



A Flexible Approach to Multi-Source Data Imputation: How it Works and Expected Results



James Hunter MSc, OCS Life Sciences, the Netherlands

Introduction



- I am a: Statistician / Programmer.
- The idea:
 - To impute missing data in multiple ADaM datasets simultaneously.
 - All parameters from each dataset would be used as covariates to impute data in each dataset.
- Issues:
 - Large amount of data.
 - Many parameters to impute.
 - Covariates used in imputation must vary based on the parameter.
- The solution: In SAS, use Macros and Proc MI to create a flexible imputation program.
- The results: A syntax to impute large amounts of data, using many sources for covariates, producing imputed data with a strong fit to the original data.

Imputation



- Imputation: A process of replacing missing data with substituted values.
- Types of imputation:
 - Last Observation Carried Forward (LOCF).
 - Worst Observation Carried Forward (WOCF).
 - Mean substitution.
 - Regression model.
 - Single imputation.
 - Multiple imputation.
- When to use imputation:
 - % of data missing.
 - Type of missing data;
 - Missing Completely At Random (MCAR).
 - Missing At Random (MAR).
 - Missing Not At Random (MNAR).
 - Clinical trial data assumes MNAR.
 - There is no universally agreed method of imputing MNAR data.





Proc MI

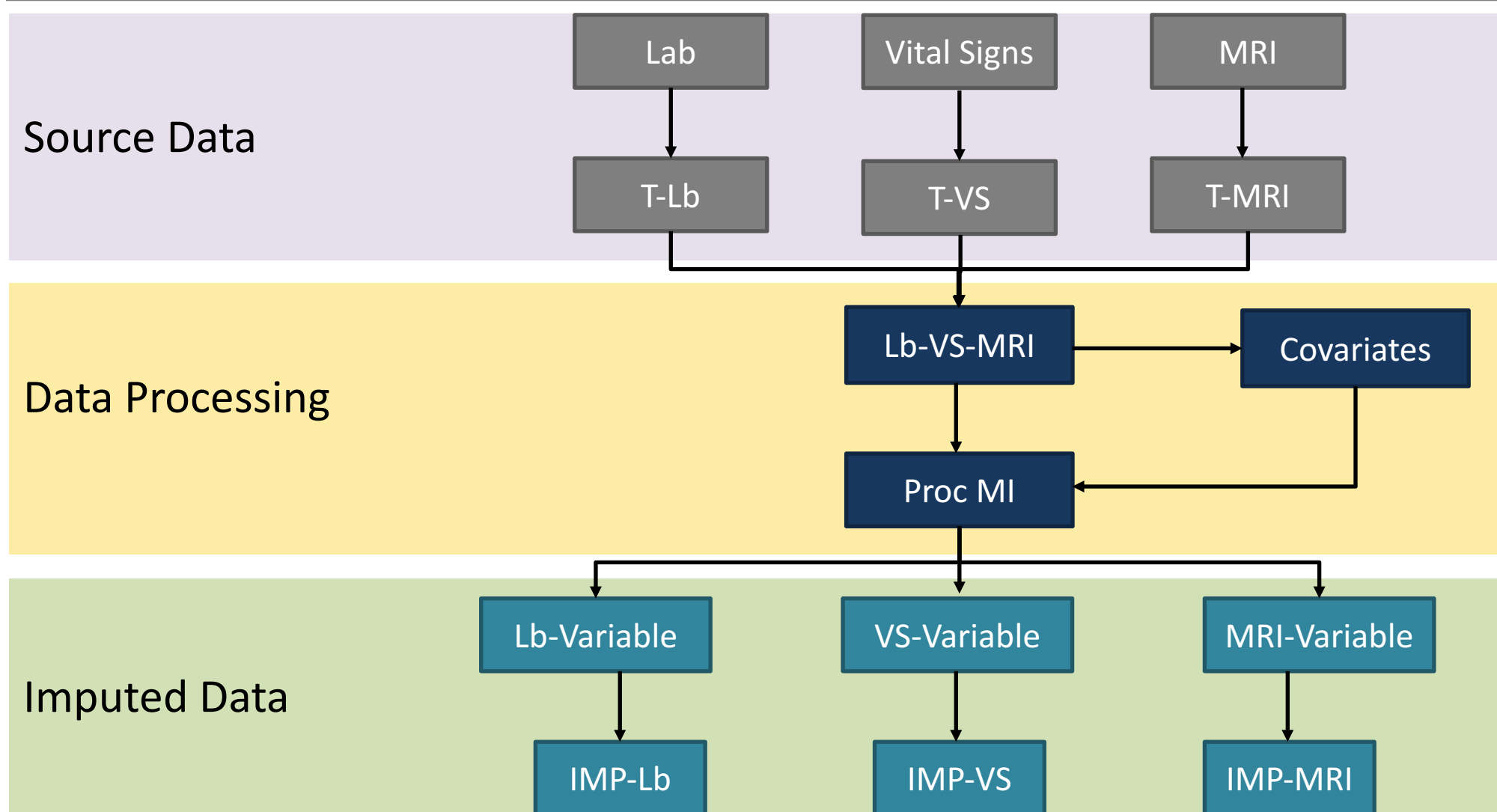
The proc MI (Multiple Imputation) procedure:

