

Efficacy Analyses with Imputations: Methods, Limitations, and Lessons Learned

Amine Khemiri

AS08 – PHUSE EU Connect 2023



AGENDA

- Introduction
- The Missing Data Conundrum
- Imputation Methods
- Case Study Example
- Summary & Conclusion



#Data #DataDriven
#DataStrategy #DataIntegration
#DataViz #DataStandards
#DataScience #Datagovernance
#HealthCare #DataAnalysis
#ClinicalTrials
#ArtificialIntelligence
#AdvancedAnalytics
#MachineLearning #Digital
#Chatbot #Biometrics #CRO
#ClinicalData
#LifeSciencesExperts
#Whodrug #MedDRA
#Pharmacovigilance
#ClinicalResearchLifeCycle
#MedicalWriting
#MedicalInformation
#MedicalReview #SDTM
#ADaM #CDISC #Innovation
#RegulatorySubmissions
#Protocol #Coding

INTRODUCTION

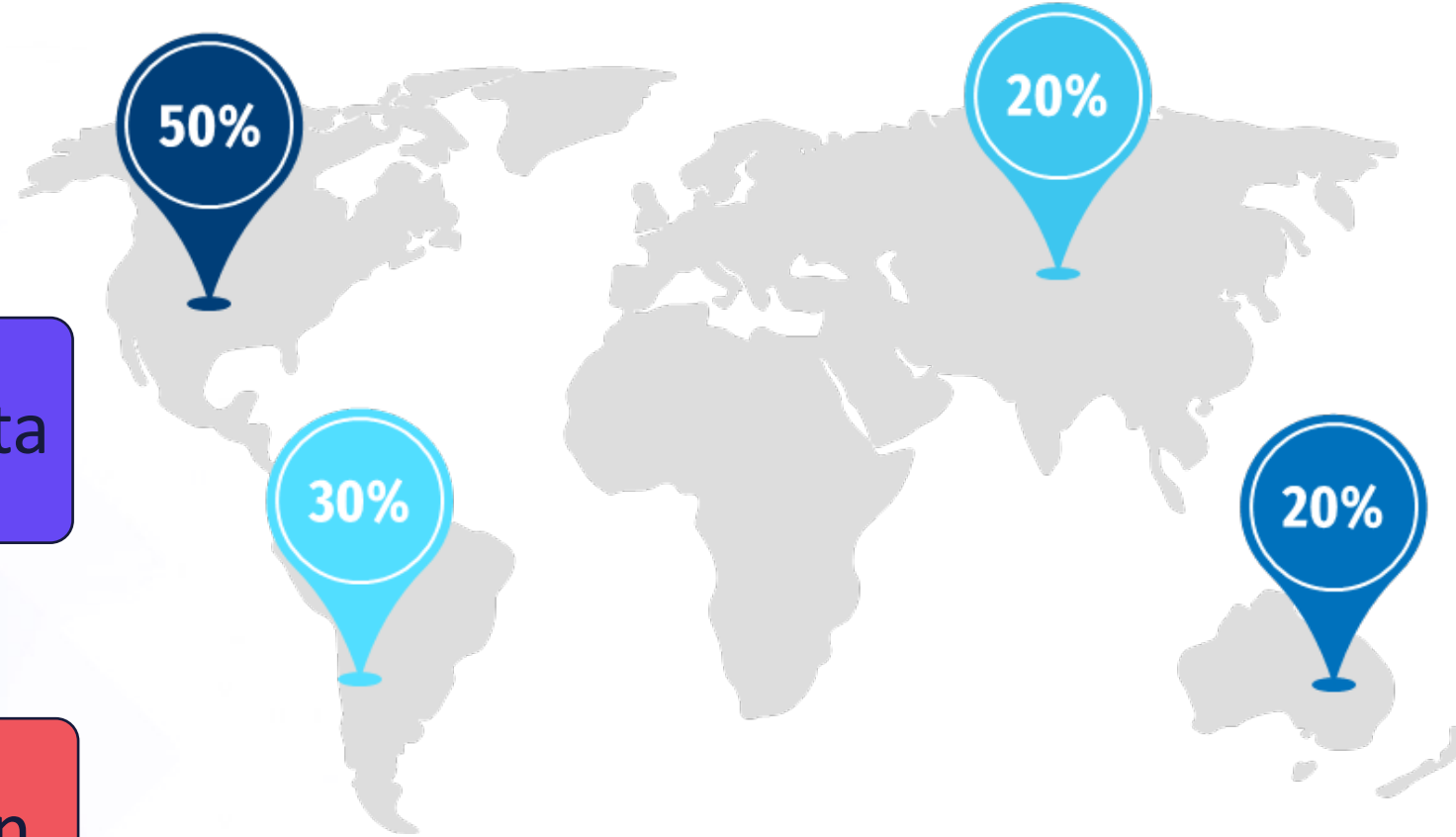
Missing (%) : Sample Example

Efficacy
analysis

Missing data

Causes of
missing data

Imputation



Imputation is like a smart helper that fills in missing information, saving us from having to gather more data. It makes our analysis stronger without extra data collection efforts.

THE MISSING DATA CONUNDRUM

- There are 3 major types of missingness to be concerned about:



- **Missing Completely at Random (MCAR):**
 - The probability of an observation being missing is unrelated to both the observed and unobserved values.
- **Missing at Random (MAR):**
 - The probability of an observation being missing is not related to the unobserved value itself, but it may be related to other observed variables.
- **Missing Not at Random (MNAR):**
 - The probability of an observation being missing is related to the unobserved value itself

IMPUTATION

- **What is Imputation?**
 - It is a statistical technique used to replace missing data with estimated values
- **Why use Imputation?:**
 - Ensures a more complete dataset for analysis
 - Preserves sample size and statistical power
 - Reduces potential bias in parameter estimates
- **Considerations for imputation:**
 - Nature of missing data (MCAR, MAR, MNAR)
