Handling missing data in practice:

Complete Case Analysis vs Multiple imputation

PV Ebasone, MD, PhD

CRENC Clinical Research Education Networking and Consultancy

IeDEA International Epidemiology Databases to Evaluate AIDS CENTRAL AFRICA

Missing Data: Causes and Why It Matters

Causes

1. Entry mistakes or skipped fields
2. Refusal to answer sensitive items
3. Missed visits or dropouts
4. Device/software failure, poor connectivity
5. Protocols too long or complex
6. Data transfer or type conversion errors
7. Merges with mismatched keys or structures

Why it matters

1. **Bias and accuracy:**

- Biase estimates and distort prevalence / effects

2. **Precision and stability:**

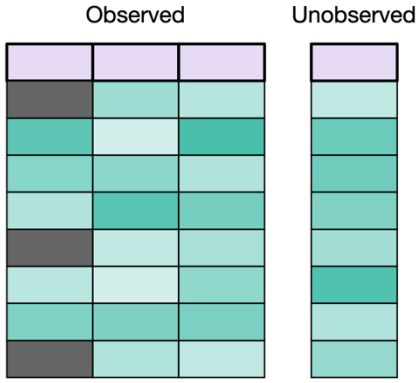
- Loss of power; wider CIs; unstable models

3. **Operational and strategic cost:**

- Slower analysis; higher cost; wasted data; poor decisions

Types of Missing Data

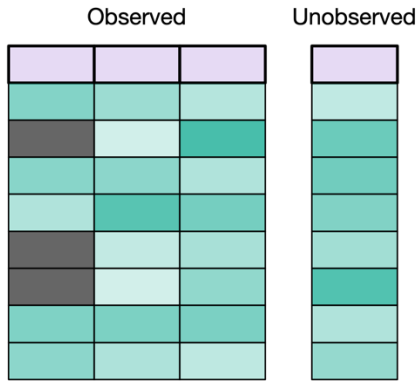
Missing Completely at Random (MCAR)



ID	Sex	Age	SBP	SBP
1	Female	27	105	105
2	Female	26	104	104
3	Female	33	NA	120
4	Male	61	127	127
5	Female	43	109	109
6	Female	34	119	119
7	Female	18	110	110
8	Male	44	NA	110
9	Female	51	119	119
10	Female	78	135	135

Missingness unrelated to Age, Sex, or SBP.

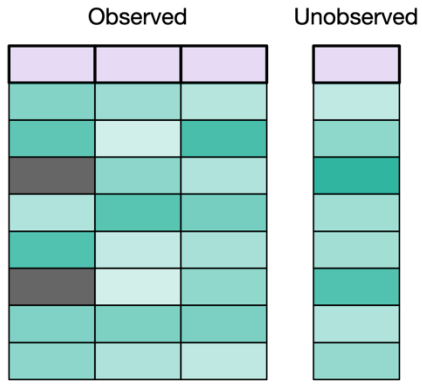
Missing at Random (MAR)



ID	Sex	Age	SBP	SBP
1	Female	27	105	105
2	Female	26	NA	104
3	Male	23	NA	101
4	Male	61	127	127
5	Female	43	109	109
6	Female	71	119	119
7	Female	18	NA	102
8	Male	44	106	106
9	Female	51	119	119
10	Female	78	135	135

SBP missing more often for Age < 35 (low-risk triage).

Missing Not at Random (MNAR)



ID	Sex	Age	SBP	SBP
1	Female	27	105	105
2	Female	26	104	104
3	Male	23	101	101
4	Male	64	NA	192
5	Female	47	NA	202
6	Female	71	119	119
7	Female	18	102	102
8	Male	44	106	106
9	Female	51	119	119
10	Female	78	135	135

Very high SBP triggers not recorded due to emergency

1. Handling Missing Data: Ignore/Delete Approaches

i. Listwise deletion
(aka **complete case analysis**): remove entire rows with any missing value.

ID	Sex	Age	SBP
1	Female	27	105
2	Female	26	104
3	Male	43	NA
4	Male	61	127
5	NA	NA	109
6	Female	34	119

ID	Sex	Age	SBP
1	Female	27	105
2	Female	26	104
4	Male	61	127
6	Female	34	119

ii. Pairwise deletion: use available data for each variable pair.

ID	Sex	Age	SBP
1	Female	27	105
2	NA	NA	127
3	Male	23	103
4	Male	61	127
5	Female	43	109
6	Female	71	119

ID	Sex	Age
1	Female	27
3	Male	23
4	Male	61
6	Female	71

iii. Column deletion: remove a variable with too much missing data (e.g., >40%)

N/B: All three methods are valid only if MCAR and missingness is small. Otherwise, they reduce power or bias results.

