Cognizant

PROC MI – A New tool to deal with Laboratory Missing Data in ADaM

Jaganadha Rao Gundu

July 14, 2018 PhUSE SDE, Hyderabad

DISCLAIMER

- This presentation is based on publicly available information (including data relating to non-Cognizant products or approaches).
- The views presented are the views of the presenter, not necessarily those of Cognizant.
- These slides are intended for educational purposes only and for the personal use of the audience. These slides are not intended for wider distribution outside the intended purpose without presenter approval.
- The content of this slide deck is accurate to the best of the presenter's knowledge at the time of production.

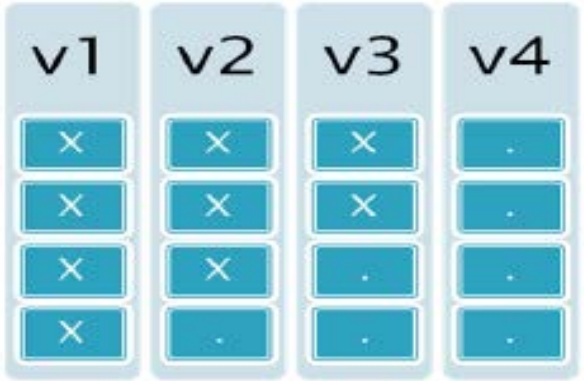
Agenda.

- Types of missing data
- How to deal with missing data
- What is Multiple Imputation
- Why Multiple Imputation
- Implementation of PROC MI procedure in ADLB
- Imputation diagnostics
- Concerns in MI
- Take home message
- Questions & Answers

Types of Missing Data Patterns

This is one of the most important considerations in planning imputation
Model for missing data

Arbitrary Pattern



Monotone Pattern

Types Of Missing data Mechanisms

Defined in order of severity, Rubin (1976)

- Missing completely at Random (MCAR)
- Missing at Random (MAR)
- Missing not at Random (MNAR)

We never really know for sure about the true nature of missing data, but we can make assumptions on how to treat datasets with missing data.

Causes of missing data

More Generally, Missing observations can arise due to

- Missed clinical appointments
- Equipment failures
- Data Entry
- Lost-to- Follow up
- Discontinue from the study
- Other circumstances that disrupt the intended measurements

How to Deal with Missing Data

- List wise Deletion
- Pairwise Deletion

- Single Imputation
  - Mean Substitution
  - Median Substitution
  - LOCF (Last Observation carry forward)
  - BOCF (Baseline Observation carry forward)

- Multiple Imputation

What is Multiple Imputation?

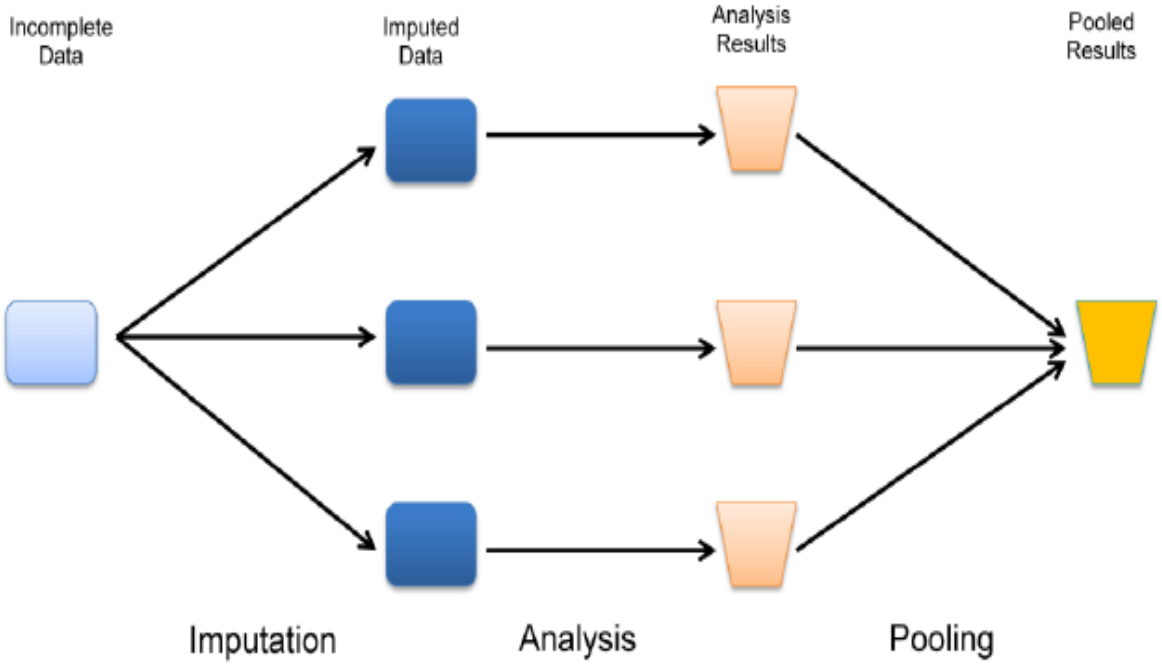
Multiple imputation involves creating multiple predictions for each missing value

