Missing data and multiple imputation

Professor Irene Petersen
Primary Care & Population Health
Today

• Missing data

• Different methods to deal with missing data

• Multiple imputation (MI) of missing data
What do we mean by missing data?

• Missing values are observations/records which were:
  – Never collected
  – Lost by accident
  – Wrongly collected and so deleted

• Missing data can be in outcomes, explanatory variables or covariates
Missing data in electronic health records

Health indicators
• Blood pressure
• Weight
• Height
• Smoking
• Alcohol
• Cholesterol
Recording of health indicators – 3 cohorts

Petersen et al. (2019) **Health indicator recording in UK primary care electronic health records: key implications for handling missing data.**
People with diabetes are more likely to have health indicators recorded.
Data can be missing in different ways….

The scale is broken

• Missing Completely at Random (MCAR): Missingness of Y is independent of Y and X.

More younger women have weight measured than younger men

• Missing at Random (MAR): Missingness of Y is independent of Y given X.

Only those with high weight have weight recorded

• Missing Not at Random (MNAR): Missingness of Y is depending on Y, even after conditioning on X.
Recording of **weight** by age and gender

What kind of missingness??
How do I know if data are MAR, MCAR or MNAR?

• We can exclude that data are missing completely at random (MCAR)
• Important to understand dataset
• Tricky to demonstrate whether data are MAR or MNAR
  – Compare mean or % to external data sources
  – E.g. Health Survey for England
**Ad hoc methods**

- *Ad hoc* = a method that is proposed for convenience

- Sometimes *ad hoc* methods are fine, but not always
Different types of ad-hoc methods

- Exclude variables with incomplete records
- Complete case analysis
- Create missing data category
- Mean imputation
- Regression imputation
Exclude variable with incomplete records

• Sometimes we have no other options
  – High proportion of missing data

• Analysis may be biased
  – If variable is a strong confounder
Complete case analysis

• Reduce sample size = reduce precision of results 😞

• Make assumption
  – Complete cases represent full dataset

• May be OK if some situations
  – Data missing completely at random
  – Small proportion of missing data
  – If missingness is not associated with outcome
• Create missing data category

• Mixed bag – results not meaningful 😞

• Severe bias can arise, in any direction

• Variable will not correctly adjust for confounding
Missing data category
Missing data category
Some real data…

- Cardiovascular risk in people with severe mental illnesses
- Sample of 42,213 people
- Risk factors:
  - Age
  - Sex
  - Smoking
  - Diabetes
  - Blood pressure
- All variables have missing data
Risk of cardiovascular diseases in 42,213 people - Complete case analysis

<table>
<thead>
<tr>
<th></th>
<th>Complete case</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N=3,736</td>
</tr>
<tr>
<td>Hazard ratio (95% CI)</td>
<td>P</td>
</tr>
<tr>
<td>Age, years (per unit increase)</td>
<td>1.05 (1.04 to 1.06)</td>
</tr>
<tr>
<td>Sex: Females v males</td>
<td>1.19 (0.86 to 1.66)</td>
</tr>
<tr>
<td>SBP, mmHg (per unit increase)</td>
<td>1.94 (1.33 to 2.82)</td>
</tr>
<tr>
<td>Diabetes: Yes v no</td>
<td>1.19 (0.82 to 1.71)</td>
</tr>
<tr>
<td>Smoking: Never</td>
<td>1</td>
</tr>
<tr>
<td>Ex</td>
<td>1.77 (1.23 to 2.53)</td>
</tr>
<tr>
<td>Current</td>
<td>1.56 (1.07 to 2.28)</td>
</tr>
</tbody>
</table>

SBP: systolic blood pressure
Mean imputation

• Impute average values for missing data

• For example replace all missing systolic blood pressure values with 140 mmHg
Issues with mean imputation
Systolic Blood pressure 10,000 observation
20% missing = 130 mmHg

Mean = 130  Variance = 319

Mean = 130  Variance = 256
Regression Imputation

- Fit a regression model
- Use all information available in existing data
- Provides a ‘best guess’

Health indicators
- Blood pressure
- Cholesterol
- Weight
- Height
- Smoking
- Alcohol

Predictors
- Age
- Gender
- Social deprivation
- Ethnicity
- Diseases/illness
- Medication
Regression imputation

