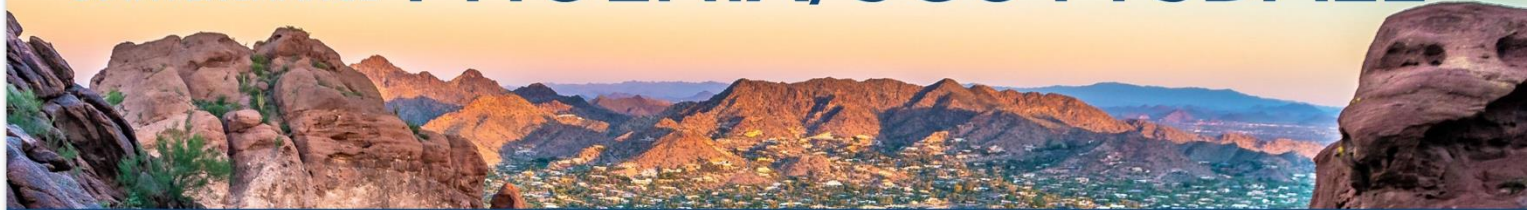




2024 CDISC + TMF
US INTERCHANGE

PHOENIX/SCOTTSDALE



23-24 OCTOBER: CONFERENCE & EXPO | 21, 22, 25 OCTOBER: TRAININGS

**Advanced Approaches to Missing Data in Rare Disease Studies -
*Using SAS and ADaM Datasets to Facilitate LOCF, MMRM and MI***

Presented by Olivia Hwang
Biostatistical Programming Manager
Amgen

Meet the Speaker

Olivia Hwang

Title: Biostatistical Programming Manager

Organization: Amgen

Olivia Hwang holds a bachelor's degree in Healthcare Administration and a master's degree in Biomedical Informatics from Taipei Medical University. She has six years of experience as a statistical programmer, having worked for IQVIA, ICON, and PPD. Currently, she is a Biostatistical Programming Manager at Amgen.





Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*
- *The author(s) have no real or apparent conflicts of interest to report.*



Outline

1. Challenges with missing data in rare disease studies
2. Introduce three types of missing data
3. Common methodologies for imputing missing data:
 1. *Last Observation Carried Forward (LOCF)*
 2. *Mixed Model for Repeated Measures (MMRM)*
 3. *Multiple Imputation (MI)*
4. How ADaM and intermediate datasets could facilitate this type of analysis and improve traceability

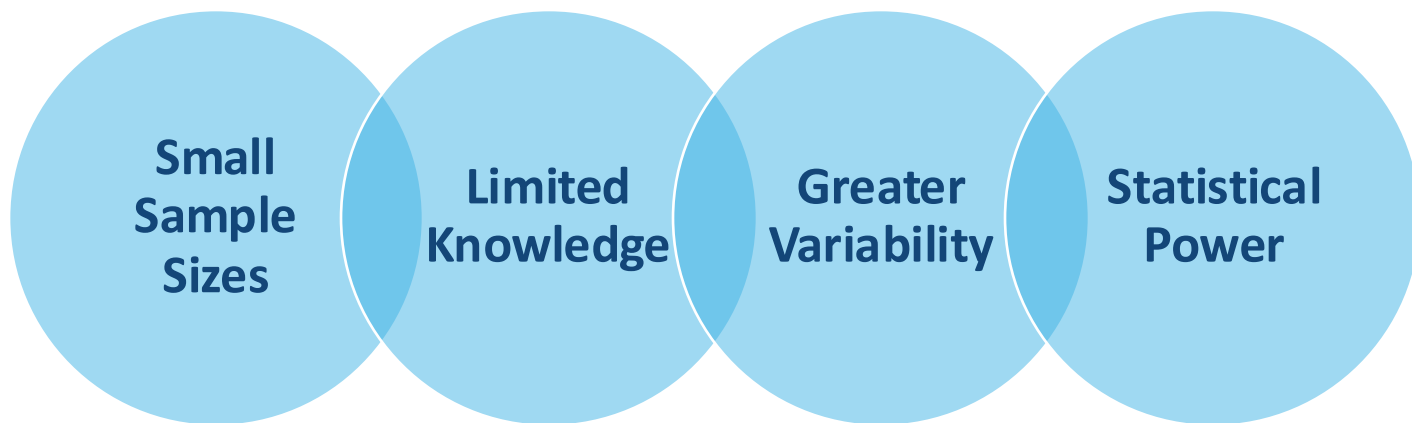


Introduction

- A longitudinal study is a type of research method that involves repeated observations of the same subjects over a long period of time.
- Six common reasons why patients withdraw from longitudinal studies:
 - (1) Recovery
 - (2) lack of improvement
 - (3) treatment-related side-effects
 - (4) unpleasant study procedures
 - (5) intercurrent health problems
 - (6) external factors unrelated to the trial
- COVID-19 pandemic has led to disruptions in clinical trial operations, further complicating data collection and the issue of missing data.