- Performs single or multiple imputation of missing data.
- Has a variety of data imputation methods;
 - Discrim
 - Logistic
 - Propensity
 - Regpredmeanmatch
- Proc MI assumes data is MAR.

```
proc mi data=work.Indata out=work.Outdata nimpute=1 seed= 8546 simple;  
  class sex treatment country;  
  fcs regpredmeanmatch (DataToImpute= age sex treatment country .....);  
  var DataToImpute age sex treatment country .....;  
run;
```



Multi-Source Proc MI



ADaM Data



- Purpose: A standard for reporting clinical trial data in a ready for analysis format.
- Structure: Follows the Basic Data Structure (BDS), one record per row.
- Example: Vital Signs (ADVS).

USUBJID	TRTA	PARAM	PARAMCD	AVISIT	AVISITN	AVAL
001-001	Placebo	Weight (kg)	WGHT	Screening	-1	70.5
001-001	Placebo	Weight (kg)	WGHT	Baseline	0	71.0
001-001	Placebo	Heart rate (bpm)	HRT	Screening	-1	50
001-001	Placebo	Heart rate (bpm)	HRT	Baseline	0	56
001-002	Drug A	Weight (kg)	WGHT	Screening	-1	68.0
001-002	Drug A	Weight (kg)	WGHT	Baseline	0	
001-002	Drug A	Heart rate (bpm)	HRT	Screening	-1	46
001-002	Drug A	Heart rate (bpm)	HRT	Baseline	0	52