SUBJID	WEIGHT	BASE	HBA1B	HBA1C
101	60.2	6.5	6.3	6.4
102	80.3	7.9	7.8	?
103	73	7	?	8.4

SUBJID	WEIGHT	BASE	HBA1B	HBA1C	IMPUTATION
101	60.2	6.5	6.3	6.4	1
102	80.3	7.9	7.8	7.3	1
103	73	7	6.6	8.4	1
101	60.2	6.5	6.3	6.4	2
102	80.3	7.9	7.8	9.1	2
103	73	7	6.1	8.4	2
101	60.2	6.5	6.3	6.4	3
102	80.3	7.9	7.8	5.6	3
103	73	7	7.5	8.4	3
101	60.2	6.5	6.3	6.4	4
102	80.3	7.9	7.8	8.1	4
103	73	7	6.9	8.4	4
101	60.2	6.5	6.3	6.4	5
102	80.3	7.9	7.8	4.9	5
103	73	7	7.7	8.4	5
--	--	--	--	--	m times

SUBJID	WEIGHT	BASE	HBA1B	HBA1C
101	60.2	6.5	6.3	6.4
102	80.3	7.9	7.8	7.0
103	73	7	6.96	8.4

MI

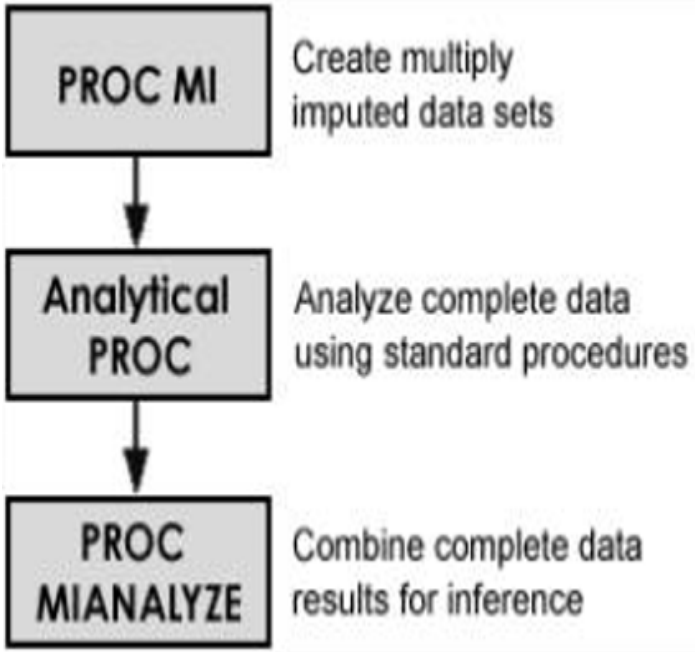


Steps in MI

Step 1: Imputation

Step 2: Analysis

Step 3: Pooling



Why Multiple Imputation ?

- It's so righteous dude!
- Minimize the bias
- Multiple imputation (MI) is an effective and responsible way to handle data which is missing at random (MAR)
- It gives more accurate Standard errors

MI Method recommendations

Choosing method depends on missing data pattern & Imputed variable type

Missing Data Pattern	Imputed Variable Type	Method	PROC MI Statement
Monotone	Continuous	Linear regression	MONOTONE REG
		Predictive mean matching	MONOTONE REGPMM
		Propensity score	MONOTONE PROPENSITY
	Binary/ordinal	Logistic regression	MONOTONE LOGISTIC
	Nominal	Discriminant function	MONOTONE DISCRIM
Arbitrary	Continuous	With continuous covariates:	
		MCMC monotone method	MCMC IMPUTE=MONOTONE
		MCMC full-data imputation	MCMC IMPUTE=FULL
		With mixed covariates:	
		FCS regression	FCS REG
		FCS predictive mean matching	FCS REGPMM
	Binary/ordinal	FCS logistic regression	FCS LOGISTIC
	Nominal	FCS discriminant function	FCS DISCRIM

Syntax: MI Procedure

```
PROC MI  DATA=      OUT=      SEED=   NIMPUTE= ;
  BY variables;
  CLASS variables;
  EM <options>;
  FCS <options>;
  FREQ variable;
  MCMC <options>;
  MNAR options;
  MONOTONE <options>;
     TRANSFORM transform (variables</options>) <…transform (variables</options>)>;
  VAR variables;
RUN;
```

Implementation of MI in ADLB

```
*Transpose to horizontal structure;
PROC TRANSPOSE DATA=adeff OUT=onepersub PREFIX=MIDI;
    BY subjid age sex trtp base paramcd param;
    ID avisitn;
    VAR aval;
RUN
    /* Multiple Imputation - 3-step process */

    *Step 1a - check missing data pattern (arbitrary vs monotone);
    PROC MI DATA=onepersub NIMPUTE=0;
        CLASS sex trtp;
        FCS;
        VAR sex trtp age base midi7 midi14 midi28 midi42 midi98;
    RUN;
 *Step 1b - select appropriate method given missing
            data pattern and variables to be imputed;
 PROC MI DATA=onepersub OUT=midi_mi SEED=1999 NIMPUTE=5
     ROUND = . . . . 1 1 1 1 1;
     CLASS sex trtp;
     FCS REG (midi7 midi14 midi28 midi42 midi98);
     VAR sex trtp age base midi7 midi14 midi28 midi42 midi98;
 RUN;
```

