GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
Instituto de Salud Carlos III
CENTRO NACIONAL DE EPIDEMIOLOGÍA
RETIC-RIS Red Española de Investigación en SIDA (RIS)
REDES TEMÁTICAS DE INVESTIGACIÓN COOPERATIVA
CoRIS Cohorte de la Red de Investigación en Sida

Dealing with missing data in practice: Methods, applications, and implications for HIV cohort studies

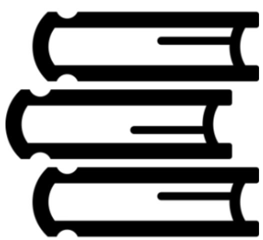
Belen Alejos Ferreras
Centro Nacional de Epidemiología
Instituto de Salud Carlos III

19 de Octubre de 2017

What is Missing or Incomplete data?

What is Missing or Incomplete data?

Missing or Incomplete data



Data that were intended to collect on observations but that due to different reasons were not collected

V1	V2	V3	V4
X	.	X	X
X	X	X	.
X	X	.	X
X	X	X	.

Do I need to be worried about missing data?

No universal rule to indicate **the proportion** of missing data producing **bias** or to **invalid** results

The **success** of a statistical analysis in the presence of missing data will depend on the reasons why data are missing (**<u>missing data mechanisms</u>**)

Which Missing data mechanisms are there?

Which Missing data mechanisms are there?

- Missing Completly At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

Missing data mechanisms

Missing completely at random (MCAR)
There is no relationship between whether an observation is missing and the unseen value nor to any values (observed or missing)

$$P(R|Y) = P(R)$$

Missing at random (MAR)
There is no relationship between whether an observation is missing and the unseen value, but it is related to some of the observed data

$$P(R|Y) = P(R|Y_{obs})$$

Missing not at random (MNAR)
Whether an observation is missing depends on the unseen value itself

R=missing data point ; Y=Variables

Methods to deal with missing data

If it is not possible to get the original value

… it is necessary to face the problem with statistical techniques

Methods to deal with missing data

Ad-hoc or conventional

Complete- Case (CC)
Indicator Method (IM)
Simple mean or regression mean imputation
Stochastic regression imputation

- **Easy** implementation
- No specific **software**
- Not based on statistical principles
- Might produce **biased** results and **loss of power**

Methods to deal with missing data

Ad-hoc or conventional

Complete- Case (CC)
Indicator Method (IM)
Simple mean or regression mean imputation
Stochastic regression imputation

- **Easy** implementation
- No specific **software**
- Not based on statistical principles
- Might produce **biased** results and **loss of power**

Advanced or complex

Multiple Imputation by Chained Equations (MICE)
Maximum likelihood estimation
Bayesian Methods
Inverse Probability weighting

- **Maximize** use of available information
- **More precise** results (higher statistical power)
- Depend on missing **data mechanism**
- Some not implemented in statistical **software**

Methods to deal with missing data

Ad-hoc or conventional

Complete- Case (CC)
Indicator Method (IM)
Simple mean or regression mean imputation
Stochastic regression imputation

- **Easy** implementation
- No specific **software**
- Not based on statistical principles
- Might produce **biased** results and **loss of power**

Advanced or complex

Multiple Imputation by Chained Equations (MICE)
Maximum likelihood estimation
Bayesian Methods
Inverse Probability weighting

- **Maximize** use of available information
- **More precise** results (higher statistical power)
- Depend on missing **data mechanism**
- Some not implemented in statistical **software**

Complete-Cases

Consists of restricting the statistical analyses to the cases with complete information for all the variables in the model

Original			
ID	**Outcome**	**Variable**	**Complete-Case**
1	5	4	Yes
2	~~4~~	~~.~~	~~No~~
3	~~.~~	~~2~~	~~No~~
4	~~3~~	~~.~~	~~No~~
5	4	5	Yes

Complete-cases			
ID	**Outcome**	**Variable**	**Complete-Case**
1	5	4	Yes
5	4	5	Yes

Indicator method

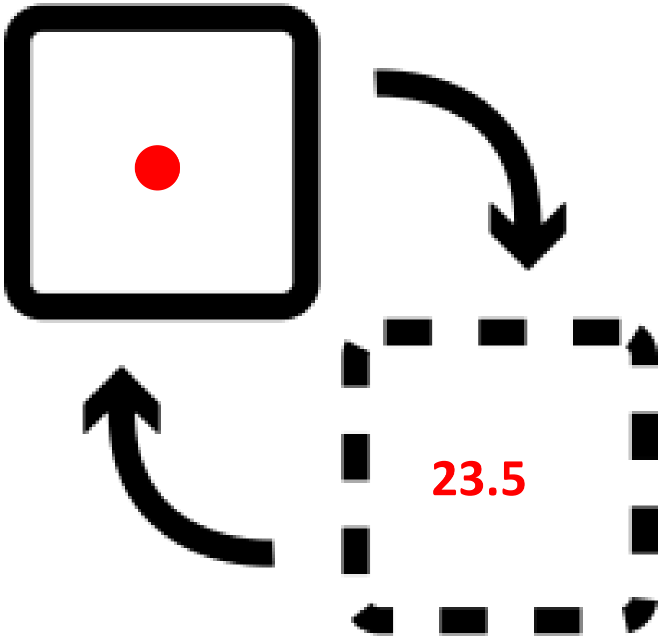
Creates an extra category for missing values in each incomplete, independent and categorical variable and therefore all the observations are included in the analyses

