Missing Data Methods for (Un)commonly Used Statistics

Emile Latour and Miguel Marino

Oregon Health & Science University

@emilelatour

@MmMiguelmM

About Us: Miguel Marino, PhD

- Associate Professor of Biostatistics, Department of Family Medicine at Oregon Health & Science University (OHSU)
- Dr. Marino has co-authored over 145 peer-reviewed publications and has served as co-investigator/site PI in over 20 federally-funded grants from a diverse set of funders (e.g. NIH, CDC, etc.).
- Dr. Marino currently serves as the Publications Officer for the Health Policy Statistics Section of the ASA and as the statistical editor for the Annals of Family Medicine journal.

About Us: Emile Latour, MS

- Emile Latour MS is a staff biostatistician with the OHSU Knight Cancer Institute working collaboratively in cancer research to provide statistical support and consultation focused in clinical trials and dermatology.
- Areas of statistical interest include missing data, multiple imputation, agreement statistics, and evaluation of diagnostic testing and screenings.
- Strong interest in statistical computing in R, data visualization, reproducible research, and open science.

Sooner or later (usually sooner), anyone who does statistical analysis runs into problems with missing data.

-Allison (2002)

Agenda

- Foundational Review of Important Missing Data Concepts
- Visual Approaches to Understanding Missing Data
- Review of Common Methods
- Missing data Methods for Uncommon Statistics
- Summary, Questions and Answers

Adapting to a Virtual Format

- In the chat box, share:
  1. Name
  2. Affiliation
  3. Why are you interested in missing data?
- If you have a question, enter it in the chat box.
  - One of us will collect the questions and address them when a natural break occurs. We may ask the person asking the question to unmute themselves and share the question with the group
- **Code of conduct from ASA:** https://www.amstat.org/ASA/Meetings/Meeting-Conduct-Policy.aspx

Workshop Materials are Available Online

- Workshop materials include:
  - Workshop Slides
  - Practice Data set
  - R code and HTML slides
  - Stata code
- Materials will be found at:
  - ASA CSP workshop website
  - Emile Latour's GITHUB:
    - https://github.com/emilelatour/CSP-2021-missing-data