Unique challenges of Rare disease studies

The significance of missing data is often magnified for *rare disease studies* compared to other therapeutic areas (TA) due to several unique challenges.



The presence of missing data can hinder the ability to demonstrate efficacy or safety, potentially impacting drug approval and patient access.

Missing Data Mechanisms

MCAR

(Missing Completely At Random)

MAR

(Missing At Random)

MNAR

(Missing Not At Random)

MCAR

(Missing Completely At Random)

- **No dependency** on observed or unobserved variables.

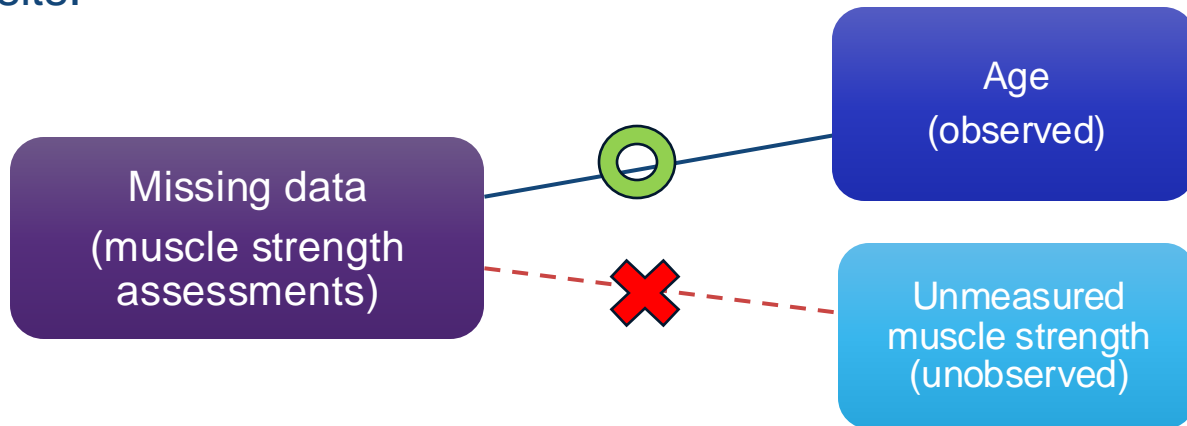
Example: A patient exits a study for reasons unrelated to the treatment or their health condition such as moving.

- MCAR is frequently impractical.

MAR

(Missing At Random)

- Missing data is **associated with observed information** but NOT with data that hasn't been observed.
- Example: In a **rare muscle disease**, where patients are regularly assessed for their muscle strength. Some older patients are more likely to miss follow-up visits.



MNAR

(Missing Not At Random)

- This category applies when missingness is associated with **unobserved data**.
- Example: In a rare neurological disorder study. One of the key outcomes you're measuring is **cognitive function**. Patients are required to attend follow-up visits every three months.



Patients with worsening
cognitive function



more likely to miss their
follow-up visits

Missing data
(cognitive function
scores)

Unmeasured cognitive function
(unobserved)

Common Methods for Analyzing Missing Data

LOCF

- Last Observation Carried Forward

MMRM




- Mixed Model Repeated Measures

MI

- Multiple imputation

Example

Table x.x.x.x Analysis of Test1 Change from Baseline at Week 28
(Full Analysis Set)