What is the problem?
Regression imputation

- Just ONE estimate - do not account for uncertainty of the missing data 😞

- Creates datasets with too small variation 😞
  - (too narrow confidence intervals)

- Bias results
Multiple imputation

• Builds on regression modelling

• Replace missing values with ‘plausible’ values
  – Based on distribution of observed values
  – Include randomness to reflect uncertainty
Multiple imputation (MI) involves three steps

- **Step 1:** Create *multiple* imputed datasets
  - We will never know the *true* values of the missing data
  - Set of values – not just a single value

- **Step 2:** Analyse each dataset separately

- **Step 3:** Combine results of *m* analyses using ‘Rubin’s rule’
Multiple imputation (MI) of missing data

- Builds on regression imputation – three stages

- Implemented in SAS, Stata, R
Rubin’s rule

• Combine results from individual analysis
  – Overall point estimate is just the average

• Variance estimate
  – Within imputation variance
  – Between imputation variance

• For details see White IR, Royston P, Wood AM (2011).
An example of multiple imputation of a single variable

<table>
<thead>
<tr>
<th>ID</th>
<th>Y</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.1</td>
<td>3.4</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>3.9</td>
</tr>
<tr>
<td>3</td>
<td>2.3</td>
<td>2.6</td>
</tr>
<tr>
<td>4</td>
<td>3.6</td>
<td>1.9</td>
</tr>
<tr>
<td>5</td>
<td>0.8</td>
<td>2.2</td>
</tr>
<tr>
<td>6</td>
<td>3.6</td>
<td>3.3</td>
</tr>
<tr>
<td>7</td>
<td>3.8</td>
<td>1.7</td>
</tr>
<tr>
<td>8</td>
<td>?</td>
<td>0.8</td>
</tr>
<tr>
<td>9</td>
<td>?</td>
<td>2.0</td>
</tr>
<tr>
<td>10</td>
<td>?</td>
<td>3.2</td>
</tr>
</tbody>
</table>
## Multiply imputed data

<table>
<thead>
<tr>
<th>ID</th>
<th>Original data</th>
<th>Imp 1</th>
<th>Imp 2</th>
<th>Imp 3</th>
<th>Imp 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>1</td>
<td>1.1</td>
<td>3.4</td>
<td>1.1</td>
<td>3.4</td>
<td>1.1</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>3.9</td>
<td>1.5</td>
<td>3.9</td>
<td>1.5</td>
</tr>
<tr>
<td>3</td>
<td>2.3</td>
<td>2.6</td>
<td>2.3</td>
<td>2.6</td>
<td>2.3</td>
</tr>
<tr>
<td>4</td>
<td>3.6</td>
<td>1.9</td>
<td>3.6</td>
<td>1.9</td>
<td>3.6</td>
</tr>
<tr>
<td>5</td>
<td>0.8</td>
<td>2.2</td>
<td>0.8</td>
<td>2.2</td>
<td>0.8</td>
</tr>
<tr>
<td>6</td>
<td>3.6</td>
<td>3.3</td>
<td>3.6</td>
<td>3.3</td>
<td>3.6</td>
</tr>
<tr>
<td>7</td>
<td>3.8</td>
<td>1.7</td>
<td>3.8</td>
<td>1.7</td>
<td>3.8</td>
</tr>
<tr>
<td>8</td>
<td>?</td>
<td>0.8</td>
<td>0.2</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>9</td>
<td>?</td>
<td>2.0</td>
<td>1.7</td>
<td>2.0</td>
<td>2.4</td>
</tr>
<tr>
<td>10</td>
<td>?</td>
<td>3.2</td>
<td>2.7</td>
<td>3.2</td>
<td>2.5</td>
</tr>
</tbody>
</table>
So far

• Missing data mechanisms

• Different ad-hoc methods to deal with missing data

• Multiple Imputation of a single variable

• Now we will touch on multiple imputation of multiple variables
FCS Multiple Imputation

- 5 variables with missing data Y1, Y2, Y3, Y4, Y5
- Breaks the problem down into individual regression models
  \[ f(Y1|Y1(\text{obs}), Y2, Y3, Y4, Y5, x1, x2) \]
  \[ f(Y2|Y1, Y2(\text{obs}), Y3, Y4, Y5, x1, x2) \]
  \[ f(Y3|Y1, Y2, Y3(\text{obs}), Y4, Y5, x1, x2) \]
  ...
- Each is a model for a single variable
- Logistic, linear model...
FCS Multiple Imputation