Imputation Data Structure

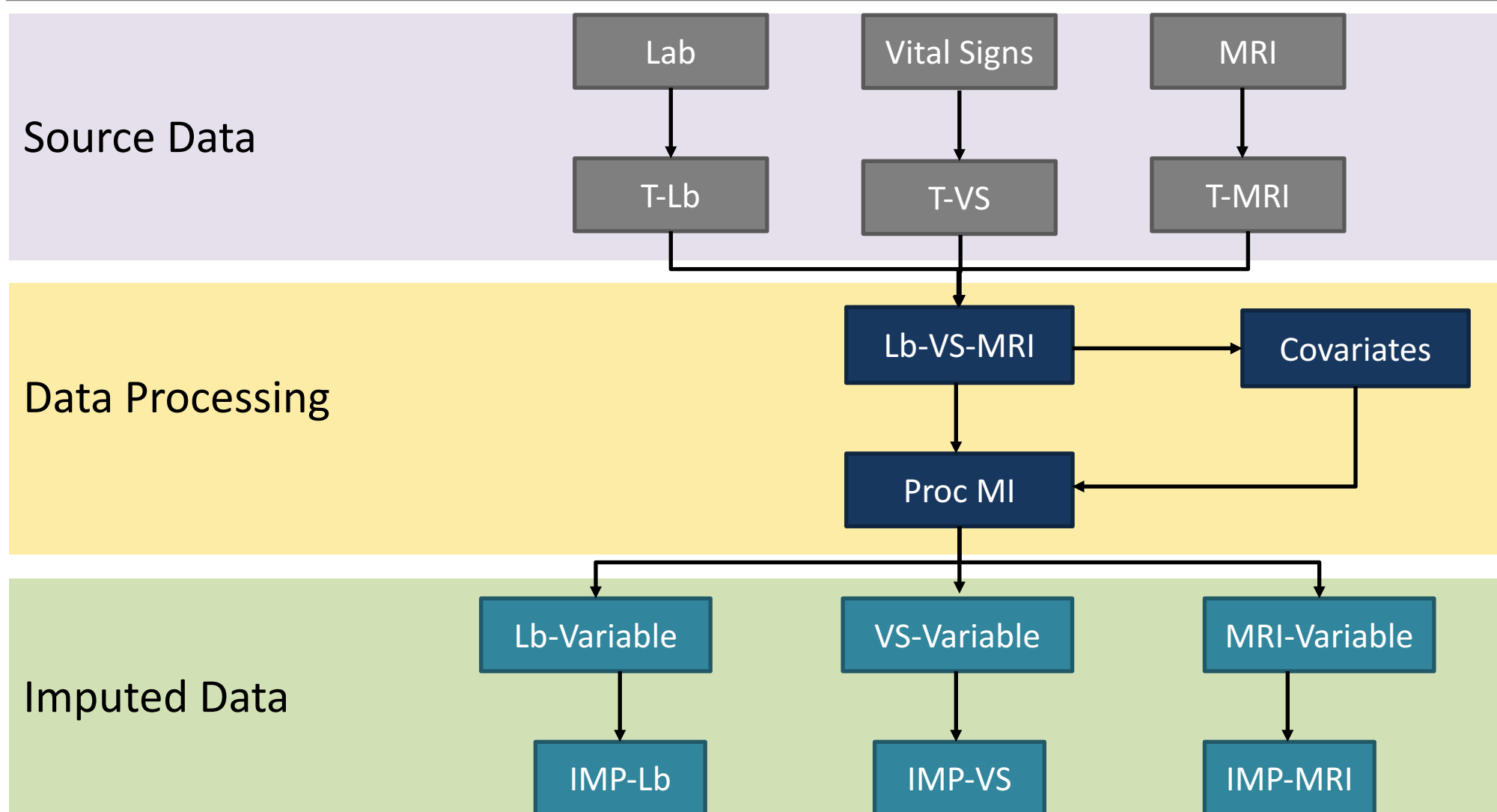


- Purpose: Readable data for Proc MI.
- Structure: Transposed ADaM data.
- Example: Transposed Vital Signs (T-VS).

USUBJID	TRTA	AVISIT	AVISITN	WGHT	HRT
001-001	Placebo	Screening	-1	70.5	50
001-001	Placebo	Baseline	0	71.0	56
001-002	Drug A	Screening	-1	68.0	46
001-002	Drug A	Baseline	0		52



Multi-Source Proc MI



Covariates



- Purpose: Directs Proc MI.
- Specifies:
 - Data to receive imputation.
 - Covariates to use in imputation.

DataToImpute	Covariate 1	Covariate 2	Covariate 3	Covariate N...
WGHT	HRT	TRTA	AVISITN	...
HRT	WGHT	TRTA	AVISITN	...