ID	Sex	Age	SBP
1	Female	27	105
2	Female	NA	104
3	Male	NA	110
4	Male	61	127
5	Male	NA	109
6	Female	34	119

ID	Sex	SBP
1	Female	105
2	Female	104
3	Male	110
4	Male	127
5	Male	109
6	Female	119

2. Handling Missing Data: Single Imputation Approaches

i. **Mean/median/mode** Imputation

ID	Sex	Age	SBP
1	Female	27	105
2	Female	26	104
3	Male	43	NA
4	Male	61	127
5	NA	NA	109
6	Female	34	119

ID	Sex	Age	SBP
1	Female	27	105
2	Female	26	104
3	Male	43	113
4	Male	61	127
5	Female	38	109
6	Female	34	119

ii. **Last/Next Observation Carried Forward/Backward** (LOCF/NOCB)

ID	Sex	Age	SBP
1	Female	27	105
2	Female	NA	104
3	Male	NA	110
4	Male	61	127
5	Male	NA	109
6	Female	34	119

ID	Sex	Age	SBP
1	Female	27	105
2	Female	27	104
3	Male	27	110
4	Male	61	127
5	Male	34	109
6	Female	34	119

NOCB

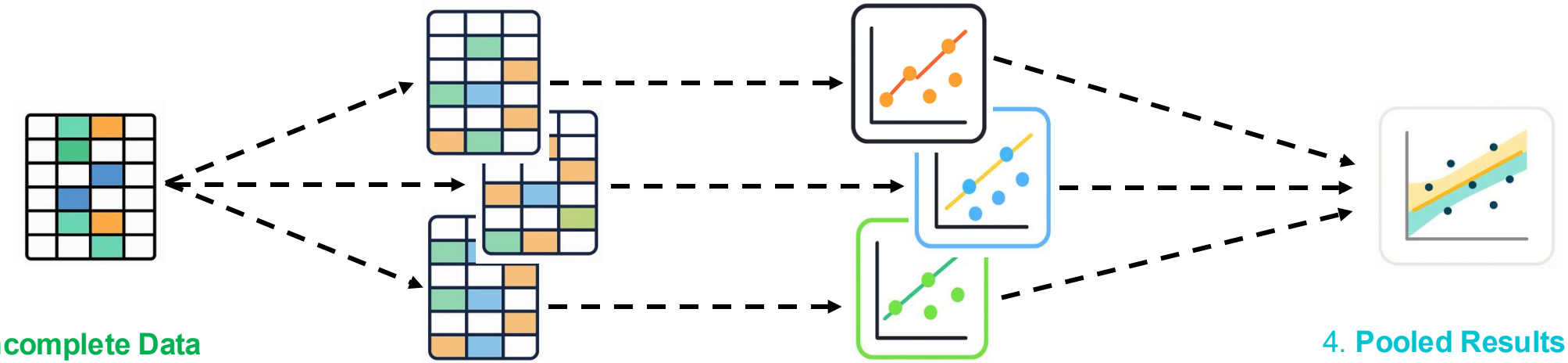
iii. **Missing Indicator Method**

ID	Sex	Age	SBP
1	Female	27	105
2	NA	37	127
3	Male	23	103
4	Male	61	127
5	NA	50	109
6	Female	71	119

ID	Sex	Age	SBP
1	Female	27	105
2	Missing	37	127
3	Male	23	103
4	Male	61	127
5	Missing	50	109
6	Female	71	119

N/B: All single imputation methods underestimate variability and hence risk biasing results.

3. Handling Missing Data: Multiple Imputation



1. Incomplete Data

- Dataset with missing values. Goal: avoid bias and data loss.

2. Imputed Data

- Generate multiple plausible datasets.
- Each missing value filled in using MICE (regression models conditional on other variables).

3. Analysis Results

- Analyze each imputed dataset separately
- Get estimates & standard errors

4. Pooled Results

- Combine results across datasets using Rubin's Rules
- Accounts for both within- and between-imputation variance
- Produces valid estimates & confidence intervals

Missingness Type to Handling Method

Missingness Type	Handling Methods
MCAR (Missing Completely at Random)	• **Listwise (CCA) / Pairwise deletion** • Simple imputation (mean/median)
MAR (Missing at Random)	• **Multiple Imputation (MICE)** • Maximum Likelihood (e.g., mixed models, EM algorithm)
MNAR (Missing Not at Random)	• **Sensitivity analyses**. • Pattern-mixture models • Selection models / **Delta adjustment**

N/B: The best method for handling missing data is preventing it at study design and data collection stage.