Endpoint	Drug A	Drug B	Placebo
Visit	(N= xx)	(N=xx)	(N=xx)
Statistic			
Test1			
Change from Baseline at Week 4 			
Treatment Comparison (Drug - Placebo)			
n	xx	xx	xx
LS Mean (SE)	xx.x (x.xx)	xx.x (x.xx)	xx.x (x.xx)
LS Mean Difference (SE)	xx.x (x.xx)	xx.x (x.xx)	
90% CI for Difference	(xx.x, xx.x)	(xx.x, xx.x)	
p-value	0.xxxxx	0.xxxxx	
Change from Baseline at Week 16 			
Treatment Comparison (Drug - Placebo)			
n	xx	xx	xx
LS Mean (SE)	xx.x (x.xx)	xx.x (x.xx)	xx.x (x.xx)
LS Mean Difference (SE)	xx.x (x.xx)	xx.x (x.xx)	
90% CI for Difference	(xx.x, xx.x)	(xx.x, xx.x)	
p-value	0.xxxxx	0.xxxxx	
Change from Baseline at Week 28 			
Treatment Comparison (Drug - Placebo)			
N	xx	xx	xx
LS Mean (SE)	xx.x (x.xx)	xx.x (x.xx)	xx.x (x.xx)
LS Mean Difference (SE)	xx.x (x.xx)	xx.x (x.xx)	
90% CI for Difference	(xx.x, xx.x)	(xx.x, xx.x)	
p-value	0.xxxxx	0.xxxxx	



LOCF (Last Observation Carried Forward)

Simple imputation method where missing values are replaced with the most recent non-missing value for each subject.

The LOCF approach is simple, but it makes two strong assumptions that:

- (1) missing data follow MCAR (*missing completely at random*)
- (2) the outcome of a participant does not change after drop out.

A snapshot of ADaM Data Set, ADEFF

Change from Baseline

Planned
Treatment:
1=Drug A,
2=Drug B,
3=Placebo

TRTP	USUBJID	PARAMCD	STRATA1	STRATA2	AVISITN	CHG
1	001	TEST1	No	< 20%	4	-0.8
1	001	TEST1	No	< 20%	16	-3.8
1	001	TEST1	No	< 20%	28	-3
1	002	TEST1	Yes	< 20%	4	-2.2
1	002	TEST1	Yes	< 20%	28	-4.6
2	003	TEST1	No	>= 20%	4	-0.2
2	003	TEST1	No	>= 20%	28	-6.2
3	004	TEST1	No	>= 20%	4	-2.8
3	004	TEST1	No	>= 20%	16	-21.2
3	005	TEST1	No	< 20%	4	13.4
3	005	TEST1	No	< 20%	16	35.2
3	005	TEST1	No	< 20%	28	27.4

Stratification Factor 1:
Prior use of therapy [Yes, No]

Stratification Factor 2:
Baseline PARAM1 [≥ 20 , < 20]

A snapshot of ADaM Data Set, ADEFF

Planned Treatment:
1=Drug A,
2=Drug B,
3=Placebo

TRTP	USUBJID	PARAMCD	STRATA1	STRATA2	AVISITN	CHG
1	001	TEST1	No	< 20%	4	-0.8
1	001	TEST1	No	< 20%	16	-3.8
1	001	TEST1	No	< 20%	28	-3
1	002	TEST1	Yes	< 20%	4	-2.2
1	002	TEST1	Yes	< 20%	28	-4.6
2	003	TEST1	No	>= 20%	4	-0.2
2	003	TEST1	No	>= 20%	28	-6.2
3	004	TEST1	No	>= 20%	4	-2.8
3	004	TEST1	No	>= 20%	16	-21.2
3	005	TEST1	No	< 20%	4	13.4
3	005	TEST1	No	< 20%	16	35.2
3	005	TEST1	No	< 20%	28	27.4

Change from Baseline

Week 16 missing

Week 16 missing

Week 28 missing

Stratification Factor 1:
Prior use of therapy [Yes, No]

Stratification Factor 2:
Baseline PARAM1 [≥ 20 , < 20]