 - Study design and objectives
 - Statistical assumptions
 - Available resources and data characteristics
- **Types of Imputation:**
 - Single Imputation
 - Multiple Imputation (MI)

IMPUTATION METHODS

- **Single imputation**

- It is a method used to estimate missing values by replacing them with a single value.
- A simple and straightforward approach

ID	Var 1	Var 2
1	12	27
2	11	x
3	x	56
4	15	37
5	10	31
6	18	40
7	15	x
...	...	...



ID	Var 1	Var 2
1	12	27
2	11	56.05511
3	15.20032	56
4	15	37
5	10	31
6	18	40
7	15	40.45120
...	...	...

IMPUTATION METHODS

- **Methods of single imputation**

Mean, Median, or Mode Imputation

Replacing missing values with the mean or the median or the mode of the observed values for continuous variables.

Last Observation Carried Forward (LOCF)

Assumes that the last observed value is carried forward for subsequent missing time points.

Baseline Observation Carried Forward (BLOCF)

This method involves imputing missing values with the baseline (initial) observation.

Worst Observed Case (WOC)

It involves imputing missing values with the worst (most severe) observed value.

Hot Deck Imputation

Replaces missing values with similar values from other respondents (donors), based on characteristics that are like the recipient.

Regression Imputation

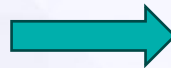
Predicts missing values using regression models based on observed variables.

IMPUTATION METHODS

- **Multiple imputation**

- It involves creating several plausible imputed datasets, each of which replaces missing values with a set of estimated values.
- These imputed datasets are then analyzed separately, and the results are combined to provide a more robust and accurate estimate of the parameters of interest.

ID	Var 1	Var 2
1	12	27
2	11	x
3	x	56
...	...	...



Dataset 1

ID	Var 1	Var 2
1	12	27
2	11	56.05511
3	15.20032	56
...	...	...

Dataset 2

ID	Var 1	Var 2
1	12	27
2	11	49.02245
3	17.35246	56
...	...	...