Assessing Missing Data in Practice

1. Background review: Check study design, protocol, and collection process

2. Explore patterns: Visualize missingness (heatmaps, missing data matrix)

3. Formal checks:

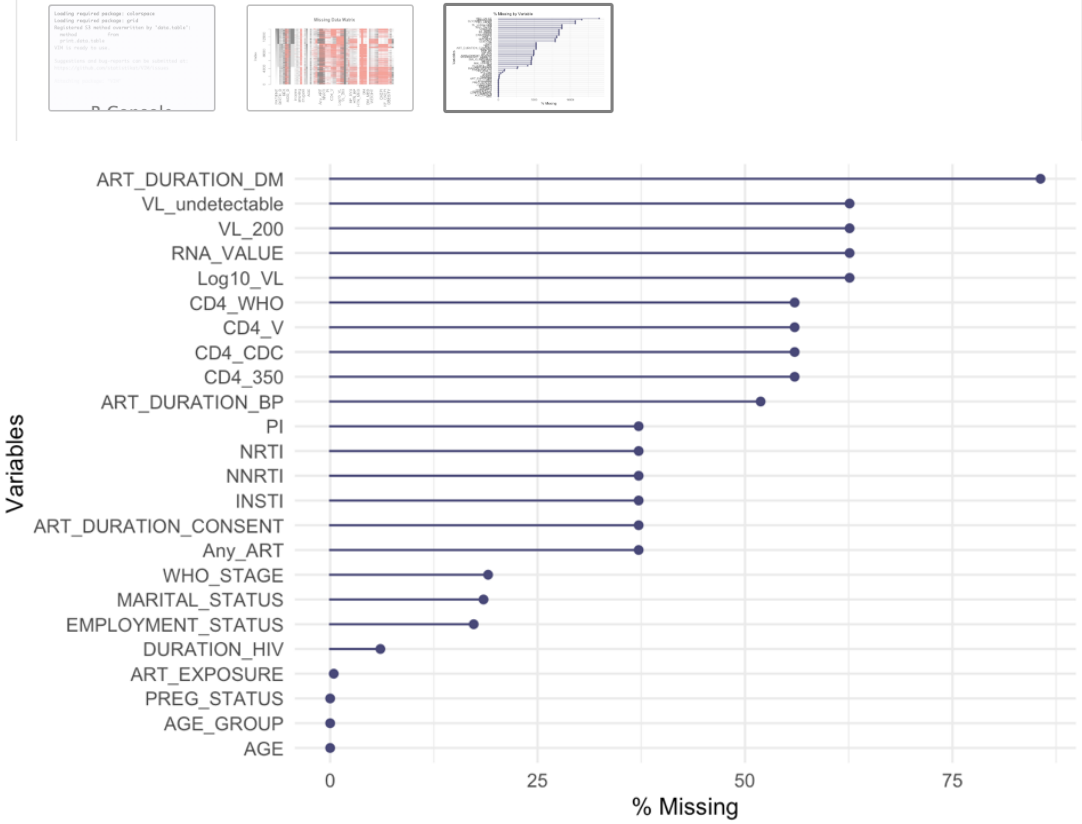
Little's MCAR test (tests MCAR assumption)	Compare observed vs missing groups (Chi-square, t-tests, SMD)