Foundational Review of Important Missing Data Concepts

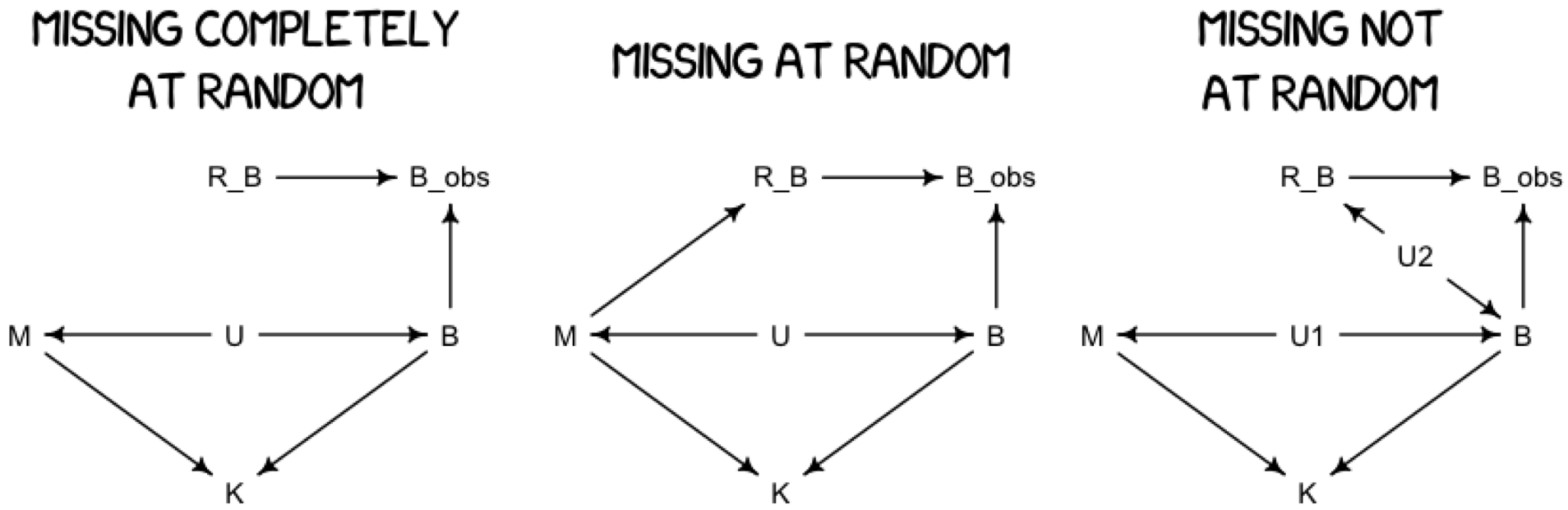
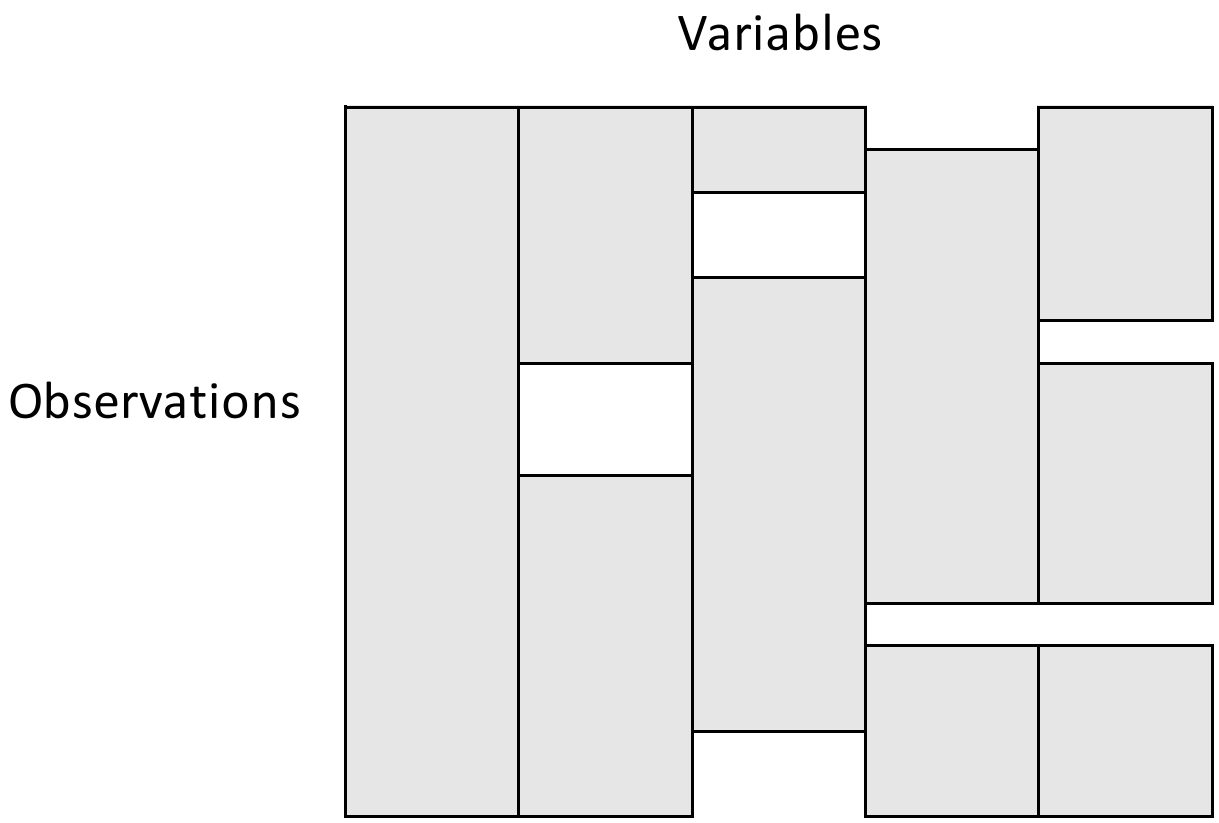


Fig. Source: https://twitter.com/rlmcelreath/status/1100763647943557121?s=20


Variables
Observations

Foundational Review Objectives

- What is missing data?
- Review pattern of missingness and implications
- Review missingness mechanism and why it matters
- How to determine mechanism
- Discuss the impacts of missing data
- Exploratory data analysis of missing data by example with code

