

APAC 2026

From Missing to Meaningful: Validating Multiple Imputation Techniques

Karthick Jayaraman
SENSAN Biosciences Pvt. Ltd.



- Missing Data Patterns and Mechanisms
- Steps in Multiple imputation
- Imputation methods
- MI implementation in SAS
- MI implementation in R
- Summary & Conclusion

Monotone missing pattern

- If a variable is missing for an individual, all subsequent variables in that order are also missing.

SUBJID	BASE	V1	V2	V3
1001	X	X	X	Missing
1002	X	X	Missing	Missing
1003	X	Missing	Missing	Missing

Arbitrary missing pattern

- No specific pattern or sequence; data may be missing anywhere.

SUBJID	BASE	V1	V2	V3
1001	X	Missing	X	X
1002	X	X	Missing	X
1003	X	Missing	X	Missing

MCAR (Missing Completely at Random)

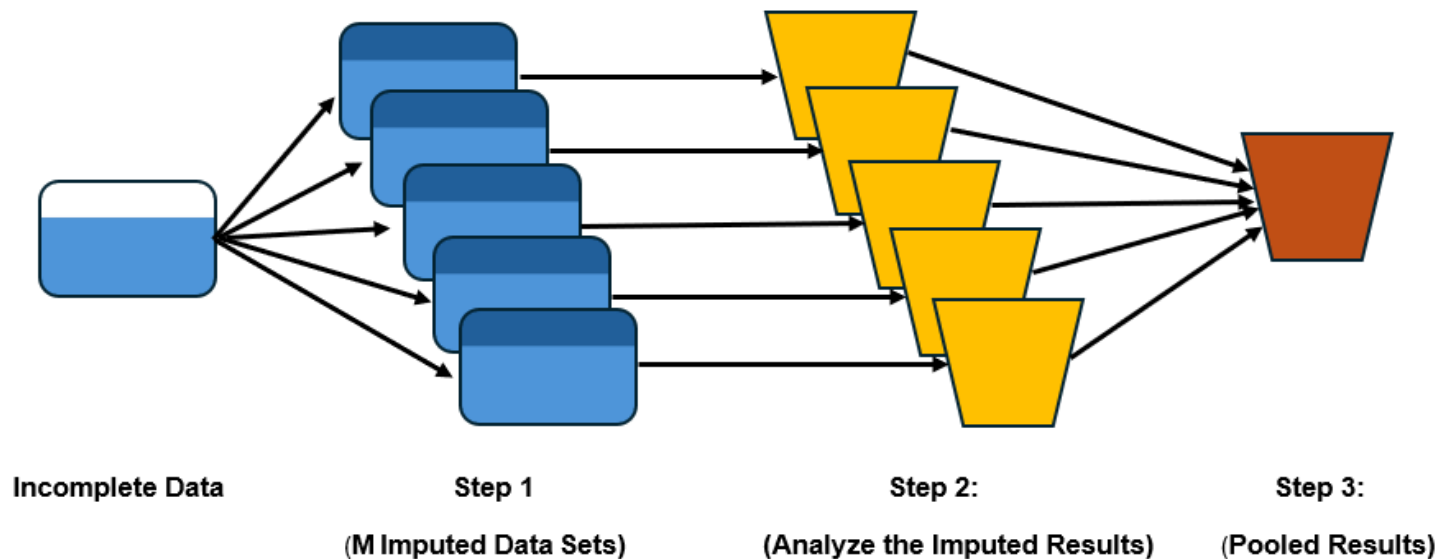
- Missingness is unrelated to any observed or unobserved data.