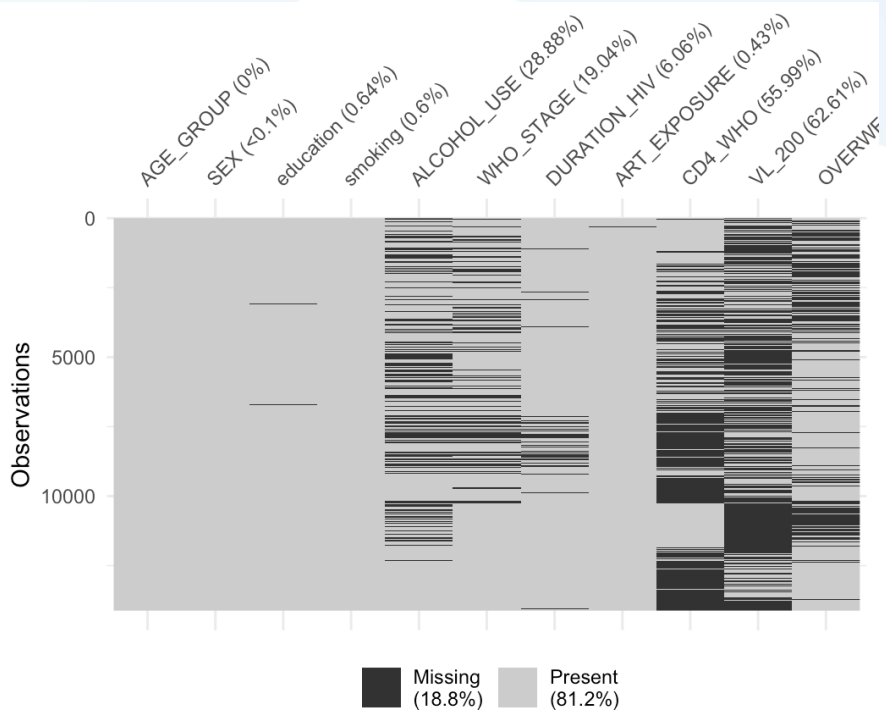
4. Modeling approach

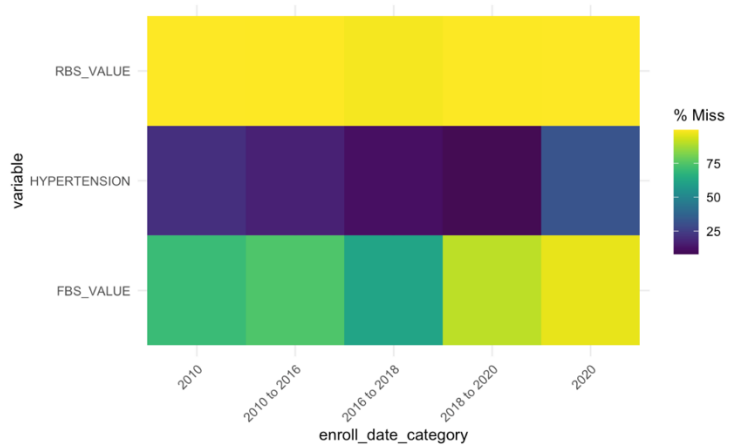
Logistic regression: outcome = "is missing" indicator, predictors = other variables	Helps identify MAR predictors

Missing data visualisations: example

```r
{r}
library(VIM)
library(naniar)
library(ggplot2)

# --- 1. Missingness matrix ---
matrixplot(BASELINE, main = "Missing Data Matrix")

# --- 2. % missing per variable ---
gg_miss_var(BASELINE) +
  labs(title = "% Missing by Variable",
       x = "Variables", y = "% Missing") +
  theme_minimal()
```







Baseline imbalance by BMI missingness

1. **Create missingness indicator**: Observed vs Missing.

2. **Compare groups:** Use chi-square/Wilcoxon tests.

3. **Quantify imbalance**: Use SMD, preferred over p-values because it reflects magnitude, not sample size.

```r
{r}

BASELINE2 <- BASELINE2 |>
  dplyr::mutate(BMI_missing = factor(BMI_missing, levels = c("Observed","Missing")))

vars <- c("AGE_GROUP","SEX","ALCOHOL_USE","CD4_V","VL_200")
tab1 <- tableone::CreateTableOne(vars = vars, strata = "BMI_missing",
                                 data = BASELINE2, test = FALSE)

plot_smd_tableone <- function(tab1, threshold = 0.10, title = "Baseline imbalance by BMI
missingness") {
  smd <- tableone::ExtractSmd(tab1)
  smd_df <- if (is.null(dim(smd))) data.frame(variable = names(smd), SMD = as.numeric(smd))
            else data.frame(variable = rownames(smd), SMD = as.numeric(smd[, 1]))
  ggplot2::ggplot(smd_df, ggplot2::aes(x = SMD, y = reorder(variable, abs(SMD)))) +
    ggplot2::geom_vline(xintercept = 0, linewidth = 0.5) +
    ggplot2::geom_vline(xintercept = c(-threshold, threshold), linetype = "dashed", linewidth =
0.5) +
    ggplot2::geom_point(size = 2.6) +
    ggplot2::labs(x = "Standardized mean difference", y = "Variable",
                  title = title, subtitle = paste0("Dashed lines at |SMD| = ", threshold)) +
    ggplot2::theme_minimal()
}

plot_smd_tableone(tab1, threshold = 0.10)
```