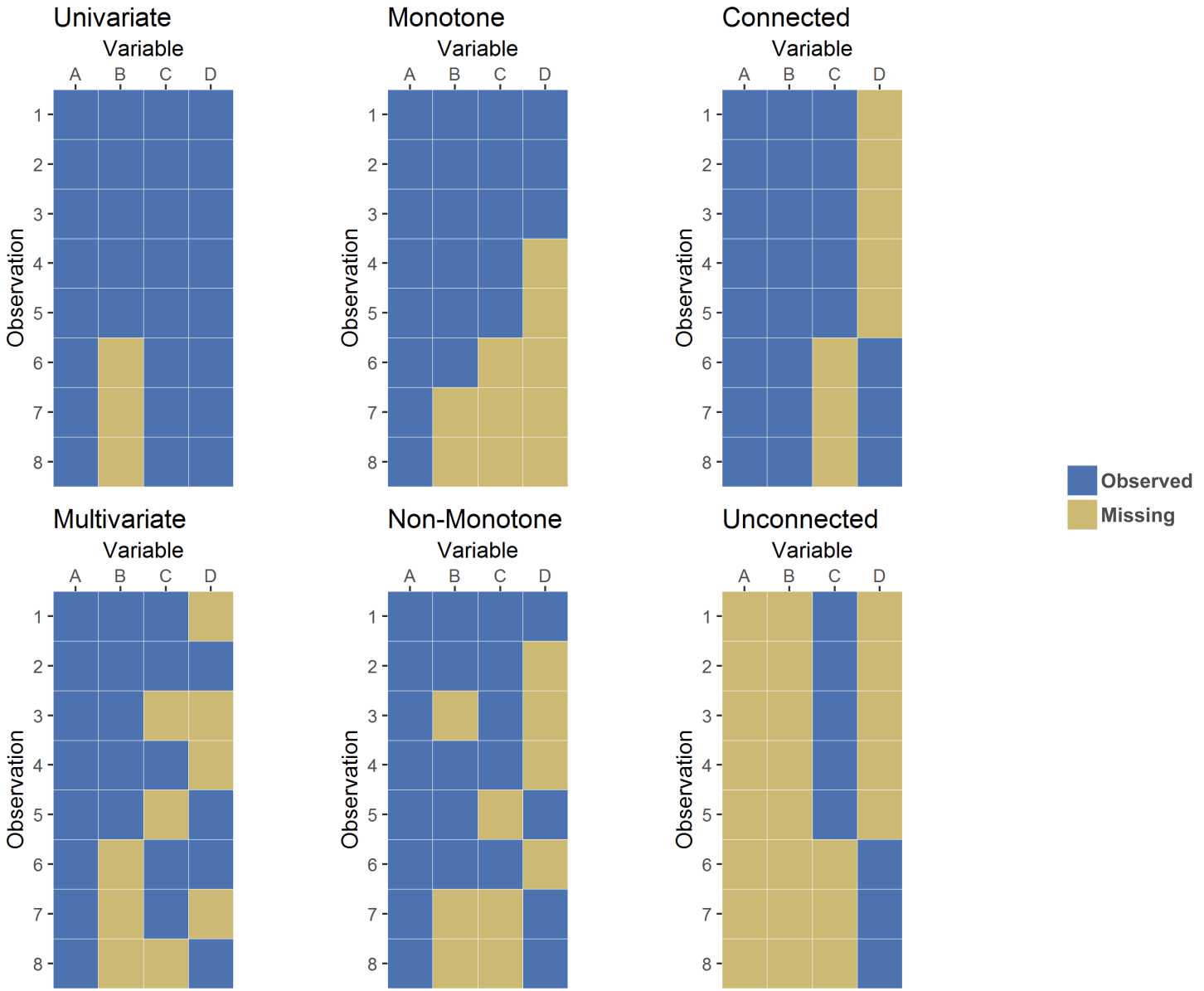
Missing data

- Values that are not available and that *would be meaningful* for analysis had they been observed
- Too often missing data is ignored or edited which leads to problems
- Missing data is here to stay whether we like it or not

Pattern and mechanism

- Working with missing data, need to consider
  - Pattern – which values are observed and which are missing
  - Mechanism – relationship between missingness and values of the variables in the data
- Pattern and mechanism dictate which missing data methods are appropriate

Pattern


Univariate
Monotone
Connected
Multivariate
Non-Monotone
Unconnected
Variable
A
B
C
D
Observation
1
2
3
4
5
6
7
8
Observed
Missing

Pattern

- Consider the horizontal and vertical “moves” that could be made from observed cell to observed cell.
- Multivariate more complicated than univariate
- Monotone typically occurs in longitudinal studies where drop out is an issue. Not covered here today, but there are computational savings with monotone data.
- Connected data are important to understand and address missing data problems. For example, to examine correlation.

Mechanism

- Examines the “reason” for missing values
- Tries to determine whether variables that are missing are related to the underlying values of the variables
- Rubin (1976) classified missing data mechanisms into three categories

Missing completely at random (MCAR)

- The missingness does really occur at random.
- Information in Y cannot predict whether the data are missing or not
- Causes of the missing data are unrelated to the data
- Probability of missing is the same for all cases
- Convenient but often unrealistic. Rarely the case.