Dataset 3

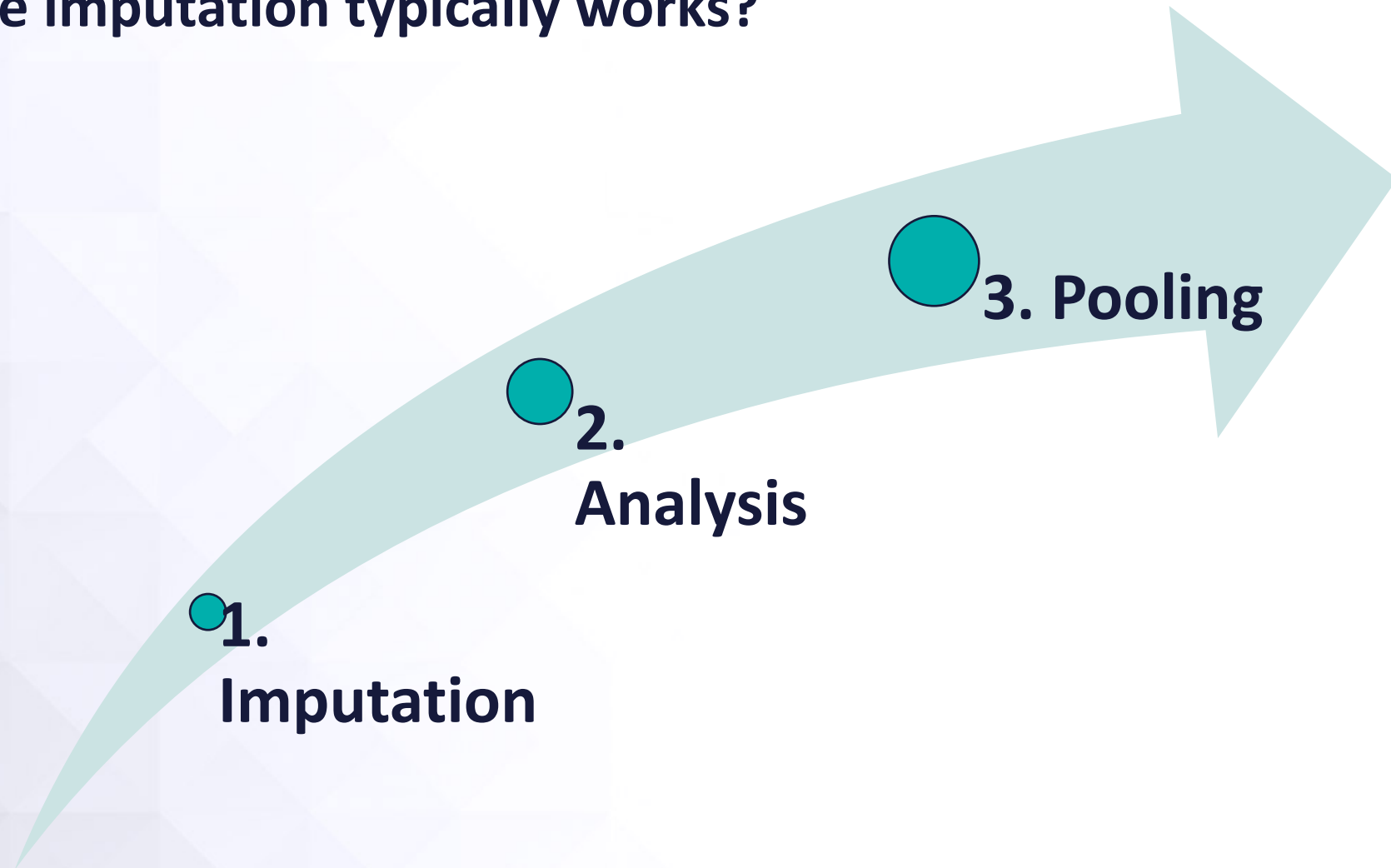
ID	Var 1	Var 2
1	12	27
2	11	32.18994
3	21.45879	56
...	...	...