Implementation of MI in ADLB – Continued

```
*Step 2a - transpose back to vertical structure;
PROC TRANSPOSE DATA=midi_mi OUT=backtobds;
    BY _imputation_ subjid age sex trtp base paramcd param;
    VAR midi1 midi7 midi14 midi28 midi42 midi98;
RUN;


*Step 2b - Analysis step;
ODS OUTPUT diffs=diffbyvis(WHERE=(avisitn=_avisitn));
PROC MIXED DATA=adeffmi (WHERE=(avisitn > 1));
    BY imputeno;
    CLASS subjid trtp avisitn ;
    MODEL aval = base trtp avisitn trtp*avisitn ;
    REPEATED avisitn / SUBJECT=subjid TYPE=un;
    LSMEANS trtp*avisitn /DIFF CL;
RUN;
ODS OUTPUT CLOSE;
```

Implementation of MI in ADLB – Continued

```
*Step 3 - Pooling step;
PROC SORT DATA=formianalyze;
    BY avisitn imputeno;
RUN;

ODS OUTPUT PARAMETERESTIMATES=diffestvisit;
PROC MIANALYZE DATA=formianalyze;
    BY avisitn;
    MODELEFFECTS estimate;
    STDERR stderr;
RUN;
ODS OUTPUT CLOSE;
```

The **MIANALYZE** Procedure

(Variance Information)

Variance Information							
	Variance						
Parameter	Between	Within	Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
intercept	0.484830	11.599514	12.132827	4658	0.045977	0.044366	0.995583
write	0.000262	0.006014	0.006302	4311	0.047879	0.046134	0.995408
female	0.079066	1.264539	1.351512	2173.3	0.068778	0.065212	0.993521
math	0.000310	0.005865	0.006207	2973.8	0.058215	0.055648	0.994466
progcat1	0.239760	1.955043	2.218779	636.99	0.134900	0.121619	0.987984
progcat2	0.256880	2.339505	2.622073	774.97	0.120781	0.110059	0.989114

Relative Efficiencies

Lambda=% of missing ness
m=Number of imputations

	λ				
m	**10%**	**20%**	**30%**	**50%**	**70%**
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

The **MIANALYZE** Procedure
(Parameter Estimates)

Parameter Estimates										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
intercept	9.881994	3.483221	3.05323	16.71076	4658	8.523732	11.007546	0	2.84	0.0046
write	0.388897	0.079385	0.23326	0.54453	4311	0.346491	0.404590	0	4.90	<.0001
female	-2.424315	1.162545	-4.70413	-0.14450	2173.3	-2.830625	-2.059445	0	-2.09	0.0372
math	0.414454	0.078783	0.25998	0.56893	2973.8	0.384558	0.446180	0	5.26	<.0001
progcat1	2.332793	1.489557	-0.59224	5.25783	636.99	1.573968	3.121862	0	1.57	0.1178
progcat2	0.302432	1.619282	-2.87627	3.48113	774.97	-0.369425	0.868353	0	0.19	0.8519

Imputation Diagnostics

It is important to examine the output from proc mianalyze

Variance Information

Variance Between (V_B)
Variance Within (V_W)
Variance Total (V_T)
Degrees of Freedom (DF)
Relative Increases in Variance (RIV/RVI)
Fraction of Missing Information (FMI)
Relative Efficiency

Parameter Estimates

Estimate
Std Error
95 % CI
DF
Minimum
Maximum
Theta0

Concerns !

- Is there any better method available other than MI
- Isn’t MI just making up data
- How to select the number of imputations
- What by variables should be used
- Options, Statements in MI
- Should I include dependent variables, auxiliary variables in model
- How much missing can I have and still get good estimates using MI
- Which stat programs should be used
- The type of imputation
- How to justify for choosing a particular imputation method

Take home message !

- MI is effective and is always superior to any of the single imputation
- There are several decisions to be made before performing a MI
- MI is one of the best tools for researchers to address the very common problems of missing data
- Produce unbiased estimates

References

- https://stats.idre.ucla.edu/sas/seminars/multiple-imputation-in-sas
- http://support.sas.com/resources/papers/proceedings12/319-2012.pdf
- https://www.lexjansen.com/phuse/2014/sp/SP04.pdf
- https://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf
- https://pharmasug.org/proceedings/2017/SP/PharmaSUG-2017-SP01.pdf

Acknowledgements

I would like to thank Cognizant colleagues, especially Sreekar Mateti, Saibaba and Jayachand Kotturu for their Support.

Cognizant

Thank you