Missing completely at random (MCAR)

- Examples of MCAR
  - Subject randomly skips questions on survey
  - Random data entry errors
  - Subject randomly doesn’t show up to appointment to provide data

Missing at random (MAR)

- Probability that observations are missing may depend on observed values but not missing values. Conditional on observed data.
- Possible that missingness can be predicted from other available data
- Known aspect of the data influences missingness
- Most commonly seen

Missing at random (MAR)

- Examples of MAR
  - Subject skips questions that are related to other questions in a survey
  - Subject is missing data that is *related* to observed demographic data
  - Subject that did poorly at a prior data collection point doesn’t come back for follow-up

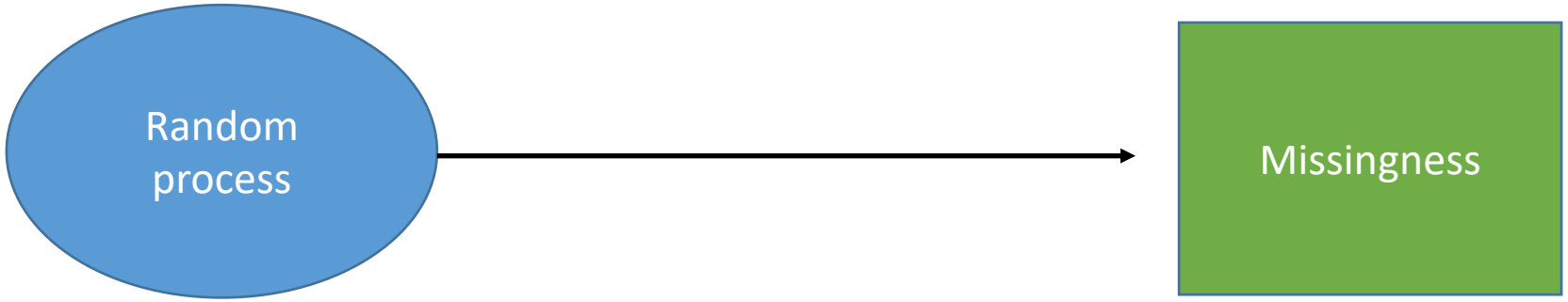
Missing not at random (MNAR)

- Not MCAR and not MAR
- Probability that a missing value is associated with the missing variable itself and with other variables
- The probability of missingness varies for reasons that are unknown
- Missing observations are related to values of unobserved data

Missing not at random (MNAR)

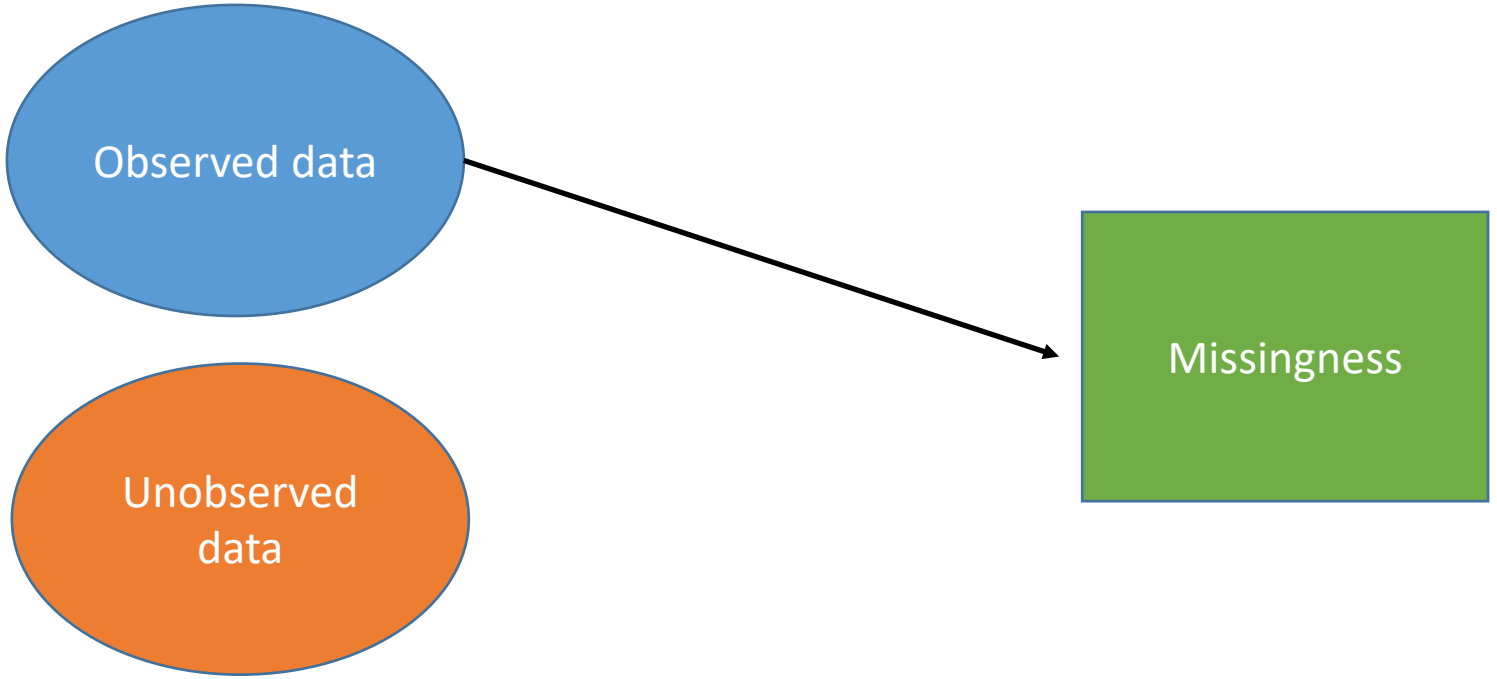
- Examples of MNAR
  - Subject doesn’t respond to incriminating questions
  - Subject is missing data that is *unrelated* to observed demographic data
  - Subject that is doing poorly in a current data collection point stops answering questions