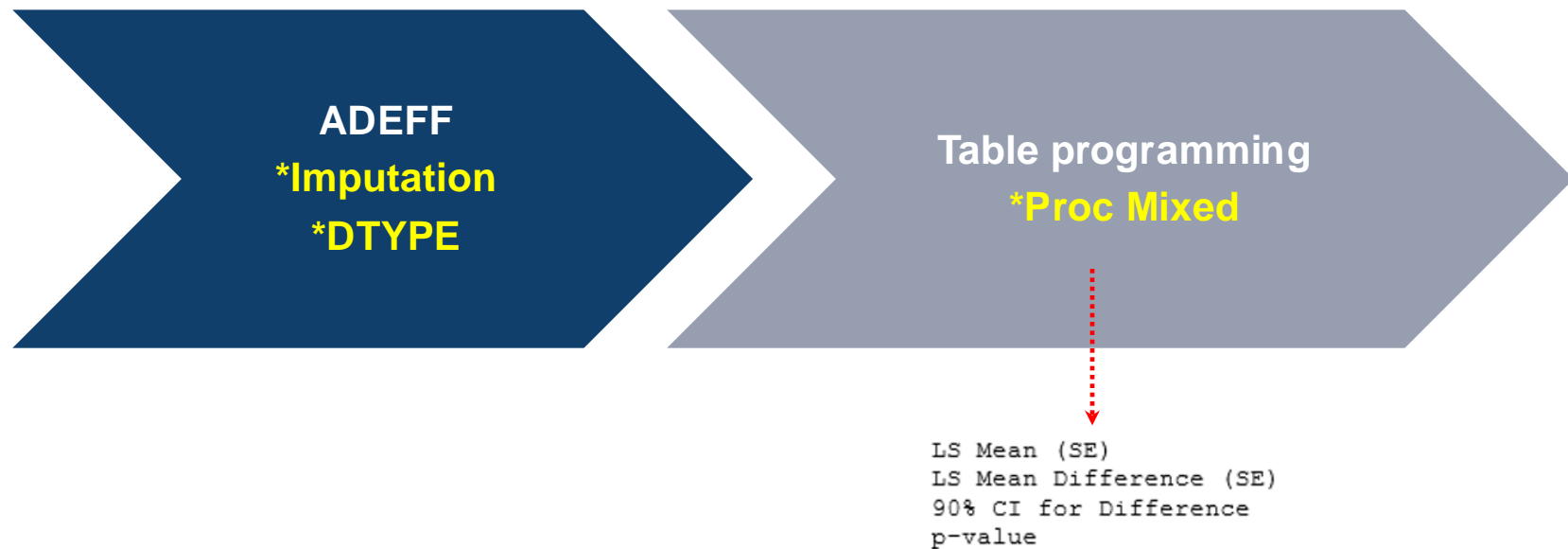
ADEFF - LOCF (Last Observation Carried Forward)

New added variable
- Derivation Type

TRTP	USUBJID	PARAMCD	STRATA1	STRATA2	AVISITN	CHG	DTYPE
1	001	TEST1	No	< 20%	4	-0.8	
1	001	TEST1	No	< 20%	16	-3.8	
1	001	TEST1	No	< 20%	28	-3	
1	002	TEST1	Yes	< 20%	4	-2.2	
1	002	TEST1	Yes	< 20%	16	-2.2	LOCF
1	002	TEST1	Yes	< 20%	28	-4.6	
2	003	TEST1	No	>= 20%	4	-0.2	
2	003	TEST1	No	>= 20%	16	-0.2	LOCF
2	003	TEST1	No	>= 20%	28	-6.2	
3	004	TEST1	No	>= 20%	4	-2.8	
3	004	TEST1	No	>= 20%	16	-21.2	
3	004	TEST1	No	>= 20%	28	-21.2	LOCF
3	005	TEST1	No	< 20%	4	13.4	
3	005	TEST1	No	< 20%	16	35.2	
3	005	TEST1	No	< 20%	28	27.4	

New added Rows!

Flowchart illustrating LOCF analysis require data imputation in ADEFF



LOCF (Last Observation Carried Forward)

- Siddiqui, et al. (2009) shows LOCF analysis can lead to substantial biases in estimators of treatment effects and can greatly **inflate Type I error rates**.
- The National Research Council (NRC) report describes that **LOCF should NOT be used as the primary analysis** unless the assumption that underline this method is scientifically justified, and other approaches can provide less biased outcomes.



MMRM (Mixed Model Repeated Measures)

- Favored for trials with longitudinal continuous outcomes
- Unbiased for MCAR and MAR
- If a participant misses a measurement, MMRM can still use the available data points to conduct analysis. **No imputation needed.**
- Controls Type I error rates better than LOCF

Flowchart illustrating MMRM analysis can use ADEFF directly without imputation.



```
*In Table program, feed ADEFF into PROC MIXED directly ;  
Proc Mixed data= adam.adeff ;  
    by paramcd;  
    where chg^=. and avisitn^=.;  
    class usubjid trtp(ref="3") strata2 strata1 avisitn ;  
    model chg = trtp avisitn trtp*avisitn strata2 strata1/solution ddfm=kr;  
    repeated avisitn / subject=usubjid type=cs ;  
    lsmeans avisitn*trtp / alpha=0.1 cl pdiff;  
    ods output lsmeans= LS Mean  
              diffs= LS Mean Difference;  
run;
```

- A further sensitivity analysis to assess the robustness of the MMRM results under various missing data mechanisms.
- FDA E9(R1): Sensitivity analysis should be planned for the main estimators of all estimands that will be important for regulatory decision making.