Dataset 4

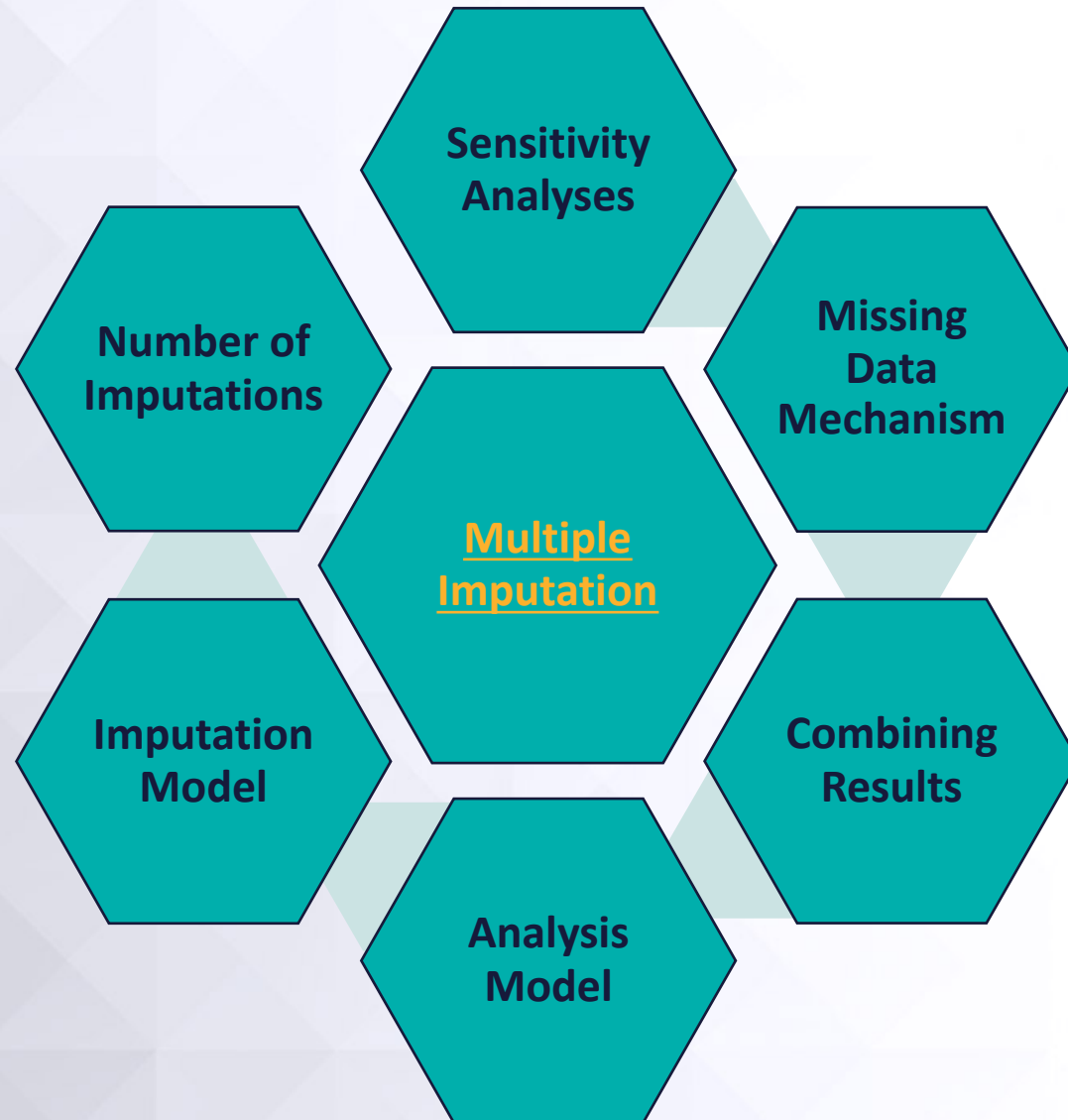
ID	Var 1	Var 2
1	12	27
2	11	46.456879
3	19.524565	56
...	...	...

IMPUTATION METHODS

How multiple imputation typically works?



CONSIDERATIONS FOR MULTIPLE IMPUTATION



- Multiple imputation is considered a powerful and flexible technique for handling missing data
- The key advantage of multiple imputation is that it properly accounts for the uncertainty associated with imputed values
- Leads to more accurate estimates and valid inferences
- Selecting the most appropriate imputation method is crucial as it can have a significant impact on the results of efficacy analysis
- The choice should be based on the data characteristics, assumptions, and limitations of the imputation method

SINGLE IMPUTATION VS. MULTIPLE IMPUTATION

	Advantages	Limitations
Single Imputation: • A common method used to handle missing data • It replaces missing value with single value from the observed data for the variable	• Preserves sample size • Does not require complex algorithm • Quick results	• May not capture the true variability of the missing values • Lead to biased results
Multiple Imputation: • Advanced method that creates multiple plausible values for missing data based on the observed data and their relationships • It considers the uncertainty associated with the missing values, providing a more accurate estimation	• Capture the uncertainty • Produces more reliable results • Facilitates Sensitivity Analyses	• Computationally demanding • Requires assumptions about data distribution • Greater Analytical Demands

CASE STUDY EXAMPLE

	Drug A (N=100)	Placebo (N=100)
Body Weight - Mean change from baseline over 6 months - Observed Values (non-missing data)		
n	70	66
Mean $\pm$ SD	3.136 $\pm$ 1.893	0.203 $\pm$ 2.563
Median	2.60	0.65
Q1; Q3	0.250; 4.500	-0.750; 1.400
Min; Max	-5.65; 10.20	-7.20; 6.62