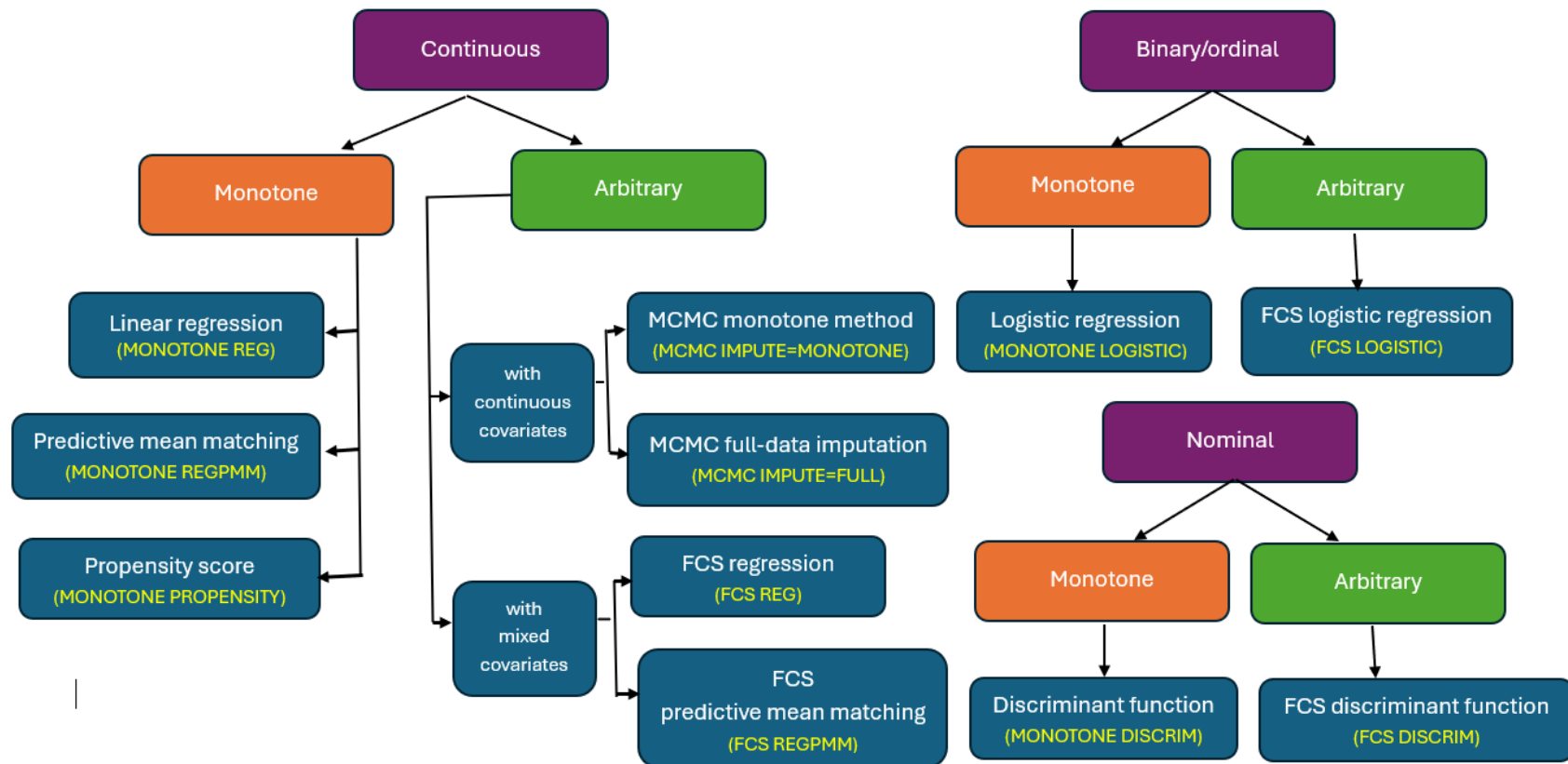
MAR (Missing at Random)

- Missingness is related to observed variables but not to the value of the missing data.

MNAR (Missing Not at Random)

- Missingness is related to the unobserved (missing) data itself.





Step 0: Examine missing data pattern

```
/* Transpose input data to be in a horizontal format */
```

```
PROC TRANSPOSE DATA= adef
```

```
OUT= indset;
```

```
BY usubjid trt;
```

```
ID avisit;
```

```
VAR aval;
```

```
RUN;
```

Obs	SUBJID	TRT	BASELINE	WK_12	WK_24	WK_36	WK_48	WK_60	WK_72
1	1001	Active	30	.	30	.	.	.	.
2	1002	Active	26	26	25	27	.	.	.
3	1003	Placebo	27	28	.	.	.	.	.
4	1004	Placebo	33	33	33	33	.	34	33

```
/* Examine missing data patterns */
```

```
PROC MI DATA = indset NIMPUTE = 0;
```

```
VAR baseline wk_12 -- wk_72;
```

```
RUN;
```

Missing Data Patterns									
Group	BASELINE	WK_12	WK_24	WK_36	WK_48	WK_60	WK_72	Freq	Percent
1	X	X	X	X	X	X	X	54	88.52
2	X	X	X	X	X	.	.	2	3.28
3	X	X	X	X	.	X	X	2	3.28
4	X	X	X	X	.	.	.	1	1.64
5	X	X	.	.	.	.	.	1	1.64
6	X	.	X	.	.	.	.	1	1.64