MI (Multiple imputation)

- Multiple imputation (MI) is a statistical technique used to handle missing data by creating multiple complete datasets, analyzing each one, and then combining the results.
- Unlike single-value imputation, MI handles missing data by estimating and replacing missing values *many times*.
- **Sensitivity analysis** – Assessing the reliability of the study results under different assumptions, parameters, or models used in the analysis
- MI offers a comprehensive analysis of missing data effects.

MI (Multiple Imputation)

- There are at least three steps for implementing MI:

Step 1. Imputation: Using imputation model to create one big dataset that includes multiple imputed datasets. (**PROC MI**)

Step 2. Analysis: Using analysis model to analyze each imputed data set. (**PROC MIXED**)

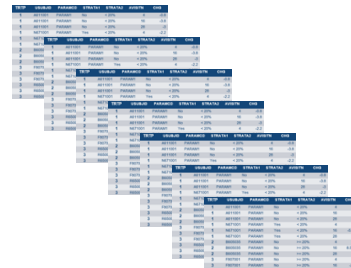
Step 3. Combining Results: Using Rubin's rule to combine results into a single set of estimates. (**PROC MIANALYZE**)

Example:

Original ADEFF:
1000 records

TRTP	USUBJID	PARAMCD	STRATA1	STRATA2	AVISITN	CHG
1	AE11001	PARAM1	No	< 20%	4	-0.8
1	AE11001	PARAM1	No	< 20%	16	-2.8
1	AE11001	PARAM1	No	< 20%	28	-3
1	NE11001	PARAM1	Yes	< 20%	4	-2.2
1	NE11001	PARAM1	Yes	< 20%	16	-0.003
1	NE11001	PARAM1	Yes	< 20%	28	-4.6
2	BS05035	PARAM1	No	≥ 20%	4	-0.2
2	BS05035	PARAM1	No	≥ 20%	16	0.5465
2	BS05035	PARAM1	No	≥ 20%	28	-6.2
3	FR07001	PARAM1	No	≥ 20%	4	-2.8
3	FR07001	PARAM1	No	≥ 20%	16	-21.8
3	FR07001	PARAM1	No	≥ 20%	28	1.7944
3	RS00021	PARAM1	No	< 20%	4	13.4
3	RS00021	PARAM1	No	< 20%	16	38.2
3	RS00021	PARAM1	No	< 20%	28	37.4

100 Multiple imputed
datasets:
 $1000 * 100 = 100000$ records



The results from Step 1 is recommended to be stored in *intermediate ADaM dataset* ADEFFMI.

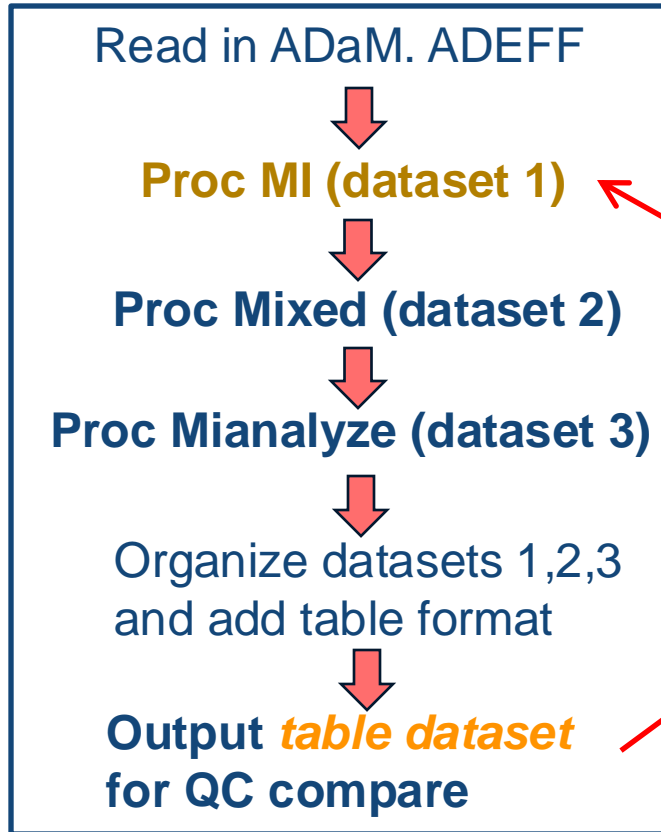
Intermediate ADaM Data Set ADEFFMI – Subject 002 as example