Missing completely at random (MCAR



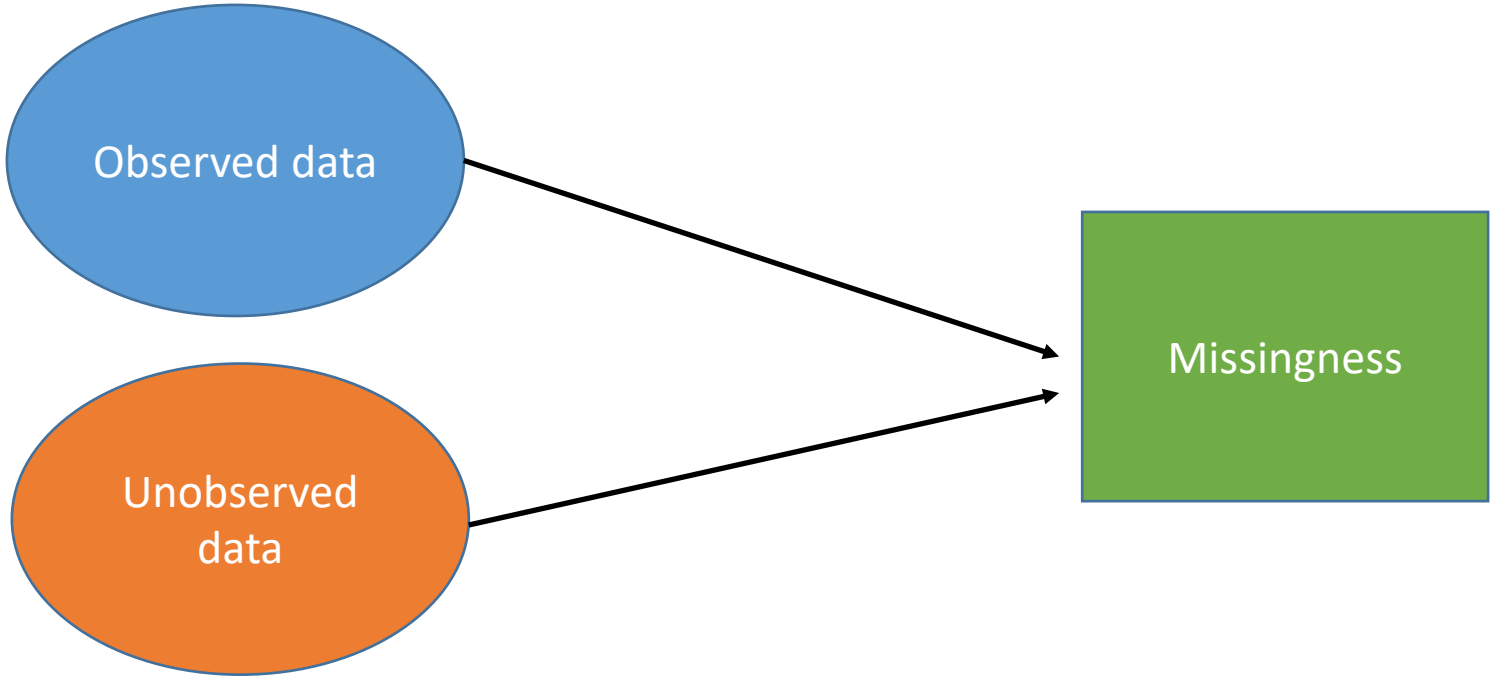
- Missingness has no relationship with observed or unobserved data

Missing at random (MAR)



- Missingness has a relationship with observed data, but *not* with unobserved data

Missing not at random (MNAR)



- Missingness has a relationship with observed data and with unobserved data

How to know which mechanism?