*** The output indicates missing data pattern is Arbitrary*

Step 1: Imputation – assuming missing values are MNAR

```
/* MCMC method to impute intermediate missing to produce 5 datasets with  
monotone missing pattern */
```

```
PROC MI data= indset OUT = indset_mono SEED=12345 NIMPUTE = 5  
  MINIMUM = . 0 0 0 0 0 0 MAXIMUM = . 36 36 36 36 36 36;  
  BY trt;  
  MCMC CHAIN = multiple IMPUTE = monotone;  
  VAR baseline wk_12 wk_24 wk_36 wk_48 wk_60 wk_72;  
RUN;
```

IMPUTATION	SUBJID	TRT	BASELINE	WK_12	WK_24	WK_36	WK_48	WK_60	WK_72
1	1001	Active	30	29.7737	30	.	.	.	.
2	1001	Active	30	31.3339	30	.	.	.	.
3	1001	Active	30	31.5557	30	.	.	.	.
4	1001	Active	30	29.6653	30	.	.	.	.
5	1001	Active	30	28.2933	30	.	.	.	.

Step 1: Imputation – assuming missing values are MNAR (cont'd)

/ Impute missing ending data using monotone regression using CCMV assuming that missing values are MNAR */*

```
PROC MI DATA= indset_mono OUT=imp_indset_monoreg NIMPUTE = 1 seed = 12345
  MINIMUM = . 0 0 0 0 0 0 MAXIMUM = . 36 36 36 36 36 36;
  BY _IMPUTATION_;
  MONOTONE REG;
  MNAR MODEL (wk_12 wk_24 wk_36 wk_48 wk_60 wk_72/ modelobs = CCMV);
  VAR baseline wk_12 wk_24 wk_36 wk_48 wk_60 wk_72;
RUN;
```

IMPUTATION	SUBJID	TRT	BASELINE	WK_12	WK_24	WK_36	WK_48	WK_60	WK_72
1	1001	Active	30	29.7737	30	31.8103	33.4266	36.0000	33.7092
2	1001	Active	30	31.3339	30	29.7793	30.7383	33.8649	34.0435
3	1001	Active	30	31.5557	30	32.6640	30.5264	32.8720	28.0543
4	1001	Active	30	29.6653	30	29.3554	29.7644	30.5748	30.0389
5	1001	Active	30	28.2933	30	29.2694	31.5780	29.2217	31.9757

Step 2: Analyze the multiple imputed datasets

```
/* Perform MMRM analysis for each imputation independently */
```

```
PROC MIXED DATA = indset1;  
  BY _IMPUTATION_;  
  CLASS usubjid trtn avisitn;  
  MODEL chg_imp = trtn base avisitn age trt*avisitn / SOLUTION;  
  REPEATED avisitn / TYPE=UN SUBJECT=usubjid;  
  LSMEANS trtn*avisit / DIFF = control('0') cl;  
RUN;
```

IMPUTATION	EFFECT	TRTN	AVISITN	_TRTN	_AVISITN	ESTIMATE	STDERR	DF	TVALUE	PROBT	ALPHA	LOWER	UPPER
1	ESTIMATE	1	9	0	9	-1.1616	0.5302	115	-2.19	0.0305	0.05	-2.2118	-0.1113
2	ESTIMATE	1	9	0	9	-1.1644	0.5364	115	-2.17	0.0320	0.05	-2.2269	-0.1019
3	ESTIMATE	1	9	0	9	-1.3932	0.5384	115	-2.59	0.0109	0.05	-2.4596	-0.3268
4	ESTIMATE	1	9	0	9	-1.1873	0.5134	115	-2.31	0.0225	0.05	-2.2042	-0.1704
5	ESTIMATE	1	9	0	9	-1.2862	0.5206	115	-2.47	0.0150	0.05	-2.3174	-0.2549

Step 3: Pooling Results

```
/* Combine estimates of each imputation using Rubin's Rule */
```

```
PROC MIANALYZE parms = lsdiffs;  
  CLASS trtn;  
  MODELEFFECTS estimate;  
  STDERR stderr;  
  ODS OUTPUT parameterestimates=lsdiffs_out;  
RUN;
```

Parameter Estimates (5 Imputations)									
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0 Pr > t
estimate	-1.238527	0.539190	-2.29587	-0.18118	2321.1	-1.393190	-1.161558	0	-2.30 0.0217

Step 0a: Load and explore the data

#Install and Load the Required Package

```
> install.packages(c("mice", "mmrm", "emmeans"))  
> library(mice) # for multiple imputation pooling  
> library(mmrm) # for MMRM analysis
```