TRTP	USUBJID	PARAMCD	STRATA1	STRATA2	AVISITN	CHG	DTYPE	IMPNUM
1	002	PARAM1	Yes	< 20%	4	-2.2		1
1	002	PARAM1	Yes	< 20%	16	-5.2 MI		1
1	002	PARAM1	Yes	< 20%	28	-4.6		1
1	002	PARAM1	Yes	< 20%	4	-2.2		2
1	002	PARAM1	Yes	< 20%	16	-4.9 MI		2
1	002	PARAM1	Yes	< 20%	28	-4.6		2
1	002	PARAM1	Yes	< 20%	4	-2.2		3
1	002	PARAM1	Yes	< 20%	16	-4.97 MI		3
1	002	PARAM1	Yes	< 20%	28	-4.6		3
⋮								
1	002	PARAM1	Yes	< 20%	4	-2.2		100
1	002	PARAM1	Yes	< 20%	16	-5.15 MI		100
1	002	PARAM1	Yes	< 20%	28	-4.6		100

'IMPNUM' refers to the Imputation Number variable from PROC MI, which, in this scenario, ranges from 1 to 100

What if we don't use an Intermediate dataset?

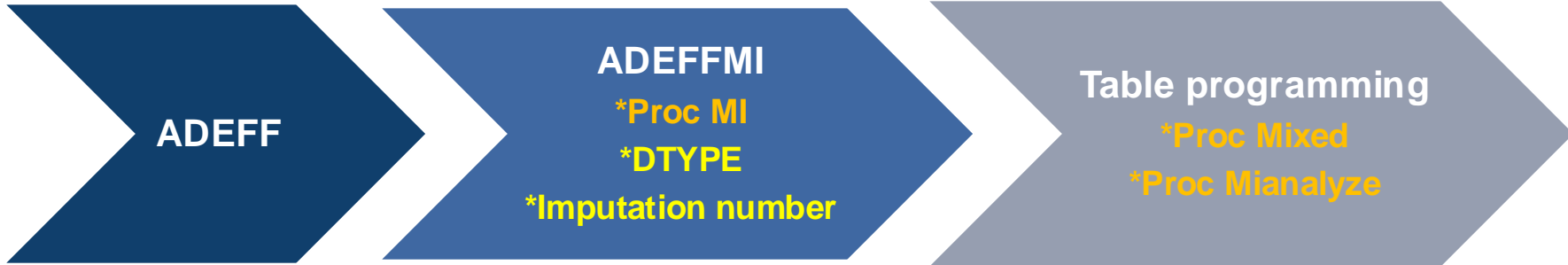
MI table
program



Validator will still need to ask
for dataset 1 to compare

Production and Validation (QC)
Not matching

ADaM Intermediate Dataset



- ★ Accelerate QC process
- ★ Better traceability



Summary

Summary

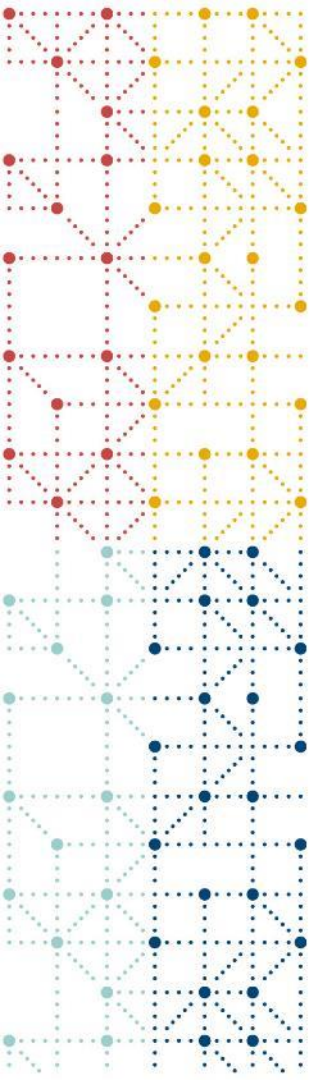
Missing data in longitudinal rare disease studies

Three missing data mechanisms: MCAR, MAR, MNAR

Common methodologies for analyzing missing data: LOCF, MMRM, MI

Sensitivity analysis

Intermediate ADaM dataset when implementing MI



Thank You!

cdisc