	Drug A (N=100)	Placebo (N=100)
Body Weight - Mean change from baseline over 6 months - Imputed Values (All data)		
n	100	100
Mean $\pm$ SD	2.444 $\pm$ 0.893	1.203 $\pm$ 1.563
Median	1.90	1.15
Q1; Q3	0.250; 4.500	-0.750; 1.400
Min; Max	-6.32; 10.20	-7.20; 7.44

CASE STUDY EXAMPLE

	Drug A (N=100)	Placebo (N=100)	Drug A vs. Placebo
Body Weight - Mean change from baseline over 6 months - LSM model estimates after imputation			
Mean ± SE	2.456 ± 0.285	0.897 ± 0.456	1.489 ± 0.320
95% CI	[1.379 ; 2.999]	[0.036 ; 1.152]	[0.718 ; 1.971]
p-value			<.0001

```
ODS exclude all ;
proc mixed data=model_ADEFF ;
  by _imputation_ ;
  class &dis_var.(ref="2") acountry sex race ;
  model IMCHG = &dis_var. acountry sex race ;
  lsmeans &dis_var. / cl diffs ;
  estimate 'Drug A - Placebo' &dis_var. 1-1 /cl;
  ods output LSMEANS=MIlsmeans;
  ods output estimates=MIancova;
run ;
```

- Think about preparing input dataset containing imputed values and imputation numbers



- CDISC Format
- Results can be reproducible
- Computational algorithms
- Easy to review
- Keep traceability

CASE STUDY EXAMPLE

ADVS

	AVAL	AVISITN ▲	IMPUTN	DTYPE	ANL01FL	ANL02FL	ANL03FL
2	37.838285323	8.0	.	AVERAGE2		Y	
3	39.10639518	8.0	1	MNAR2		Y	
4	36.897612038	8.0	2	MNAR2		Y	
5	37.989073934	8.0	3	MNAR2		Y	
6	38.2	8.0	.	WOV4			
7	38.2	8.0	.	WOV3			Y
8	38.2	8.0	.	WOV	Y		
9	38.2	8.0	.	WOV5			
10	39.790948658	9.0	4	MNAR2		Y	

CASE STUDY EXAMPLE

ADEFF

	AVISITN	PARAM	IMPUTN	ATYPE	OMCHG ▲	IMCHG
	8	Weight, kg	65	Main Analysis	.	-2.021703753
	9	Weight, kg	77	Sensitivity Analysis - C	.	-1.138821307
	8	Weight, kg	81	Sensitivity Analysis - C	.	-1.528462294
	9	Weight, kg	93	Main Analysis	.	-2.186049153
	8	Weight, kg	49	Sensitivity Analysis - C	.	-2.191087972
	7	Weight, kg	37	Main Analysis	.	-1.037454934
	8	Weight, kg	1	Main Analysis	.	-2.23721932
	6	Weight, kg	89	Sensitivity Analysis - C	.	-1.110154695
	9	Weight, kg	29	Main Analysis	.	-2.817717694

CASE STUDY EXAMPLE

ADEFF

	imputation	Estimate	StdErr	Probt
1	1	10.0063	0.4766	<.0001
2	2	9.8508	0.4610	<.0001
3	3	9.9164	0.4592	<.0001
4	4	10.0402	0.4581	<.0001
5	5	10.0641	0.4547	<.0001
6	6	9.6029	0.4651	<.0001
7	7	9.9211	0.4572	<.0001
8	100	9.7954	0.4615	<.0001

Pool estimates, standard errors, and p-values across imputations

CASE STUDY EXAMPLE

ADEFF

AVISITN	IMPUTN	IMCHG - Weight	IMCHG - QoL	CCR	SRCDOM
3	1	-2.12	-1.32	Responder	ADVS AND ADQOL
3	2	-2.32	0.46	Non-Responder	ADVS AND ADQOL
4	1	-2.12	-1.32	Responder	ADVS AND ADQOL
4	2	-2.32	-1.32	Non-Responder	ADVS AND ADQOL
5	1	-2.12	-1.32	Responder	ADVS AND ADQOL
5	2	-2.12	-0.55	Non-Responder	ADVS AND ADQOL

MINIMIZING MISSING DATA

- **Prevention is Key:**

- Rigorous data collection protocols
- Minimize dropout through participant engagement
- Utilize follow-up strategies

- **Training and Standardization:**

- **Investigator Training:** Provide thorough training to investigators and site staff on data collection protocols and procedures.
- **Regular Site Monitoring:** Conduct regular site visits to ensure compliance with data collection protocols.