Original			
ID	**Outcome**	**Variable**	**Complete-Case**
1	5	0	1
2	4	.	0
3	4	1	1
4	3	.	0
5	4	1	1

Indicator Method			
ID	**Outcome**	**Variable**	**Complete-Case**
1	5	0	1
2	4	9	0
3	4	1	1
4	3	9	0
5	4	1	1

Simple imputation methods

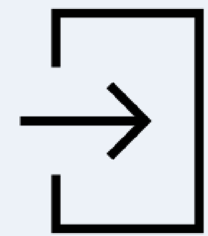
The information collected in the sample is used to assign one value to those variables with missing values



Simple imputation methods

Simple mean imputation

replaces each missing observation by the completers mean



Regression mean imputation

replaces each missing observation with the predicted values from a regression model

Random or stochastic regression imputation

to create an imputed value, an appropriate random residual is added to the value predicted using regression mean imputation.

Simple imputation methods

Simple mean imputation
replaces each missing observation by the completers mean

Regression mean imputation
replaces each missing observation with the predicted values from a regression model

Random or stochastic regression imputation
to create an imputed value, an appropriate random residual is added to the value predicted using regression mean imputation.

Simple imputation methods

PROBLEM:
- Underestimated variances

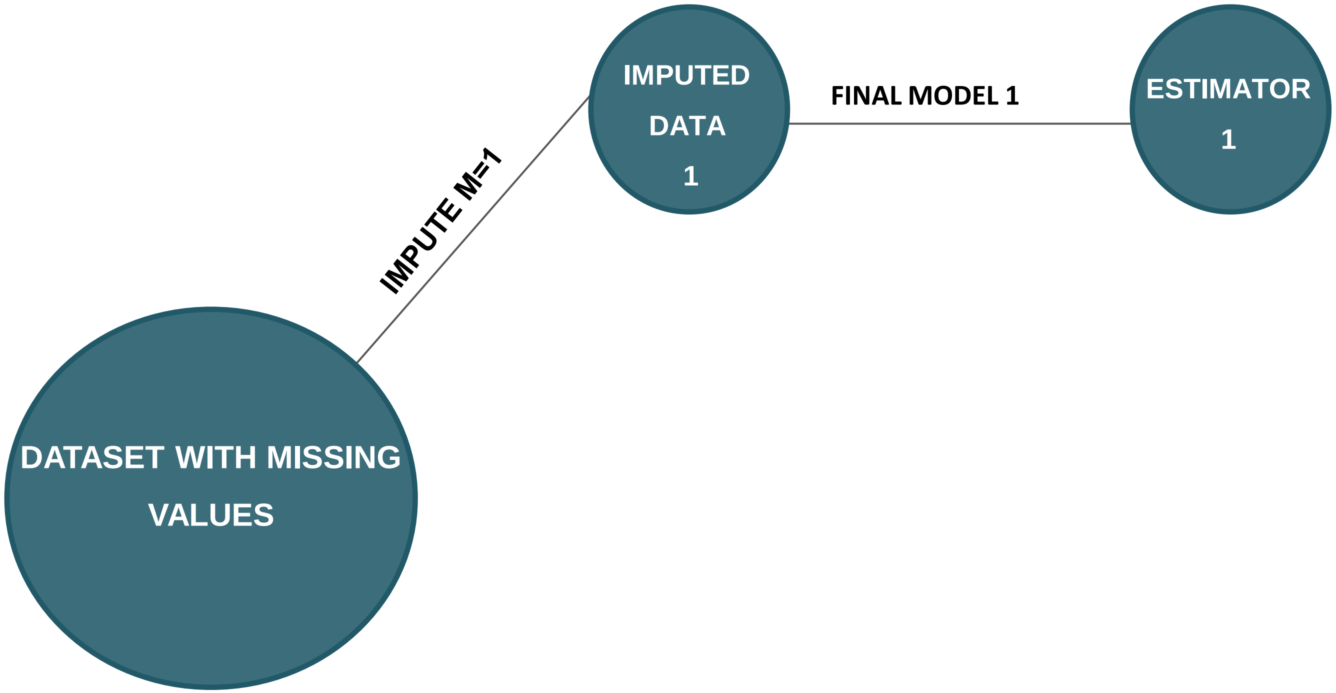
SOLUTION:
Multiple Imputation

Multiple Imputation methods

Imputation techniques that assign several imputed values to each missing value using the following procedure:

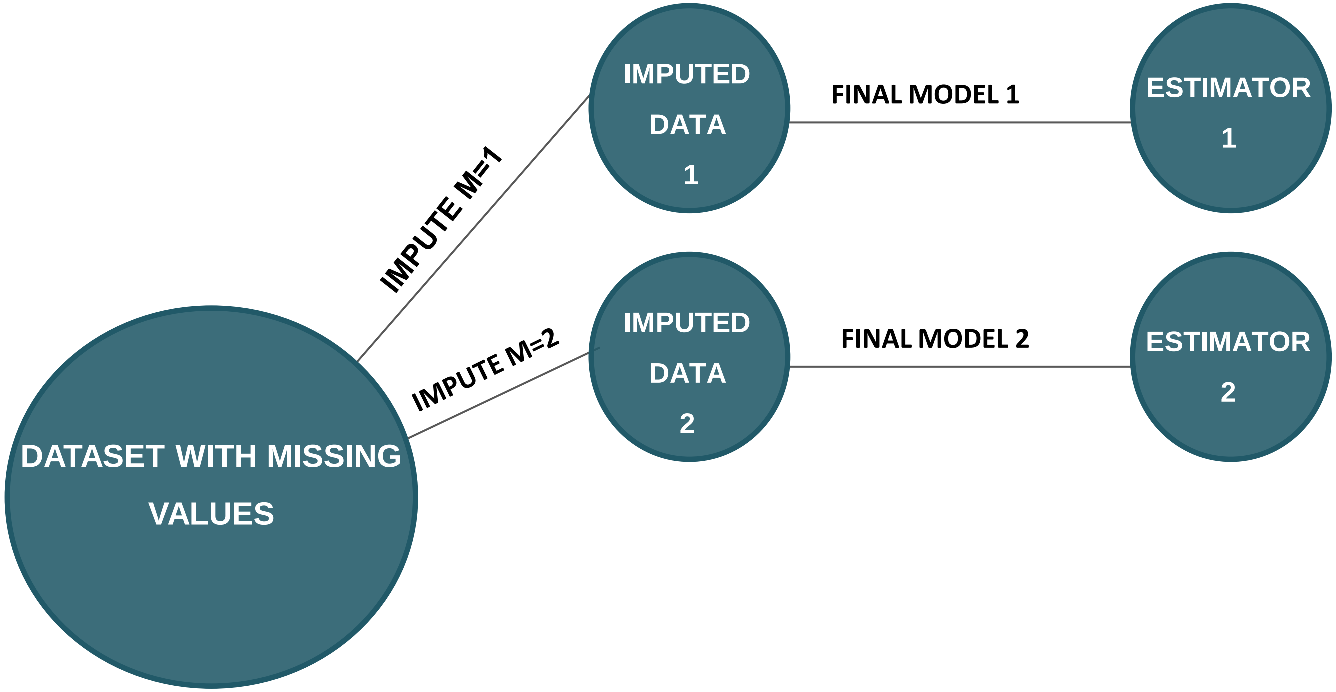
Multiple Imputation methods

Imputation techniques that assign several imputed values to each missing value using the following procedure:



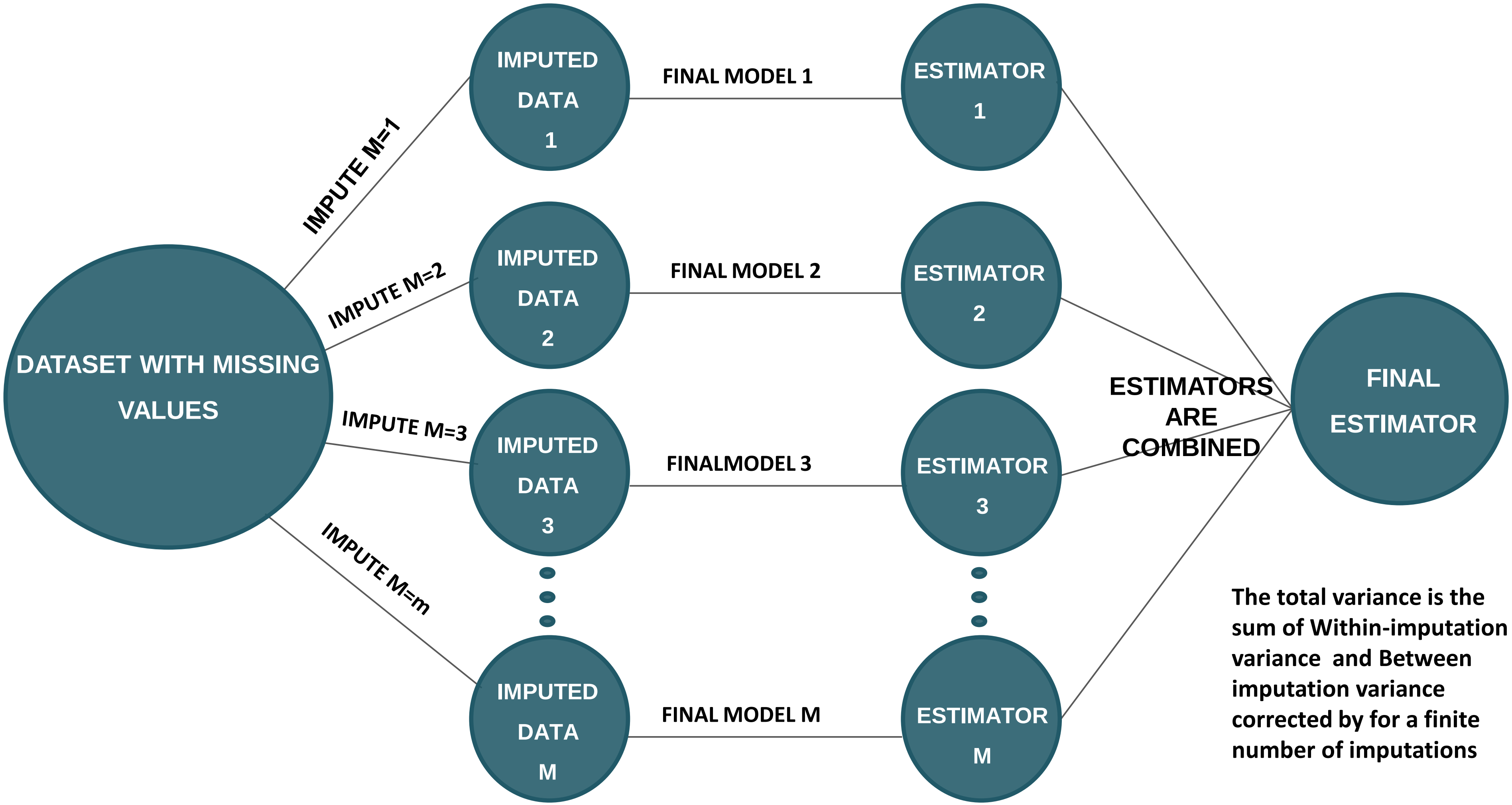
Multiple Imputation methods

Imputation techniques that assign several imputed values to each missing value using the following procedure:



Multiple Imputation methods

Imputation techniques that assign several imputed values to each missing value using the following procedure:



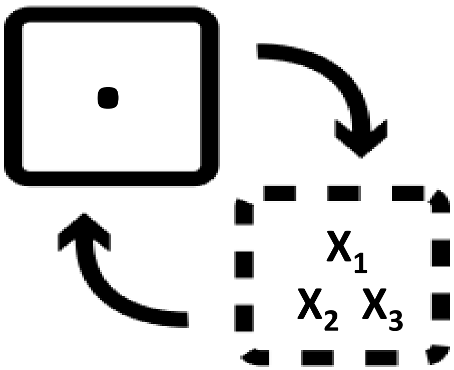
Multiple Imputation by Chained Equations (MICE)

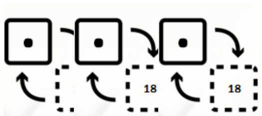
Multiple Imputation methods

Multiple Imputation by Chained Equations (MICE)

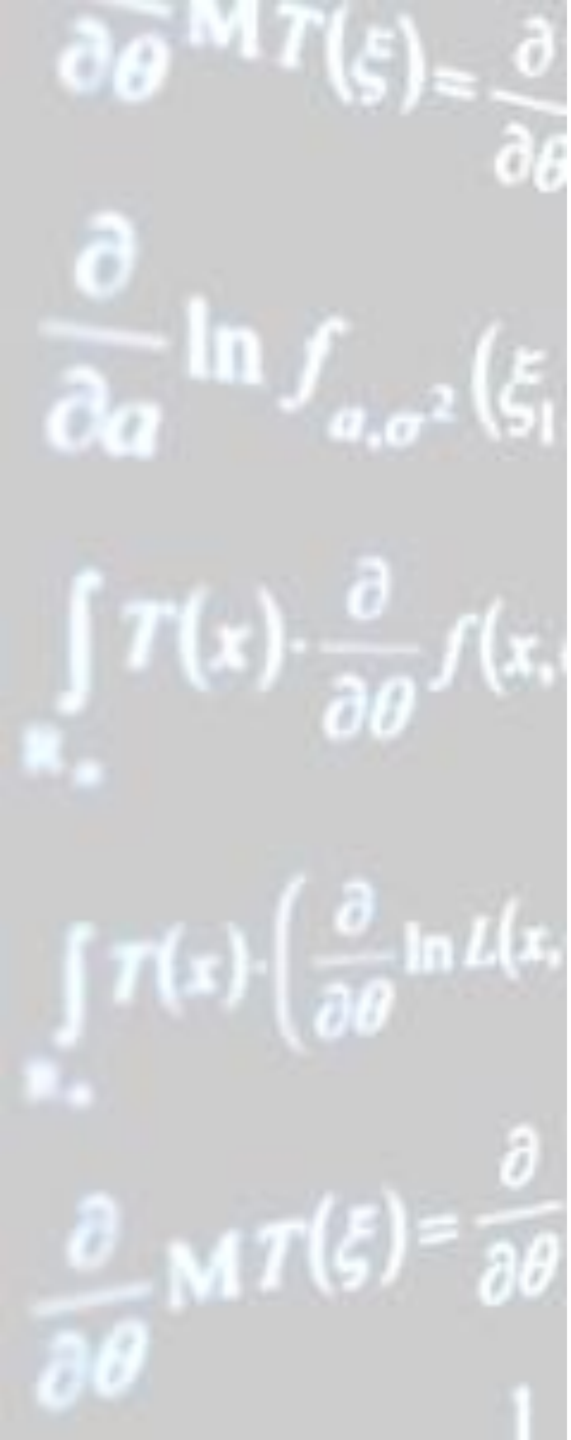
A particular multiple imputation technique that allows to **impute missing values in multiple variables** under MAR assumption. Logistic, multinomial or ordered regression can be used instead linear regression for non-normal variables