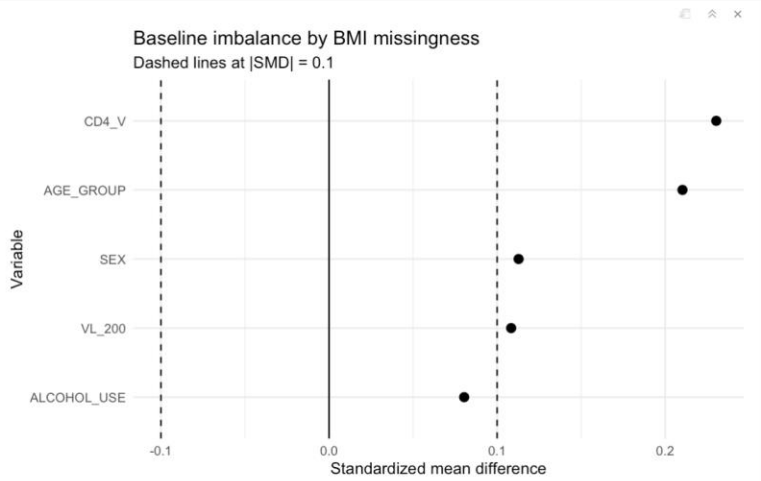


Fig 1: Compute standardized mean differences (**SMDs**) and plot imbalance between “BMI observed” and “BMI missing”. Vertical lines mark |SMD| = 0.10.

```r
{r}

library(dplyr); library(gtsummary); library(tidyselect)

BASELINE2 <- BASELINE %>%
  mutate(BMI_missing = factor(if_else(is.na(BMI), "Missing", "Observed"),
                              levels = c("Observed","Missing")))

tbl_baseline <- BASELINE2 %>%
  dplyr::select(all_of(c("AGE_GROUP","SEX","ALCOHOL_USE","CD4_V","VL_200")), BMI_missing) %>%
  gtsummary::tbl_summary(by = BMI_missing,
    type = gtsummary::all_continuous() ~ "continuous2",
    statistic = gtsummary::all_continuous() ~ "{median} ({p25}, {p75})",
    missing = "ifany") %>%
  gtsummary::add_p() %>%
  gtsummary::add_difference(test = gtsummary::everything() ~ "smd") %>%
  gtsummary::bold_labels() %>%
  gtsummary::modify_spanning_header(gtsummary::all_stat_cols() ~ "**Baseline by BMI missingness**")

tbl_baseline
```

Characteristic	Baseline by BMI missingness Observed N = 9,557[1]	 Missing N = 4,562[1]	p-value[2]
AGE_GROUP			<0.001
19-29	1,247 (13%)	365 (8.0%)	
30-39	2,816 (29%)	1,218 (27%)	
40-49	3,072 (32%)	1,508 (33%)	
≥ 50	2,422 (25%)	1,471 (32%)	
SEX			<0.001
Male	3,517 (37%)	1,435 (31%)	
Female	6,038 (63%)	3,125 (69%)	
Unknown	2	2	
ALCOHOL_USE			0.006
Never	2,262 (31%)	892 (32%)	
Monthly or less	2,375 (33%)	950 (34%)	
2-3 times per month	1,048 (14%)	417 (15%)	
Once a week or more	1,583 (22%)	515 (19%)	
Unknown	2,289	1,788	
CD4_V			<0.001
Median (Q1, Q3)	336 (164, 530)	414 (247, 594)	
Unknown	6,110	1,795	
VL_200			<0.001
≥ 200	790 (23%)	324 (18%)	
< 200	2,710 (77%)	1,455 (82%)	
Unknown	6,057	2,783	

[1] n (%)
[2] Pearson’s Chi-squared test; Wilcoxon rank sum test

Fig 2: Baseline characteristics stratified by BMI missingness.