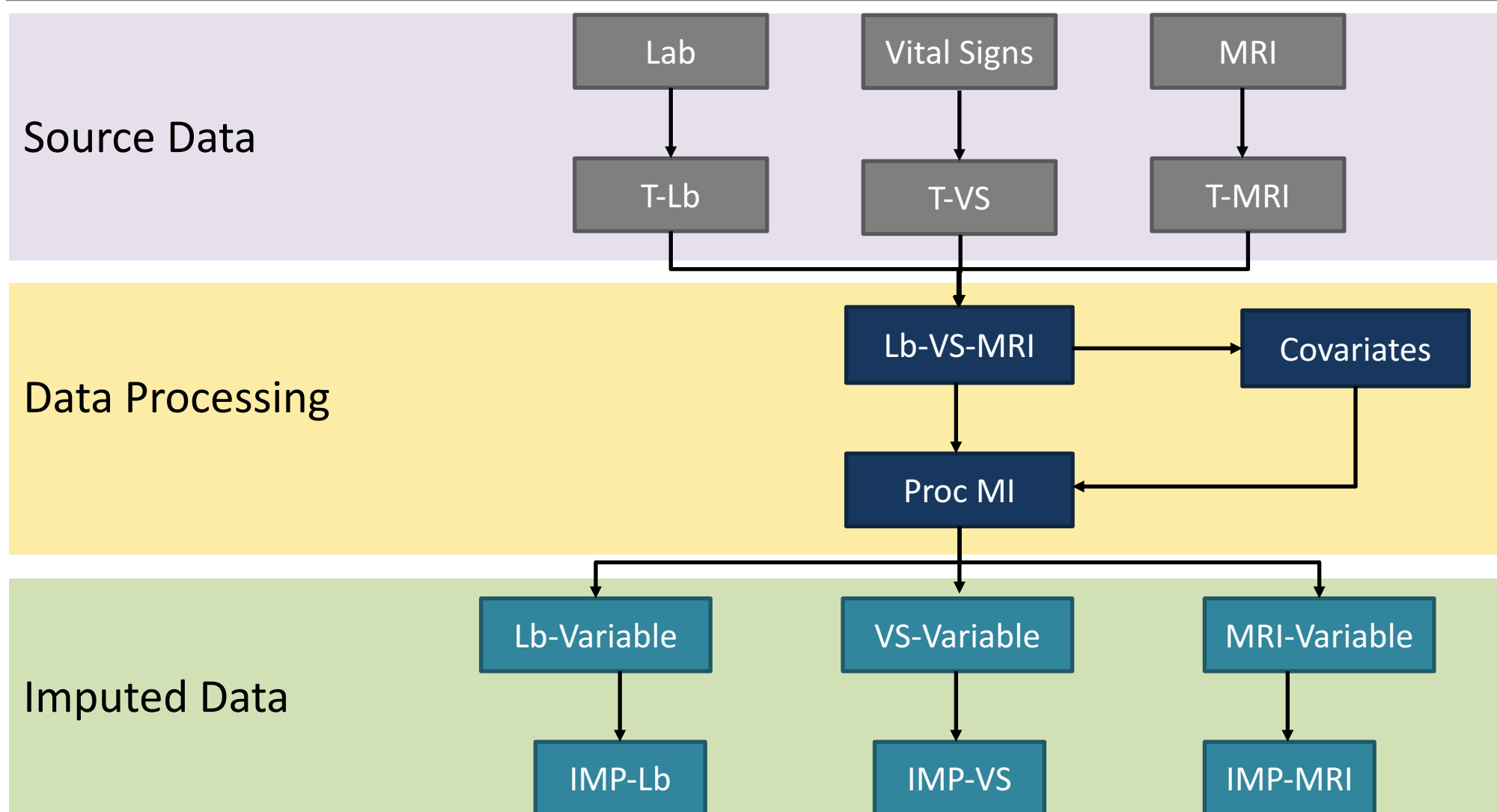
Multi-Source Proc MI

- Indata = Lb-VS-MRI.
- &DataToImpute. = Covariates.
- &list. = Covariates.
- Outdata&x. = Lb-Variable / VS-Variable / MRI-Variable.

```
proc mi data=work.Indata out=work.Outdata&x. nimpute=1 seed= 8546 simple;  
  class sex treatment country;  
  fcs regpredmeanmatch (&DataToImpute.= age sex treatment country &list.);  
  var age sex treatment country &DataToImpute. &list.;  
run;
```



Multi-Source Proc MI



Outputs



- Number: One dataset per parameter imputed.
- Example: VS-Variable - Weight (WGHT).