Missing values in X_1 , X_2 , X_3





Multiple Imputation: The complete process is repeated *m* times



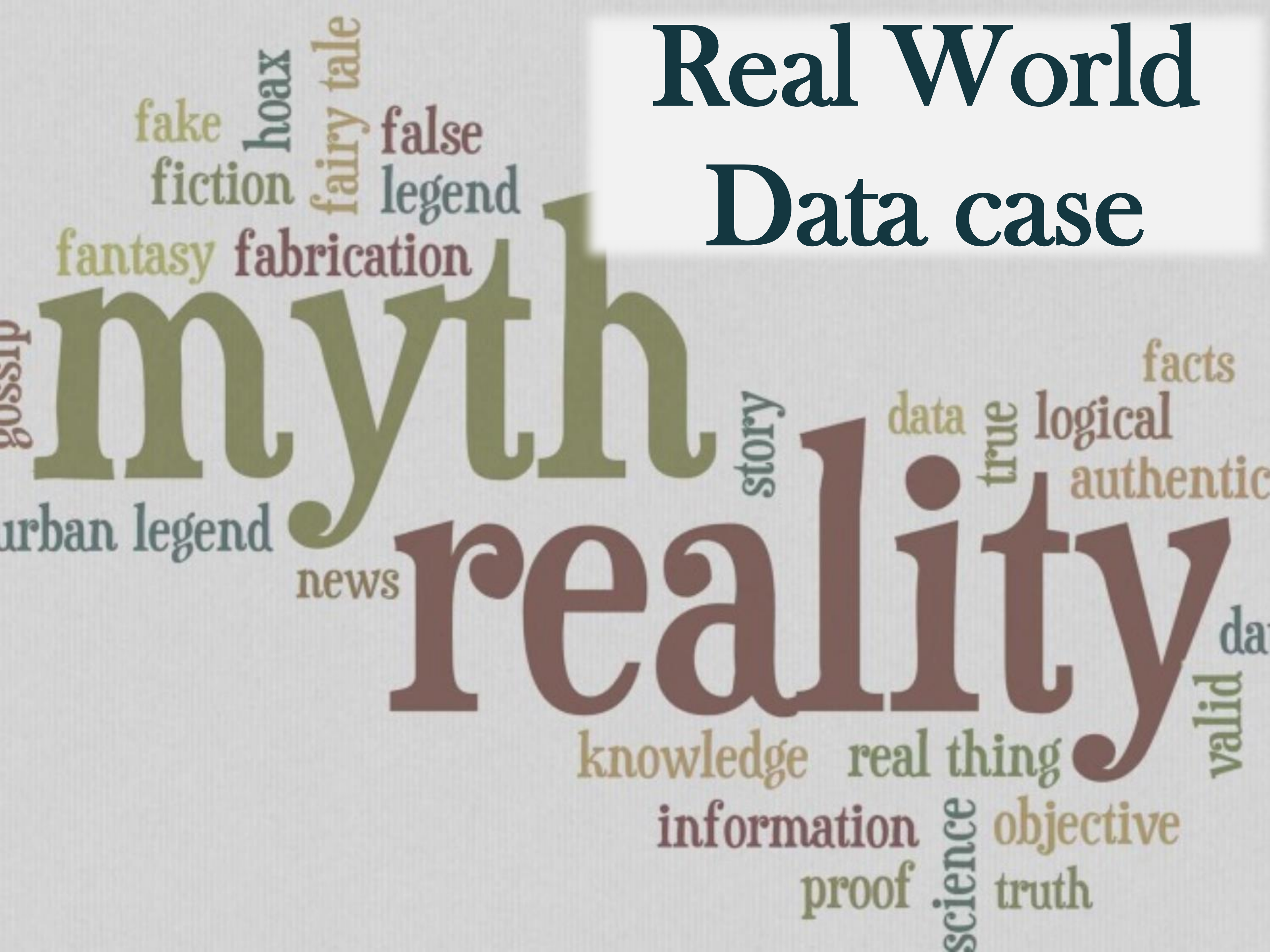
Other advanced methods

Maximum likelihood estimation
models simultaneously the outcome and the reason why data are missing

Bayesian methods
estimate a statistical model for full data (including missingness mechanism and the outcome)

Inverse Probability Weighting
calculates the predicted probability for certain variable to be observed of each patient and use these weights in the outcome model

Real World Data case



Different Approaches to Account for Missing Data in a Cohort of HIV-Positive Patients

CoRIS

To compare three different methods to deal with missing data in both outcome (cause of death) and covariates in a cohort of HIV-Positive patients (CoRIS)

- **CoRIS (**N=10,469)
- **Cancer mortality**

Poisson regression mortality rates and rate ratios for the effect of Hepatitis C Virus coinfection

