td>
<td>3.48</td>
</tr>
<tr>
<td>Genotype 3</td>
<td>29.5</td>
<td>0.3</td>
<td>1.44 [1.06 ; 1.96]</td>
<td>1.43 [1.07 ; 1.92]</td>
<td>0.51</td>
</tr>
</tbody>
</table>
Discussion

• Analysing epidemiological surveys with missing data requires making untestable assumptions, so it is crucial to test the robustness of the results to departures from these assumptions.

• Among several more complex sensitivity analysis methods (e.g. pattern-mixture), Carpenter’s approach can be easily implemented. So we believe that this method could be used by epidemiologists if clear recommendations were provided.

• Our ambition was to make it feasible for an epidemiologist to apply the method, in particular to:
  - determine the range of values for \( \delta \) (to avoid situations with high weights associated to few databases)
  - calculate the relative variation to conclude that the results are reliable if \( \text{variation} < 10\% \).