• Combine thousands of regression models
FCS Multiple Imputation

We need to think....
A few things to consider before doing MI

• Why are the data missing?
• What variables may explain missing data?
  – Age, gender, deprivation, diseases, drug treatment

• Clear idea of your analysis model (the analyses that you will perform after MI)
  – Outcome
  – Any interactions, non-linear relationships
Example from the real world - QRisk

Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study

Julia Hippisley-Cox, professor of clinical epidemiology and general practice,1 Carol Coupland, senior lecturer in medical statistics,1 Yana Vinogradova, research fellow in medical statistics,1 John Robson, senior lecturer in general practice,2 Margaret May, research fellow in medical statistics,3 Peter Brindle research and development strategy lead4
First Qrisk model

• Cox model for prediction of cardiovascular events in 1.3m UK patients aged 35–74 from GP data
• Found **no** association between Total/HDL cholesterol and CVD outcome (1.001 (95% CI 0.999 to 1.002))

• Outcome was **not** properly included in the imputation model.
  – Included time to event, but not the variable indicating an event
• MI of ratios can be tricky (see Morris et al)
Clear idea of your analysis model
Clear idea of your analysis model
How to select the variables for MI

• All variables that will go into analysis model

• Add extra variables (not in analysis model)
  – Auxiliary variables
  – Increase the likelihood that data are MAR
  – Improve precision and decrease potential bias
Longitudinal multiple imputation – Twofold FCS algorithm

- Impute data at a given time block
- Use information available +/- one time block
- Move on to next time block
- Repeat procedure x times

Two-fold FCS algorithm implemented in Stata

The Stata Journal (2014)
14, Number 2, pp. 418–431

Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data

Catherine Welch
University College London
London, UK
catherine.welch@ucl.ac.uk

Jonathan Bartlett
London School of Hygiene & Tropical Medicine
London, UK
jonathan.bartlett@lshtm.ac.uk

Irene Petersen
University College London
London, UK
i.petersen@ucl.ac.uk

Abstract. Electronic health records of longitudinal clinical data are a valuable resource for health care research. One obstacle of using databases of health records in epidemiological analyses is that general practitioners mainly record data if they are clinically relevant. We can use existing methods to handle missing data,
Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data

Catherine A. Welch, Irene Petersen, Jonathan W. Bartlett, Ian R. White, Louise Marston, Richard W. Morris, Irwin Nazareth, Kate Walters and James Carpenter
Before we finish we will go back to our example
<table>
<thead>
<tr>
<th></th>
<th>Complete case</th>
<th>After MI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N=3,736</td>
<td>N= 42,313</td>
</tr>
<tr>
<td><strong>Hazard ratio (95% CI)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, years (per unit increase)</td>
<td>1.05 (1.04 to 1.06)</td>
<td>1.06 (1.06 to 1.06)</td>
</tr>
<tr>
<td>Sex: Females v males</td>
<td>1.19 (0.86 to 1.66)</td>
<td>0.74 (0.68 to 0.81)</td>
</tr>
<tr>
<td>SBP, mmHg (per unit increase)</td>
<td>1.94 (1.33 to 2.82)</td>
<td>1.87 (1.67 to 2.09)</td>
</tr>
<tr>
<td>Diabetes: Yes v no</td>
<td>1.19 (0.82 to 1.71)</td>
<td>1.60 (1.38 to 1.86)</td>
</tr>
<tr>
<td>Smoking: Never Ex</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Current</td>
<td>1.77 (1.23 to 2.53)</td>
<td>1.36 (1.24 to 1.50)</td>
</tr>
<tr>
<td>Current</td>
<td>1.56 (1.07 to 2.28)</td>
<td>1.55 (1.40 to 1.71)</td>
</tr>
</tbody>
</table>
Literature

- Petersen et al. Health indicator recording in UK primary care electronic health records: key implications for handling missing data. *Clinical Epidemiology* 2019
- Pedersen et al. Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology* 2017
- Sterne et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls *BMJ* 2009; 338:b2393

Books
- Van Buuren *Flexible Imputation of Missing Data* 2012
- Carpenter and Kenward *Multiple Imputation and its Application* 2013