- Complete-case
- Indicator- Method
- MICE

Missing data Summary

```
. misstable sum CD4_6M VL_6M EDUCATION HIV_RISK ORIGIN HCV_6M CoD AIDS survtime
age sex
                                                               Obs<.
                                                +------------------------------
               |                                | Unique
      Variable |     Obs=.     Obs>.     Obs<.  | values          Min          Max
  -------------+--------------------------------+------------------------------
        CD4_6M |       787               9,682  |   >500            0         8246
         VL_6M |       823               9,646  |   >500            0     6.54e+07
     EDUCATION |     1,371               9,098  |      4            0            8
      HIV_RISK |       246              10,223  |      4            1           90
        ORIGEN |       220              10,249  |      4            0            3
        HCV_6M |     1,103               9,366  |      2            0            1
         CoD   |        49              10,420  |      6            0            5
  -----------------------------------------------------------------------------
```

Variables: AIDS survtime age sex are complete

Missing data Summary

```
. misstable patterns CD4_6M VL_6M EDUCATION HIV_RISK ORIGIN HCV_6M CoD AIDS_6
survtime age sex, freq

Missing-value patterns
(1 means complete)

                  |    Pattern
    Frequency     |   1   2   3   4     5   6   7
  ----------------+------------------------
  7,382 (71%)     |   1   1   1   1     1   1   1
                  |
          889     |   1   1   1   1     1   1   0
          699     |   1   1   1   1     1   0   1
          434     |   1   1   1   0     0   1   1
          166     |   1   1   1   1     1   0   0
          117     |   1   1   0   1     1   1   1
                …
                …
                …

 Variables are  (1) CoD  (2) origen  (3) HIV_RISK  (4) CD4_6M  (5) VL_6M  (6)
HCV_6M (7) EDUCATION
```

Complete-Cases

```
. use mortality_data, clear

. mi set flong
. mi register imputed CD4_6M VL_6M HIV_RISK origin CoD EDUCATION HCV_6M
. keep if _mi_miss==0
. mi unset

. stset survtime, fail(CoD==2) scale(365.25)
. strate , per(1000)

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(7384 records included in the analysis)
  +-----------------------------------------------+
  |  D          Y       Rate      Lower     Upper |
  |-----------------------------------------------|
  | 32    26.6981    1.19859    0.84761   1.69489 |
  +-----------------------------------------------+

. gen tpo =_t-_t0
. poisson _d i.HCV_6M , exp(tpo) irr

Poisson regression                              Number of obs     =      7,384
                                                LR chi2(1)        =      10.70
                                                Prob > chi2       =     0.0011
Log likelihood = -219.82597                     Pseudo R2         =     0.0238

------------------------------------------------------------------------------
          _d |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      HCV_6M |
    Positive |   3.640965   1.329493     3.54   0.000     1.779925    7.447859
       _cons |   .0008726   .0001951   -31.50   0.000     .0005629    .0013525
     ln(tpo) |          1  (exposure)
------------------------------------------------------------------------------
```

Indicator method

```
. use mortality data, clear

. recode CD4_6M VL_6M HIV_RISK origin CoD EDUCATION HCV_6M (. =9)

. stset survtime, fail(CoD==2) scale(365.25)
. strate , per(1000)

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(10469 records included in the analysis)

  +--------------------------------------------+
  |  D        Y       Rate     Lower     Upper |
  |--------------------------------------------|
  | 52   37.4372    1.3890    1.0584    1.8228 |
  +--------------------------------------------+

. gen tpo =_t-_t0
. poisson _d i.HCV_6M , exp(tpo) irr

Poisson regression                              Number of obs     =     10,469
                                                LR chi2(2)        =       9.48
                                                Prob > chi2       =     0.0087
Log likelihood = -359.94411                     Pseudo R2         =     0.0130

------------------------------------------------------------------------------
          _d |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      HCV_6M |
   Positive  |   2.792667   .8831188     3.25   0.001     1.502608    5.190303
    Unknown  |   1.622859    .681196     1.15   0.249     .7128344    3.694649
             |
       _cons |     .00106   .0001935   -37.52   0.000     .0007412    .0015161
     ln(tpo) |          1  (exposure)
```

MICE

Variables with missing values						
Education	**Mode**	**Origin**	**CD4**	**VL**	**HCV**	**CoD**

MAR

- Several predictors for the probability of being missing in each covariate
- No evidence against assuming data are MAR

Multiple imputation model for each variable with missing values including:

- Other incomplete variables (education, mode, origin, CD4, VL, HCV & CoD)
- Complete variables (AIDS at entry, age and sex)
- The outcome (log survival time and CoD)