USUBJID	TRTA	AVISIT	AVISITN	WGHT
001-001	Placebo	Screening	-1	70.5
001-001	Placebo	Baseline	0	71.0
001-002	Drug A	Screening	-1	68.0
001-002	Drug A	Baseline	0	68.5

Outputs



- Number: One dataset per parameter imputed.
- Example: VS-Variable - Weight (WGHT).
- Dataset: Imp-VS.

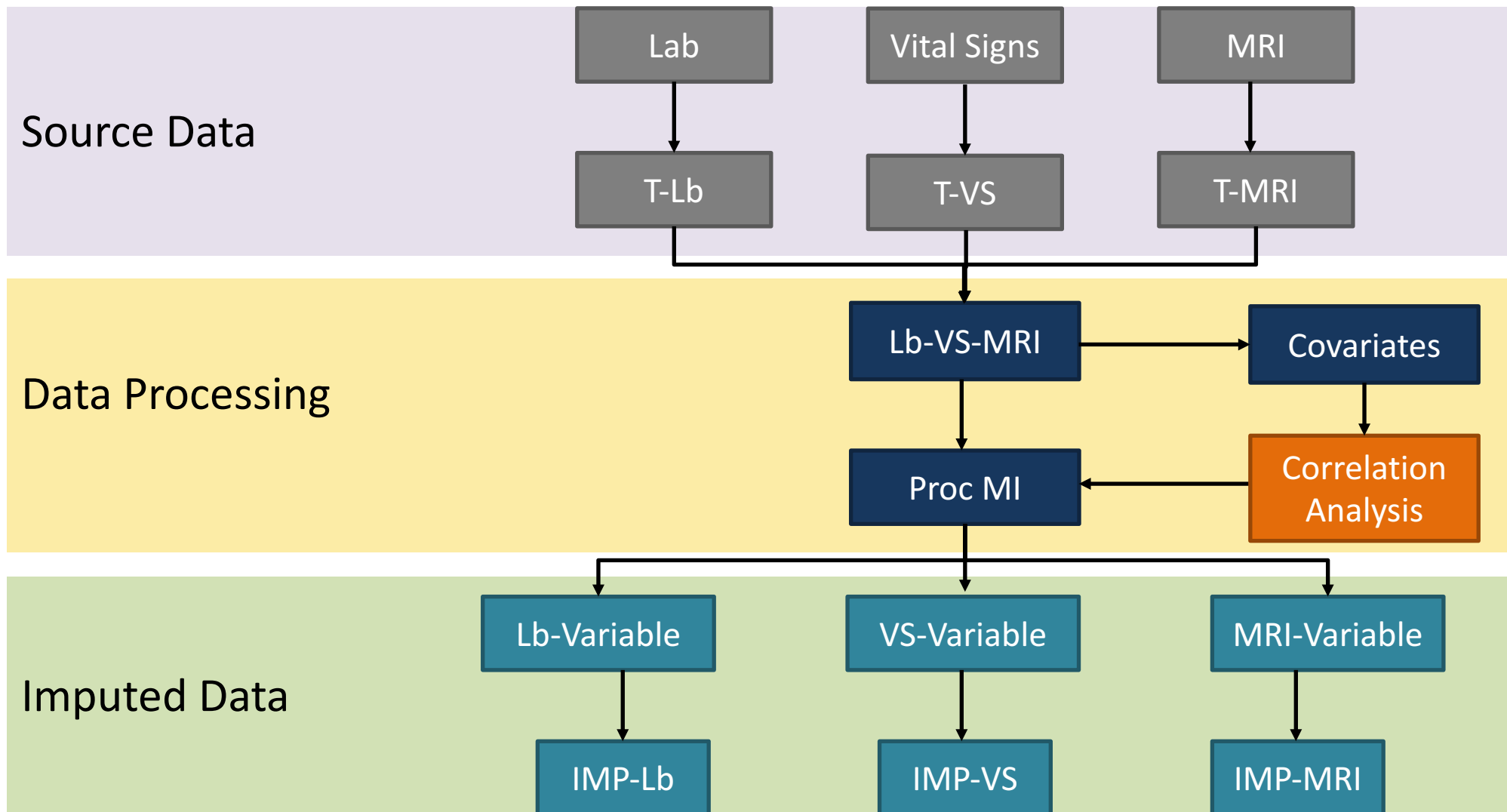
USUBJID	TRTA	PARAM	PARAMCD	AVISIT	AVISITN	AVAL
001-001	Placebo	Weight (kg)	WGHT	Screening	-1	70.5
001-001	Placebo	Weight (kg)	WGHT	Baseline	0	71.0
001-001	Placebo	Heart rate (bpm)	HRT	Screening	-1	50
001-001	Placebo	Heart rate (bpm)	HRT	Baseline	0	56
001-002	Drug A	Weight (kg)	WGHT	Screening	-1	68.0
001-002	Drug A	Weight (kg)	WGHT	Baseline	0	68.5
001-002	Drug A	Heart rate (bpm)	HRT	Screening	-1	46
001-002	Drug A	Heart rate (bpm)	HRT	Baseline	0	52