- **Redundancy in Data Collection:**

- **Multiple Data Sources:** Collect data from multiple sources whenever possible to cross-verify and minimize missing information.
- **Use of Proxy Measures:** In cases where direct data collection is challenging, consider using validated proxy measures.

MINIMIZING MISSING DATA

- **Use of Electronic Data Capture (EDC) Systems**
 - **Real-Time Data Entry:** Implement EDC systems for immediate data entry, reducing the likelihood of missing data due to delayed or manual entry.
 - **Built-In Data Validation Checks:** EDC systems can incorporate checks to flag potential data entry errors, reducing the chance of missing data.
- **Patient Engagement and Compliance**
 - **Clear Patient Instructions:** Provide clear and concise instructions to participants on data collection procedures.
 - **Regular Follow-Up and Support:** Maintain open communication with participants, addressing any concerns or issues promptly.

CONCLUSION

KEY TAKEAWAYS

1. Imputations Enhance Analytical Rigor:

- Imputations are valuable tools for handling missing data, preserving sample size, and reducing bias in efficacy analyses.

2. Diverse Imputation Methods

- Various imputation methods, from simple mean imputation to sophisticated regression techniques, offer flexibility in addressing different types of missing data.

3. Considerations for Imputation

- Understand the missing data mechanism, choose appropriate imputation models, and validate assumptions for robust results.

CONCLUSION

LIMITATIONS AND CAVEATS

1. Assumptions Matter

- Imputations rely on assumptions about the missing data mechanism. Sensitivity analyses are crucial for assessing the impact of these assumptions.

2. Imputation Uncertainty

- While imputation improves estimates, it introduces variability. This must be considered in result interpretation and reporting.

3. Not a One-Size-Fits-All Approach

- The choice of imputation method should be tailored to the study context, and researchers should be aware of the strengths and limitations of each approach.

CONCLUSION

CHALLENGES

1. Impact on Timelines

- **Increased Processing Time:** Imputation, especially for large datasets, can be computationally intensive and time-consuming. This may lead to extended timelines for data analysis and result reporting.
- **Iterative Process:** Fine-tuning imputation models and validating results can require multiple iterations, potentially causing delays in trial completion and subsequent phases.

2. Resource Allocation

- **Computational Resources:** Imputation, particularly with complex models or large datasets, may demand substantial computing power. This can strain available resources and necessitate investment in additional hardware or cloud services.
- **Skilled Personnel:** Implementing advanced imputation methods may require specialized statistical expertise, leading to increased costs for hiring or training personnel proficient in these techniques.

CONCLUSION

LESSONS LEARNED

1. Validation is Key

- Rigorously validate imputation models to ensure they accurately represent the missing data.

2. Transparency and Reporting

- Clearly document the imputation process, including methods used and any auxiliary information incorporated.

3. Sensitivity Analyses are Essential

- Explore the robustness of results to different imputation strategies to provide a more complete picture of the findings.



THANK YOU FOR YOUR ATTENTION
Q&A

CONTACT



Caroline Pelckmans

Statistical Programmer/ Biostatistician

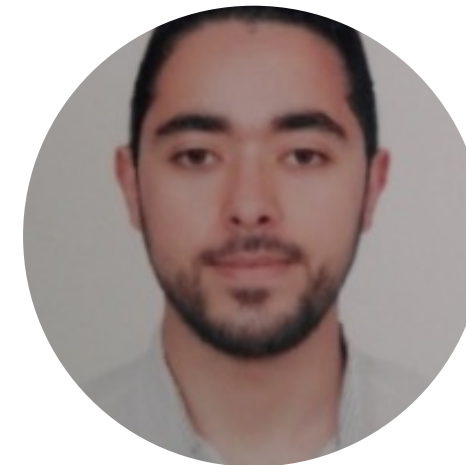
Caroline.pelckmans@businessdecision.be



Amine Khemiri

**Project Manager / Statistical
Programmer**

amine.khemiri@businessdecision.be



Oussama Zineddine

Statistical Programmer

Oussama.Zineddine@businessdecision.be