MICE

```
. use mortality_data, clear
. gen lsurvtime=log(survtime)

. mi set flong
. mi register imputed CD4_6M VL_6M HIV_RISK origin CoD EDUCATION HCV_6M
. mi register regular  AIDS_6M lsurvtime TRAN_AGE sex

. mi impute chained ///
(regress, include (i.AIDS_6M c.lsurvtime TRAN_AGE i.sex)) TRAN_CV_6M ///
(regress, include (i.AIDS_6M c.lsurvtime TRAN_AGE i.sex)) TRAN_CD4_6M ///
(mlogit, include (i.AIDS_6M c.lsurvtime TRAN_AGE i.sex))  origin ///
(mlogit, include (i.AIDS_6M c.lsurvtime TRAN_AGE i.sex))  HIV_RISK ///
(mlogit, conditional(if exitus==1) include (i.AIDS_6M c.lsurvtime TRAN_AGE  i.sex )) CoD ///
(ologit, include (i.AIDS_6M c.lsurvtime TRAN_AGE )) EDUCATION  ///
(logit, include (i.AIDS_6M c.lsurvtime TRAN_AGE i.sex )) HCV_6M  ///
, add(12) rseed(10)  burnin(10)  augment savetrace(impstats,replace)

Conditional models:
            CoD: mlogit CoD i.origen i.HIV_RISKTRAN_CD4_6M TRAN_VL_6M i.HCV_6M
                     i.EDUCATION i.AIDS_6M lsurvtime i.sex , augment conditional(if exitus==1)
      origen: mlogit origen i.CoD i.HIV_RISKTRAN_CD4_6M TRAN_VL_6M i.HCV_6M
                     i.EDUCATION i.AIDS_6M lsurvtime i.sex , augment
         HIV_RISK: mlogit HIV_RISKi.CoD i.origen TRAN_CD4_6M TRAN_VL_6M i.HCV_6M
                     i.EDUCATION i.AIDS_6M lsurvtime i.sex , augment
       TRAN_CD4_6M: regress TRAN_CD4_6M i.CoD i.origen i.HIV_RISK TRAN_VL_6M i.HCV_6M
                     i.EDUCATION i.AIDS_6M lsurvtime i.sex
        TRAN_VL_6M: regress TRAN_VL_6M i.CoD i.origen i.HIV_RISK TRAN_CD4_6M i.HCV_6M
                     i.EDUCATION i.AIDS_6M lsurvtime i.sex
            HCV_6M: logit HCV_6M i.CoD i.origen i.HIV_RISK TRAN_CD4_6M TRAN_VL_6M
                     i.EDUCATION i.AIDS_6M lsurvtime i.sex , augment
         EDUCATION: ologit EDUCATION i.CoD i.origen i.HIV_RISK TRAN_CD4_6M TRAN_VL_6M
                     i.HCV_6M i.AIDS_6M lsurvtime i.sex , augment
```

MICE

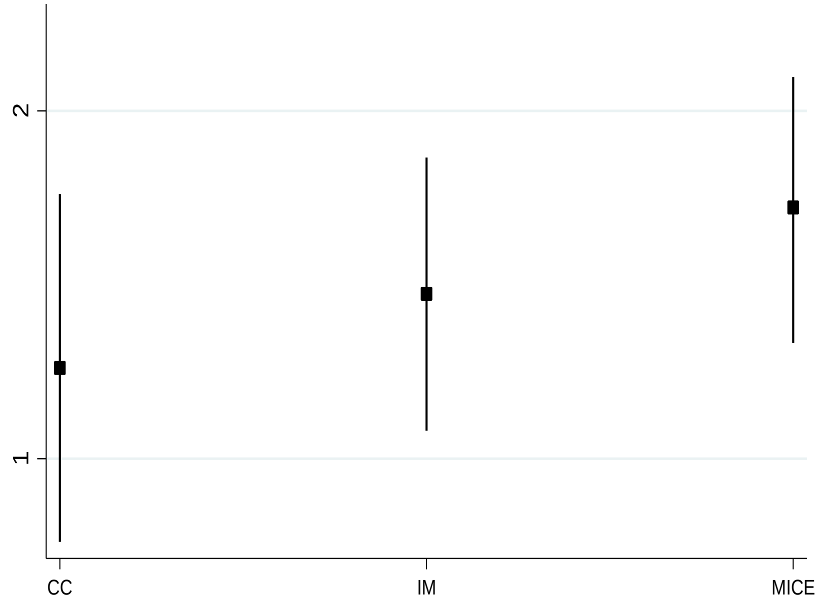
```
. gen tpo= (L_ALIVE-ENROL_D)/365.25

. mi estimate , irr: poisson cause_tumo , exp(tpo)

Multiple-imputation estimates                 Imputations       =         12
Poisson regression                            Number of obs     =     10,469
                                              Average RVI       =     0.1166
                                              Largest FMI       =     0.1062
                                              DF:     min       =   1,009.20
                                                      avg       =   1,009.20
DF adjustment:   Large sample                         max       =   1,009.20
                                              F(   0,       .)  =          .
Within VCE type:          OIM                 Prob > F          =          .

------------------------------------------------------------------------------
  cause_tumo |        IRR   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .0016503   .0002219   -47.64   0.000     .0012675    .0021487
     ln(tpo) |          1  (exposure)
------------------------------------------------------------------------------

. mi estimate , irr: poisson cause_tumo i.HCV_6M, exp(tpo)
…
…
------------------------------------------------------------------------------
  cause_tumo |        IRR   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      HCV_6M |
   Positive  |   2.593291   .7609617     3.25   0.001     1.457445    4.614347
       _cons |   .0013245   .0002133   -41.15   0.000     .0009657    .0018165
     ln(tpo) |          1  (exposure)
------------------------------------------------------------------------------
```

Which is the best method to deal with missing data?