Results

- This method was successfully used to:
 - Impute data from 4 ADaM datasets, containing ≈ 100 parameters.
 - Using ≈ 40 out of 100 parameters as covariates.
 - Imputed data showed good trends compared to source data.
- This method is applicable for:
 - Single or multiple imputation.
 - Continuous or categorical data.
- Observations:
 - Proc MI will ignore covariates that provide the same effect as combinations of other covariates.
 - Proc MI supplied with large amounts of covariates will reach the limit of data a single dataset can contain. Approximately 2 billion observations for a 32-bit system.

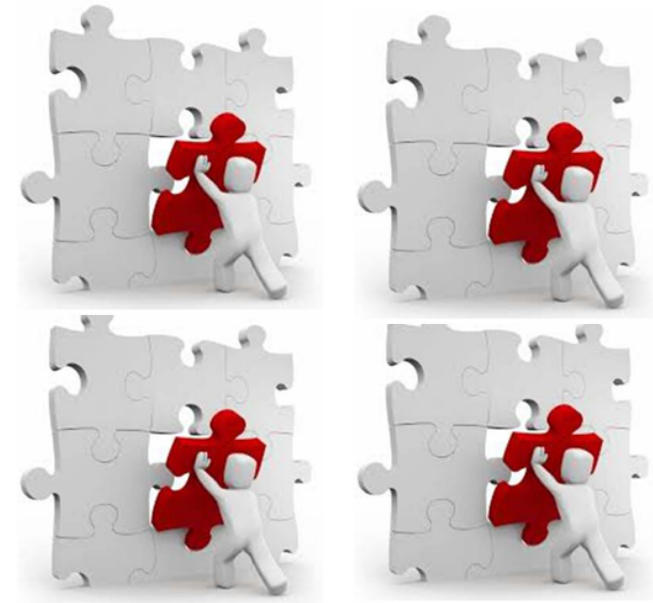
Fine-Tuning



Conclusion



- It is possible:
 - To impute missing data from multiple sources simultaneously.
 - For large amounts of data.
 - With many covariates.
 - By using a combination of Macros and Proc MI.
 - To achieve reliable results with good data trends.
 - Correlation analysis on covariates should be included.



Questions