- MCAR
  - Unlikely in practice, don’t assume
  - Tests exist. Don’t rely on them.
- MAR
  - Fair assumption most of the time
  - Can create missing/present indicator and check for associations. Logistic regression with missingness (Y/N) outcome. Don’t rely on this either.
- MNAR
  - Consult with collaborators and use knowledge of data and data collection
  - Go out and try to get data from non-respondents. If different among non-respondents, then it’s potentially MNAR.

Mechanism

- Never know for sure if missingness is MAR or MNAR
- Need substantive knowledge of what might have led to the missing values
- **Ignorability** – the missing data mechanism is said to be ignorable if
  - The data are MAR, and
  - The parameters that govern the missing data process are not related to the parameters to be estimated
- This implies that the distribution of the data is the *same* for the response and non-response groups

Consequences of missing data

- Quality of the collected data is affected
  - Less observations lead to decreases in reliability and validity
- Ability to make inferences is affected
  - Differences in those with observed data and missing data (selection bias) can threaten validity; obscures relationships
- Ability to generalize is affected
  - Missing data can make data less representative

Consequences of missing data

- Statistical methods are affected
  - Loss of power
  - Affect on distributional assumptions
  - Inaccurate results in hypothesis tests and parameter estimates

- Ignoring the missing data or editing leads to problems
  - Inefficiency – loss of information leading to loss of power,
  - Systematic difference – leading to biased results, and
  - Unreliable results

Visual Approaches to Understanding Missing Data

See slides named: **eda-for-missing-data.html**

Fig. Source: https://jenslaufer.com/data/analysis/visualize_missing_values_with_ggplot.html