Model comparisons: R analysis output

Complete case analysis model

vs

Multiple imputed data model

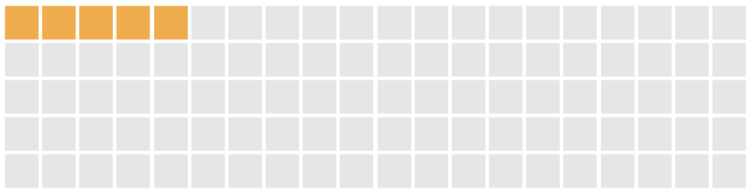
```
Call:  glm(formula = OVERWEIGHT_OBESITY ~ AGE_GROUP + SEX + education +
    ALCOHOL_USE + smoking + WHO_STAGE + DURATION_HIV + ART_EXPOSURE +
    CD4_WHO + VL_200, family = "binomial", data = BASELINE)
```

```
Coefficients:
                  (Intercept)                AGE_GROUP30-39                AGE_GROUP40-49                AGE_GROUP≥ 50
                     -1.97766                       0.19611                       0.80593                       0.51840
                    SEXFemale                  education.L                   education.Q                   education.C
                      0.34585                       0.41795                      -0.15974                       0.17662
                  education^4   ALCOHOL_USEMonthly or less ALCOHOL_USE2-3 times per month  ALCOHOL_USEOnce a week or more
                     -0.07011                       0.10948                       0.32448                       0.27023
  smokingCurrent smoker                smokingFormer smoker                   WHO_STAGE.L                   WHO_STAGE.Q
                     -0.07384                      -0.20378                      -0.46530                      -0.02976
                  WHO_STAGE.C                  DURATION_HIV                ART_EXPOSUREyes                CD4_WHO200 - 349
                      0.02037                       0.06334                      -0.18254                       0.97778
          CD4_WHO350 - 500          CD4_WHOMore than 500                      VL_200< 200
                      0.74708                       1.10592                       0.03990
```

```
Degrees of Freedom: 606 Total (i.e. Null);  584 Residual
  (8950 observations deleted due to missingness)
Null Deviance:      840.9
Residual Deviance: 778.8        AIC: 824.8
```

In this CCA model, only **606** obeservations are used

CCA: 606 / **9556** (6.3%) retained



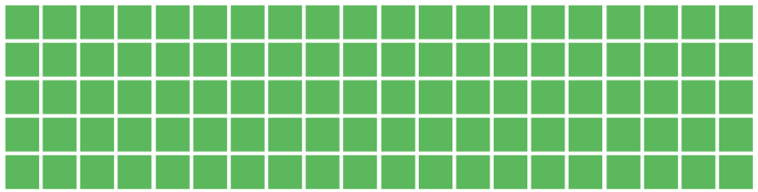
```
> mva_ovo
call :
with.mids(data = miImp, expr = glm(OVERWEIGHT_OBESITY ~ AGE_GROUP +
    SEX + education + ALCOHOL_USE + smoking + WHO_STAGE + DURATION_HIV +
    ART_EXPOSURE + CD4_WHO + VL_200, family = "binomial"))
```

```
call1 :
mice(data = imp_data, m = 10, printFlag = F)

nmis :
       AGE_GROUP              SEX        education          smoking       ALCOHOL_USE          WHO_STAGE
               0                2               67               64             2289               1354
    DURATION_HIV      ART_EXPOSURE           CD4_WHO           VL_200 OVERWEIGHT_OBESITY
             734               46             6110             6057               72

analyses :
[[1]]

Call:  glm(formula = OVERWEIGHT_OBESITY ~ AGE_GROUP + SEX + education +
    ALCOHOL_USE + smoking + WHO_STAGE + DURATION_HIV + ART_EXPOSURE +
    CD4_WHO + VL_200, family = "binomial")

Coefficients:
                  (Intercept)                AGE_GROUP30-39                AGE_GROUP40-49                AGE_GROUP≥ 50
                     -2.50307                       0.67609                       0.97757                       0.81473
                    SEXFemale                  education.L                   education.Q                   education.C
                      0.53606                       0.17740                       0.05581                      -0.13230
                  education^4   ALCOHOL_USEMonthly or less ALCOHOL_USE2-3 times per month  ALCOHOL_USEOnce a week or more
                     -0.06315                       0.22755                       0.63531                       0.58483
  smokingCurrent smoker                smokingFormer smoker                   WHO_STAGE.L                   WHO_STAGE.Q
                     -0.46871                      -0.19523                      -0.42838                       0.14934
                  WHO_STAGE.C                  DURATION_HIV                ART_EXPOSUREyes                CD4_WHO200 - 349
                      0.06939                       0.05978                       0.11253                       0.41539
          CD4_WHO350 - 500          CD4_WHOMore than 500                      VL_200< 200
                      0.62328                       0.70696                      -0.03036
```

```
Degrees of Freedom: 9556 Total (i.e. Null);  9534 Residual
Null Deviance:      12670
Residual Deviance: 11580        AIC: 11620
```