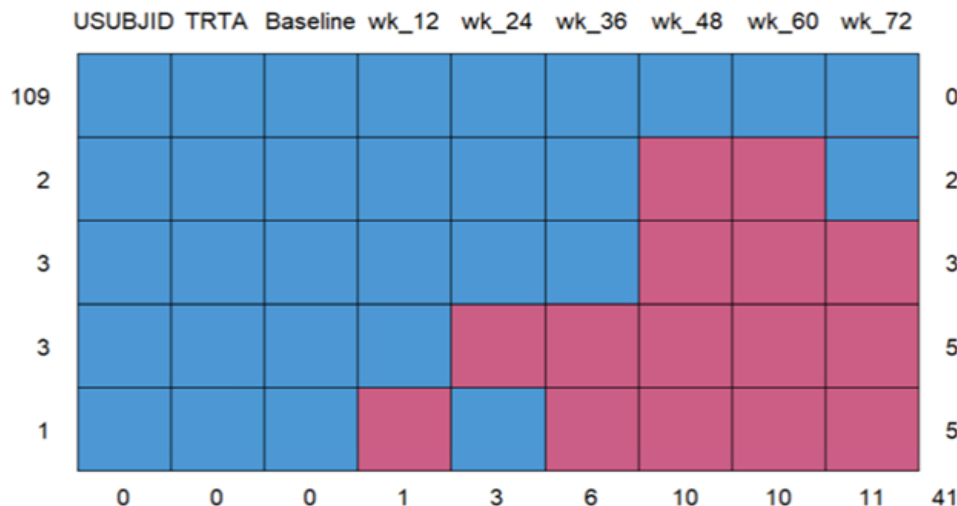
#Import the dataset

```
> indset <- read_sas("../ADaM/adeff.sas7bdat")
```

Step 0b: Examine missing data pattern

```
# Using mice for visualizing missingness pattern
```

```
> md.pattern(wide_data)
```



Step 1: Impute Missing Data

#Perform Multiple Imputation

```
> imp <- mice(wide_data,  
             m=5,  
             method = "pmm",  
             seed = 12345)
```

Variable type	Imputation Method
Continuous	Predictive mean matching (pmm)
Binary	Logistic regression (logreg)
Nominal	Polytomous Regression (polyreg)
Ordinal	Proportional Odds Logistic Regression (polr)

#Check imputation methods used in each variable

```
> imp$method
```

```
> imp$method  
USUBJID      TRTA Baseline      wk_12      wk_24      wk_36      wk_48      wk_60      wk_72  
      ""      ""      ""      "pmm"      "pmm"      "pmm"      "pmm"      "pmm"      "pmm"
```

** Customize imputation methods for individual variables

```
method=c("", "logreg", "", "pmm", "pmm", "pmm", "pmm", "", "")
```

Step 2: Analyze Imputed Data

MMRM Analysis for each imputed dataset

```
> model_fit <- with(imp_long_data,  
  mrm(formula=chg_imp ~ TRTA + BASE + AVISITN + TRTA * AVISITN  
    + us(AVISITN | USUBJID))  
> summary(model_fit)
```

Filter							
imp	contrast	AVISITN	estimate	SE	df	t.ratio	p.value
1	Active - Placebo	72	-1.381307	0.4329601	426.4050	-3.190380	0.001525819
2	Active - Placebo	72	-1.229529	0.4203232	431.3233	-2.925199	0.003623998
3	Active - Placebo	72	-1.278652	0.4215958	428.1106	-3.032885	0.002569600
4	Active - Placebo	72	-1.317373	0.4308671	419.6259	-3.057492	0.002374762
5	Active - Placebo	72	-1.336878	0.4257732	426.9728	-3.139884	0.001807542

Step 3: Pool Results

```
# Combine results of each analysis
```

```
> pooled_results <- pool(model_fit)
```

```
> summary(pooled_results)
```

Filter							
term	AVISITN	estimate	std.error	conf.low	conf.high	df	p.value
Active - Placebo	72	-1.308748	0.430991	-2.153593	-0.4639028	8653.782	0.002399741

- Always examine missing data patterns before applying imputation.
- Choose suitable imputation methods based on missing pattern and variable type.
- Use sufficient imputations to improve stability.
- MINIMUM= and MAXIMUM= options to specify logical lower and upper bounds for imputed values.
- Analyze each imputed dataset independently using BY `_IMPUTATION_` statement in the statistical procedure.
- Combine each parameter estimates and standard errors using Rubin's rules.

Thank you!

Karthick Jayaraman

Karthickj@sensanbio.com



The Global Healthcare Data Science Community

Contact Channels

📞 UK +44 1843 609600
✉ office@phuse.global
🌐 phuse.global

Social Media

🐦 [@phusetwitta](https://twitter.com/phusetwitta)
📘 [/phusebook](https://www.facebook.com/phusebook)
▶ [/phusetube](https://www.youtube.com/channel/UCphusetube)
🌐 [/company/phuse](https://www.linkedin.com/company/phuse)