Complete-case (CC)
N=7,384
n=32

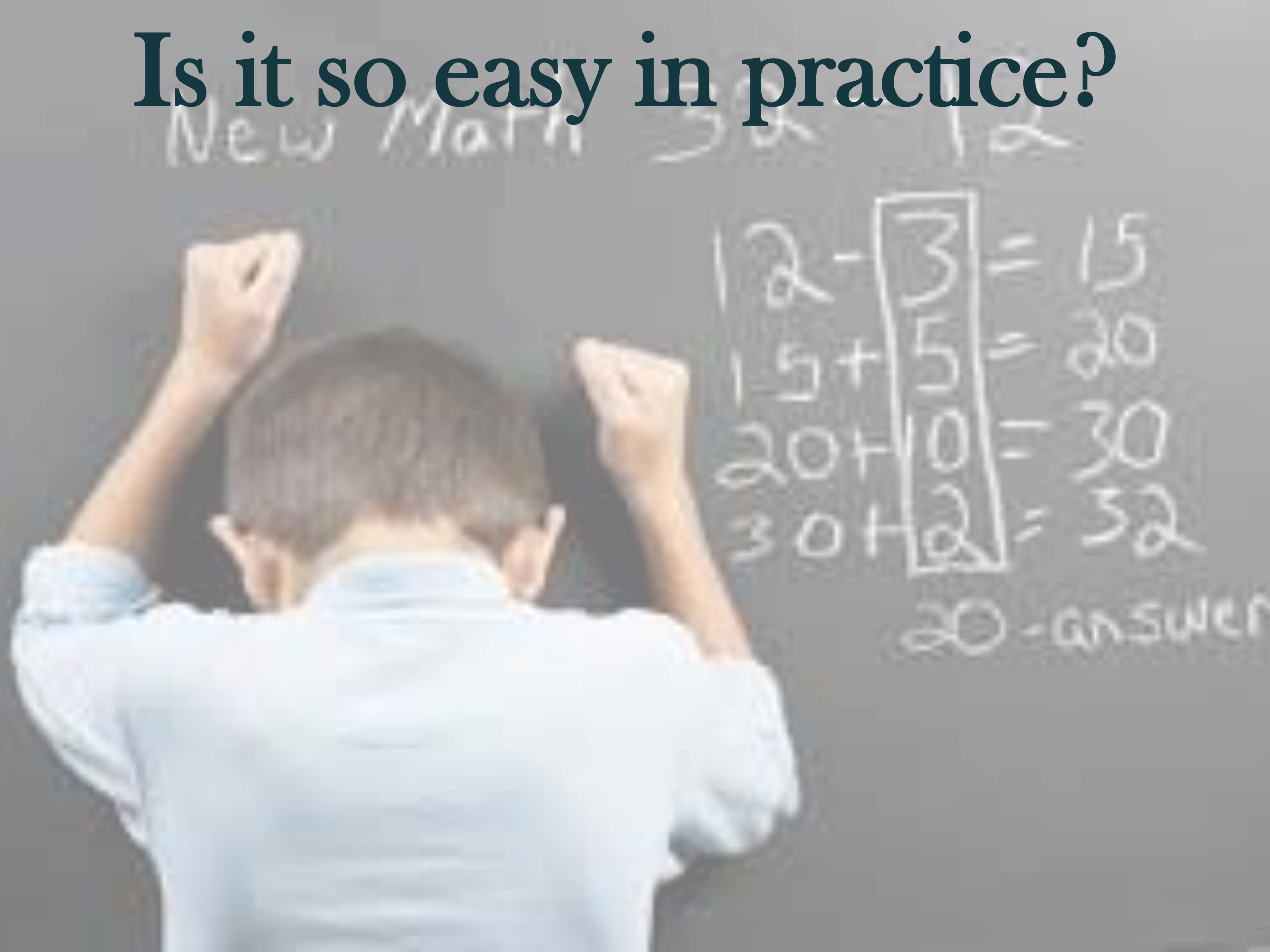
Indicator Method (IM)
N=10,469
n=52

MICE
N=10,469
n=62



	CC	IM	MICE
Death rate x1000	1.20 (0.84; 1.69)	1.39 (1.06; 1.82)	1.65 (1.26; 2.14)
HCV rate ratio	3.64 (1.78; 7.45)	2.79 (1.50; 5.19)	2.59 (1.46; 4.61)

Is it so easy in practice?



Dealing with missing data in practice….

Difficulties with….

Interactions

It is not possible to include interactions between variables with missing data in the imputation model

Interaction II

. mi estimate: lincom not working

. mi stset

Not working when the outcome has been imputed

Conclusions

Conclusions

- STATA provides multiple options to deal with missing data
- In our case-study of an HIV cohort, the application of different methods to deal with missing data in both covariates and cause of death did not produce results that differed to the extent that would vary the fundamental interpretation of the study conclusions
- MICE is a powerful approach. However, it rests on the assumption that incomplete values are Missing At Random


!Muchas gracias!
Thank you very much!