in this MI model, all **9556** observations are included

MI: 9556 / **9556** (100%) retained



Compare CCA versus MI Results

Complete case analysis model

Variable		Fully adjusted		
		OR[1]	95% CI	p-value
Age	9,485			
19-29		1		
30-39		1.22	0.48, 3.24	0.68
40-49		**2.24**	**0.89, 5.92**	**0.09**
≥ 50		1.68	0.65, 4.53	0.29
Sex	9,483			
Male		1		
Female		1.41	0.90, 2.23	0.13
Education	9,418			
Never went to school		1		
Primary		1.52	0.82, 2.83	0.18
Secondary		0.85	0.50, 1.44	0.55
High School		1.19	0.72, 1.98	0.49
University		0.93	0.63, 1.39	0.73
Smoking	9,423			
Never smoked		1		
Current smoker		0.93	0.34, 2.48	0.88
Former smoker		0.82	0.45, 1.45	0.49
Alcohol consumption	7,211			
Never		1		
Monthly or less		1.12	0.72, 1.72	0.62
2-3 times per month		1.38	0.73, 2.65	0.32
Once a week or more		1.31	0.78, 2.21	0.31
WHO stage	8,141			
WHO Stage I		1		
WHO Stage II		0.63	0.40, 0.98	0.04
WHO Stage III		0.97	0.65, 1.44	0.88
WHO Stage IV		1.02	0.74, 1.41	0.90
Time since HIV diagnosis, (years)	8,757	1.07	1.01, 1.12	0.01\
ART use	9,441			
no		1		
yes		0.83	0.55, 1.25	0.38
CD4 level, cells/mm^3	3,418			
Less than 200		1		
200 - 349		2.66	1.55, 4.61	<0.001
350 - 500		2.11	1.23, 3.66	0.007
More than 500		3.02	1.81, 5.11	<0.001
Viral load level, copies/mL	3,475			
≥ 200		1		
< 200		1.04	0.68, 1.59	0.86

Multiple imputed model

Variable	Fully adjusted		
	OR[1]	95% CI	p-value
Age			
19-29	1		
30-39	2.04	1.72, 2.42	<0.001
40-49	2.73	2.30, 3.23	<0.001
≥ 50	2.30	1.92, 2.76	<0.001
Sex			
Male	1		
Female	1.71	1.53, 1.92	<0.001
Education			
Never went to school	1		
Primary	1.19	1.02, 1.38	0.028
Secondary	1.05	0.93, 1.20	0.426
High School	0.88	0.77, 1.00	0.045
University	0.94	0.85, 1.04	0.220
Smoking			
Never smoked	1		
Current smoker	0.63	0.48, 0.81	<0.001
Former smoker	0.83	0.71, 0.97	0.018
Alcohol consumption			
Never	1		
Monthly or less	1.28	1.12, 1.46	<0.001
2-3 times per month	1.84	1.57, 2.17	<0.001
Once a week or more	1.84	1.56, 2.16	<0.001
WHO stage			
WHO Stage I	1		
WHO Stage II	0.65	0.57, 0.74	<0.001
WHO Stage III	1.18	1.06, 1.31	0.003
WHO Stage IV	1.08	0.98, 1.19	0.143
Time since HIV diagnosis, (years)	1.06	1.04, 1.07	<0.001
ART use			
no	1		
yes	1.10	0.99, 1.23	0.068
CD4 level, cells/mm^3			
Less than 200	1		
200 - 349	1.64	1.39, 1.94	<0.001
350 - 500	2.03	1.69, 2.44	<0.001
More than 500	2.47	2.11, 2.89	<0.001
Viral load level, copies/mL			
≥ 200	1		
< 200	0.93	0.75, 1.15	0.481

Notice differences in:

- The estimate (ORs) sizes
- The 95% CI widths
- The p-values for each variable for CCA versus MI

Key takeaways:

- **CCA wastes data** → weaker results
- **MI keeps all data** → stronger, more precise estimates

- **CCA loses power** → wide CIs, unstable p-values
- **MI improves power** → narrow CIs, smaller p-values, more reliable associations

Sensitivity Analysis in Missing Data

1. It is difficult to know the true missingness mechanism, so:

2. Check if results change when:

Comparing missing vs observed groups	Using different imputation methods/models (add/remove predictors)