Review of Common Approaches

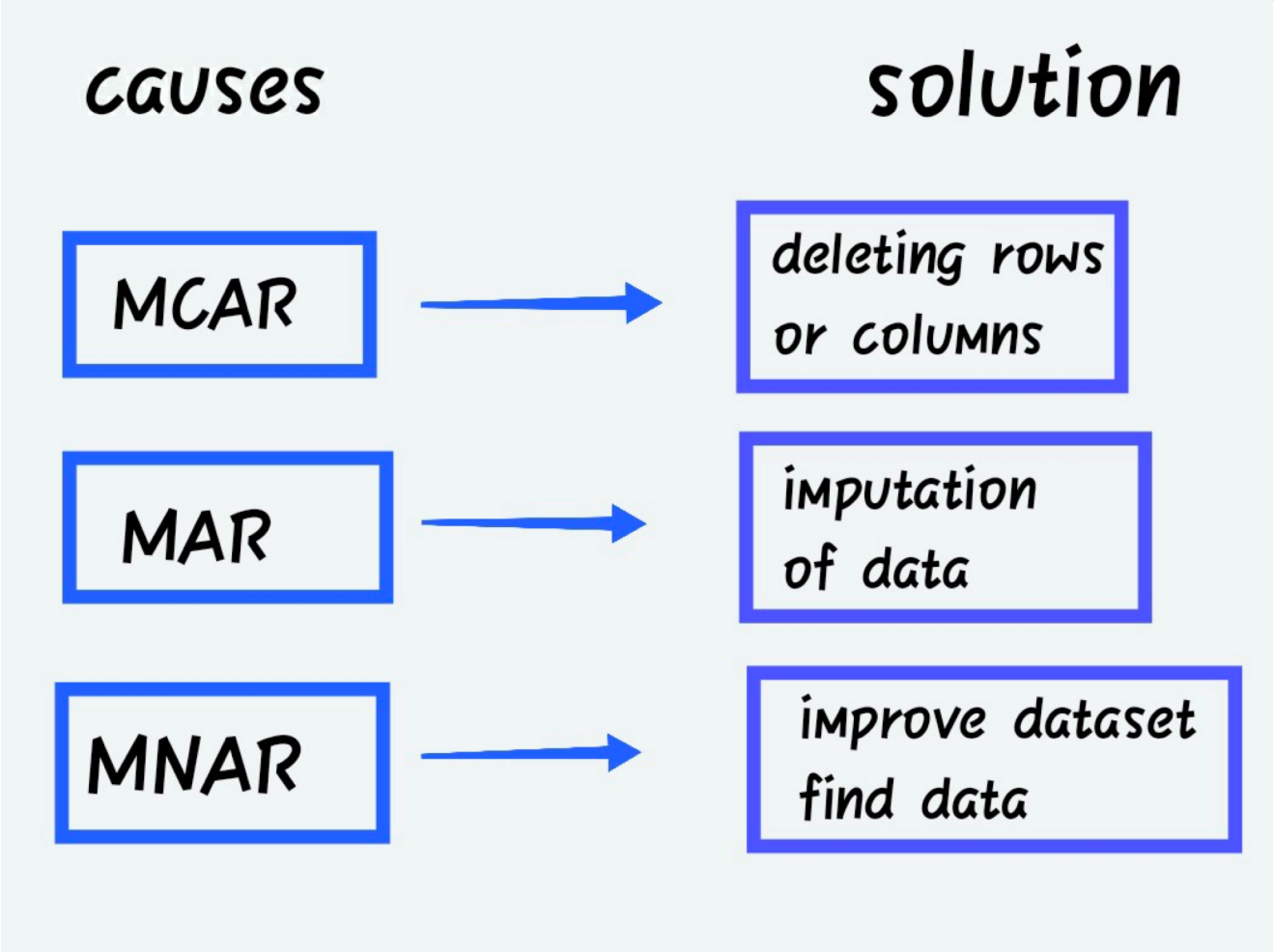


Fig. Source: https://medium.com/@codingpilot25/data-cleaning-types-of-missingness-40655a8b235e

Cholesterol Screening Example

pat_id	age	sex	race_eth	language	fpl	cholesterol_eligibilit	cholesterol_ehr	cholesterol_total
5094		Female	Non-Hispanic, White	English	<=138% FPL	1	1	152.9439156
5711	48	Female	Non-Hispanic, White	English	<=138% FPL	1	0	
1013	26	Female	Non-Hispanic, Black	English	<=138% FPL	1	0	
11608	59	Male	Non-Hispanic, White	English	<=138% FPL	1	1	193.5783138
12443	61	Female	Hispanic	Spanish	<=138% FPL	1	0	
12895	34	Male	Hispanic	English		1	1	226.7803733
8530	54	Male	Non-Hispanic, White	Other	<=138% FPL	1	1	244.2103728
7310	59	Female	Non-Hispanic, White	English	<=138% FPL	1	0	
2637	50	Male	Non-Hispanic, White	English		1	0	

- Research Question: What patient characteristics are associated with receipt of cholesterol screening (as determined by the EHR)?
- Model:

$$logit(p) = \beta_0 + \beta_1 Age + \beta_2 I(sex) + \cdots + \beta_8 I(FPL)$$

Ways to handle missing data

- Common approaches
  - Complete case analysis
  - Single imputation
  - Missing indicator approach

Ways to handle missing data

- Common approaches
  - **Complete case analysis**
  - Single imputation
  - Missing indicator approach

Complete Case Analysis: The Good

pat_id	age	sex	race_eth	language	fpl	cholesterol_eligibilit	cholesterol_ehr	cholesterol_total
5094		Female	Non-Hispanic, White	English	<=138% FPL	1	1	152.9439156
5711	48	Female	Non-Hispanic, White	English	<=138% FPL	1	0	
1013	26	Female	Non-Hispanic, Black	English	<=138% FPL	1	0	
11608	59	Male	Non-Hispanic, White	English	<=138% FPL	1	1	193.5783138
12443	61	Female	Hispanic	Spanish	<=138% FPL	1	0	
12895	34	Male	Hispanic	English		1	1	226.7803733
8530	54	Male	Non-Hispanic, White	Other	<=138% FPL	1	1	244.2103728
7310	59	Female	Non-Hispanic, White	English	<=138% FPL	1	0	
2637	50	Male	Non-Hispanic, White	English		1	0	

- Approach: The procedure eliminates all cases with one or more missing values on the analysis variables.
- Default way of handling missing data in many statistical packages

Complete Case Analysis: Example

```
. tab cholesterol_eligibility
```

cholesterol_eligibility	Freq.	Percent	Cum.
0	304	2.32	2.32
1	12,797	97.68	100.00
Total	13,101	100.00	

```
. keep if cholesterol_eligibility==1
(304 observations deleted)
```

```
. tab cholesterol_ehr
```

