MISSUITE: A Shiny Application for Clinical Trial Missing Data Analysis

GASP 2018

Chenguang Wang

Johns Hopkins University

Motivation

Missing data

- Missing data is *ubiquitous* in clinical trials
- *Validity* of statistical analysis results are *threatened* by missing data
- Inference requires *untestable assumptions* about missing data mechanism
- Rigorous *sensitivity analyses* examining sensitivity to missing data mechanism assumptions are *crucial* and should even be mandatory

Global sensitivity analysis

- Apply *benchmark* assumptions to identify the full data model
- Consider *deviations* from the benchmark assumptions and examine the robustness
- Exploring the basics of the missing data helps to *design* the sensitivity analysis

Goal

- To develop a statistical software that is *user-friendly* with *interactive* features
- To aid users to *efficiently* apply missing data *imputation* methods in existing software packages
- To *explore* the nature of the missing data
- To serve as the *first step* of rigorous missing data sensitivity analysis

Imputation Algorithms

General setting

- Z: treatment assignment
- $X_1, \ldots X_P$: baseline covariates
- $Y_1, \ldots, Y_K$: post-randomization outcomes
- $D = \{D_1, \ldots, D_J\} = \{X_1, \ldots, X_p, Y_1, \ldots, Y_K\}$: all data
- $M = \{M_1, \ldots, M_J\}$: missing data indicator
- D_{obs}: observed data
- D_{mis}: missing data
- $D_{-j} = \{D_1, \ldots, D_{j-1}, D_{j+1}, \ldots D_J\}$

Missing at random

- $M|D = M|D_{obs}$
- $D_{mis}|M, D_{obs} = Y_{mis}|D_{obs}$

Data type

- Binary
- Unordered-Categorical
- Ordered-Categorical
- Continuous
  - Proportion
  - Non-Negative

Multiple imputation software packages

- *MICE*: Multivariate Imputation by Chained Equations
- *Amelia*: A Program for Missing Data
- *missForest*: Nonparametric Missing Value Imputation using Random Forest
- *Hmisc*: Harrell Miscellaneous
- *mi*: Missing Data Imputation and Model Checking

MICE

- Multiple imputation using *Fully Conditional Specification* (FCS), also known as *multiple imputation using chained equations* (MICE)
- Imputation models specified conditionally for each variable

$$f(D_1|D_{-1}, \theta_1)$$
$$f(D_2|D_{-2}, \theta_2)$$
$$\vdots$$

- At tth iteration

$$\theta_j^{(t)} \sim \pi(\theta_j|D_{j,obs}, D_{-j}^{(t-1)})$$
$$D_{j,mis}^{(t)} \sim f(D_j|D_{-j}^{(t-1)}, \theta_j^{(t)})$$

Amelia

- Assume $D \sim N(\mu, \Sigma)$
- Imputation by EM with bootstrapping (*EMB*) algorithm
  - Apply EM to find the mode of the posterior given the bootstrapped sample
  - Draw D_{mis} from $f(D_{mis}|D_{obs}, \mu, \Sigma)$
- Ordinal data are considered continuous
- Nominal data are re-coded using dummy variables that are further considered continuous

missForest

- An implementation of non-parametric *random forest* (RF) algorithm
- For j, train an *RF* on the observed data $D_{obs,j}|D_{obs,-j}$, then predict the missing values $D_{mis,j}|D_{mis,-j}$
- Proceed iteratively until convergence
- By averaging over trees, random forest intrinsically constitutes a multiple imputation scheme

Hmisc

- A multiple purpose package for data analysis, graphics, model fitting, etc.
- Provides function `aregImpute` for multiple imputation using *additive regression, bootstrapping, and predictive mean matching*
  - continuous variables: restricted cubic splines
  - categorical variables: Fisher's optimum scoring method
  - each imputation uses a different bootstrap sample

MI

- Also implements the *chained equation approach*
- Implements *Bayesian* imputation models such as Bayesian generalized linear models
- Provide diagnostic tools for checking the fit of the imputation models

Visualization

Software package

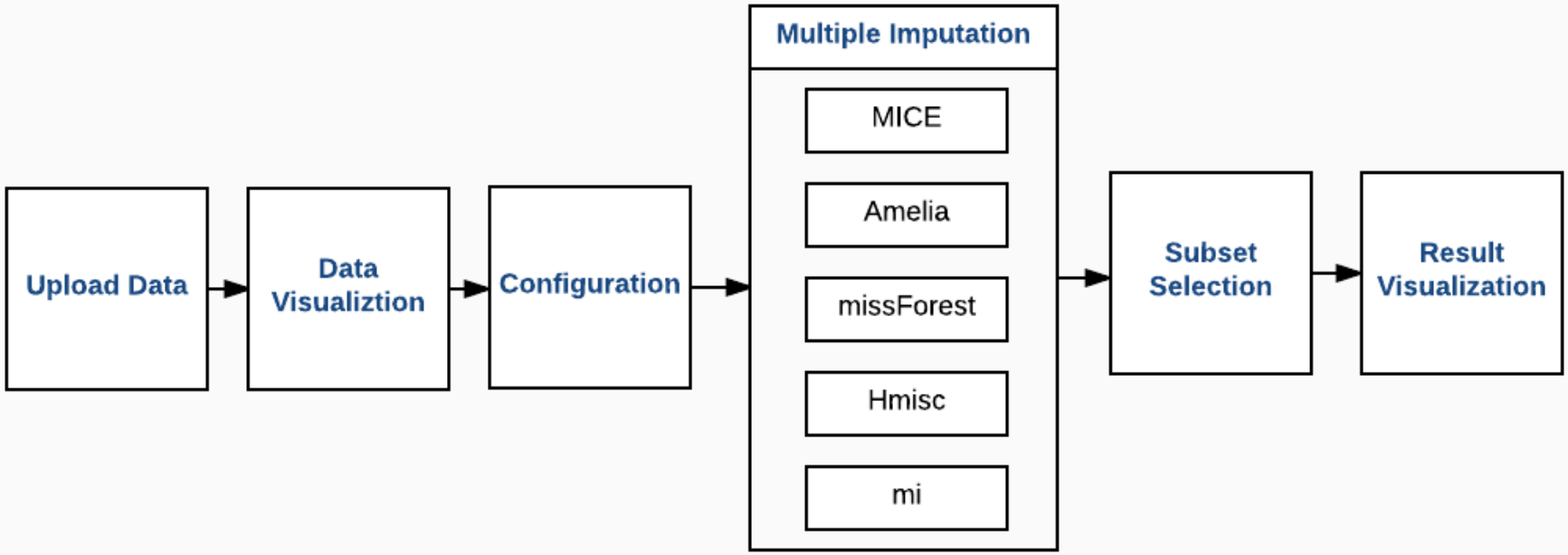
- *VIM*: Visualization and Imputation of Missing Values
- Different type of plots
  - Aggregation plot
  - Histogram
  - Spinogram
  - Marginal plot
  - Scatter plot
  - Jitter plot
  - Matrix plot
  - Spaghetti plot

Missuite

Shiny

- RStudio product
- A web application framework for R
- Turn R code into interactive web applications
- No HTML, CSS, or JavaScript knowledge required

Architecture



Access Missuite

- Demo on https://olssol.shinyapps.io/missuite/

Discussion

Statistical software for regulatory applications

- Communication
- Efficiency
- Reproducible research
- Education

Additional example

- IDEM
- Composite Endpoint Death Truncated Data Analysis
- Available on CRAN
- Demo on https://olssol.shinyapps.io/idem/

The End.