3. Outcomes

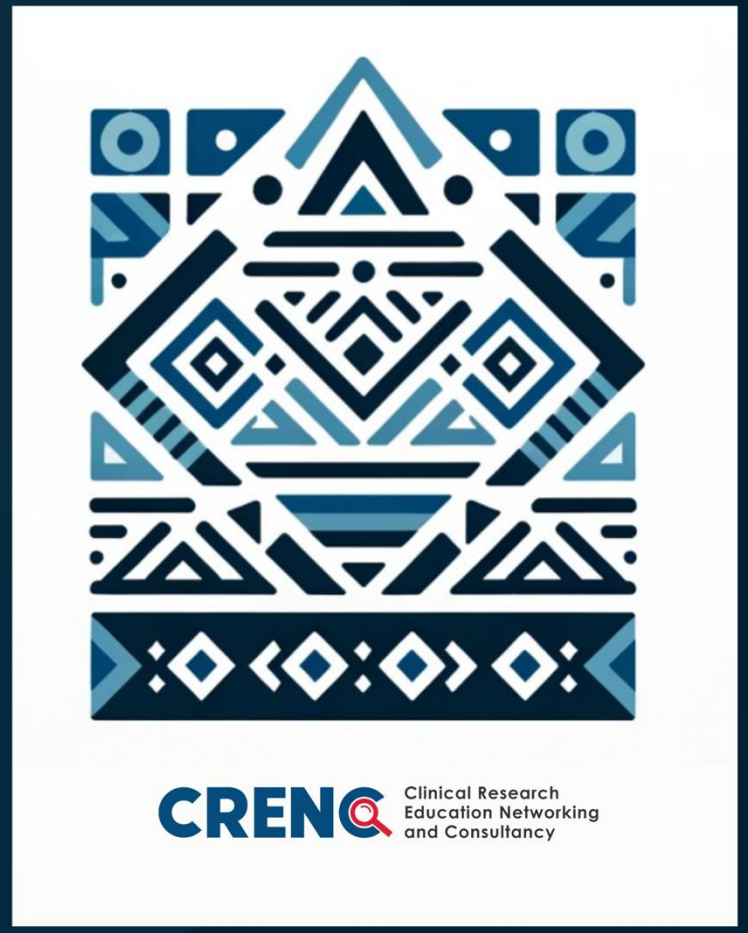
Changing results = less reliable conclusions.	Stable results = more confidence.

RESEARCH Open Access

Reporting and handling of missing data in published studies of co-morbid hypertension and diabetes among people living with HIV/AIDS: a systematic review

Peter Vanes Ebasone[1,2*], Nasheeta Peer[1,3], Anastase Dzudie[2,4,5], Johney Melpsa[2], Merveille Foaleng[2] and Andre Pascal Kengne[1,2,3]

- We found that only 34.4% of studies reported missing data.
- Missingness was mostly in the exposure variables, notably, CD4 count and viral load.
- Few studies discussed how missingness biased results and conclusions.
- Of the studies that reported missing data, less than half of these studies reported how they handled missing data, and for those that did, they largely used complete case analysis followed by multiple imputation methods.



Bibliography

1. Thakur, P. (2025, Apr 14). *Mastering Missing Data: A Comprehensive Guide to Handling Gaps in Your Dataset*. Medium. https://medium.com/@preethithakur/mastering-missing-data-a-comprehensive-guide-to-handling-gaps-in-your-dataset-86af800fcbdd

2. Alayo, B. (2024, Apr 27). *Missing Data: Causes, Types, and Handling Techniques*. LinkedIn. https://www.linkedin.com/pulse/missing-data-causes-types-handling-techniques-bilikis-alayo-ho9if

3. Chawla, A. (2024, Jul 7). *3 Types of Missing Values… and how to impute them*. Daily Dose of Data Science. https://blog.dailydoseofds.com/p/3-types-of-missing-values

4. Poppelaars, E. (2019, Jul 21). *How to Manage Missing Data, for Data Scientists*. LinkedIn. https://www.linkedin.com/pulse/how-manage-missing-data-scientists-eefje-poppelaars

5. Dancuk, M. (2021, Jul 1). *Handling Missing Data in Python: Causes and Solutions*. phoenixNAP. https://phoenixnap.com/kb/handling-missing-data-in-python

6. *The Epidemiologist R Handbook*. (Last updated Sep 18, 2024). https://epirhandbook.com/en/

7. El-Masri, M.M., & Fox-Wasylyshyn, S.M. (2005). Missing data: an introductory conceptual overview for the novice researcher. *Canadian Journal of Nursing Research, 37*(4), 156–171. PMID: 16541824.

8. Ebasone, P.V., Peer, N., Dzudie, A. et al. (2025). Reporting and handling of missing data in published studies of co-morbid hypertension and diabetes among people living with HIV/AIDS: a systematic review. *BMC Medical Research Methodology, 25*, 180. https://doi.org/10.1186/s12874-025-02630-1

Thanks You

CRENC Clinical Research Education Networking and Consultancy