cholesterol_ehr	Freq.	Percent	Cum.
0	7,777	60.77	60.77
1	5,020	39.23	100.00
Total	12,797	100.00	

```
. *Run a logistic regression of cholesterol screening on important covariates
. xi: logistic cholesterol_ehr age i.sex i.race_eth i.language i.fpl
i.sex              _Isex_1-2           (_Isex_1 for sex==Female omitted)
i.race_eth         _Irace_eth_1-4      (_Irace_eth_1 for rac~h==Hispanic omitted)
i.language         _Ilanguage_1-3      (_Ilanguage_1 for lan~e==English omitted)
i.fpl              _Ifpl_1-2           (_Ifpl_1 for fpl==<=138% FPL omitted)

Logistic regression                               Number of obs     =      7,742
                                                  LR chi2(8)        =     921.27
                                                  Prob > chi2       =     0.0000
Log likelihood = -4602.0741                       Pseudo R2         =     0.0910
```

cholesterol_ehr	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.057068	.0022685	25.86	0.000	1.052631	1.061523
_Isex_2	1.50517	.0789311	7.80	0.000	1.358152	1.668102
_Irace_eth_2	1.008237	.134977	0.06	0.951	.775548	1.31074
_Irace_eth_3	1.068467	.1571712	0.45	0.653	.8008468	1.425517
_Irace_eth_4	.835929	.0958823	-1.56	0.118	.6676291	1.046655
_Ilanguage_2	1.201161	.1050015	2.10	0.036	1.012026	1.425642
_Ilanguage_3	.9392862	.1479226	-0.40	0.691	.6898382	1.278936
_Ifpl_2	.9659661	.1728814	-0.19	0.847	.6801742	1.37184
_cons	.0563482	.007655	-21.17	0.000	.043176	.0735388

Complete Case Analysis: The Bad

- If the data are not MCAR, can produces biased estimates of means, variances, regression coefficients and correlations.
- Loss of information through loss of study sample.
- King et al. (2001) showed that % of incomplete data in the political sciences >50% on average
  - Some had >90% incomplete records.
- Little and Rubin (2002, 41–44) showed that the bias in the estimated mean increases with the difference between means of the observed and missing cases, and with the proportion of the missing data.

Ways to handle missing data

- Common approaches
  - Complete case analysis
  - **Single imputation**
  - Missing indicator approach

Single Imputation: The Good

pat_id	age	sex	ethnicity	languge	race	fpl	age_f
5094	63	Female	Non-Hispanic, white	English	White	<=138% FPL	51-<64
5711	48	Female	Non-Hispanic, white	English	White	<=138% FPL	35-<50
1013	26	Female	Non-Hispanic, other	English	Black	<=138% FPL	19-<34
11608	39.5	Male	Non-Hispanic, white	English	White	<=138% FPL	19-<34
12443	61	Female	Hispanic	Spanish	White	<=138% FPL	51-<64
12895	34	Male	Hispanic	English	White	<=138% FPL	19-<34
8530	39.5	Female	Non-Hispanic, white	Other	White	<=138% FPL	19-<34
7310	59	Female	Non-Hispanic, white	English	White	<=138% FPL	51-<64
2637	50	Male	Non-Hispanic, white	English	White	<=138% FPL	51-<64

- Approach: Replace the observation with a good estimate
  - Example: Replace continuous variables by the mean. Use the mode for categorical data
  - Other Examples: hot deck imputation, regression imputation, etc.

Single Imputation: Example

```
. *Single Impute Age using its mean

. summarize age
```

Variable	Obs	Mean	Std. Dev.
age	10,852	40.04893	12.64449

```
. replace age=40.05 if age==.
variable age was byte now float
(1,945 real changes made)

. *Single Impute Sex using its mode

. tab sex
```

sex	Freq.	Percent	Cum.
Female	7,554	65.76	65.76
Male	3,933	34.24	100.00
Total	11,487	100.00	

```
. replace sex="Female" if sex==""
(1,310 real changes made)
```

DO THIS FOR ALL VARIABLES!!!

```
. *Run a logistic regression of cholesterol screening on important covariates
. xi: logistic cholesterol_ehr age i.sex i.race_eth i.language i.fpl
i.sex              _Isex_1-2           (_Isex_1 for sex==Female omitted)
i.race_eth         _Irace_eth_1-5      (_Irace_eth_1 for rac~h==Hispanic omitted)
i.language         _Ilanguage_1-3      (_Ilanguage_1 for lan~e==English omitted)
i.fpl              _Ifpl_1-2           (_Ifpl_1 for fpl==<=138% FPL omitted)

Logistic regression                               Number of obs     =     12,797
                                                  LR chi2(9)        =   1236.39
                                                  Prob > chi2       =    0.0000
Log likelihood =  -7952.682                       Pseudo R2         =    0.0721
```

cholesterol_ehr	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.055726	.0018349	31.20	0.000	1.052135	1.059328
_Isex_2	1.254353	.0513613	5.53	0.000	1.15762	1.359169
_Irace_eth_2	1.119751	.1260718	1.00	0.315	.8980186	1.396232
_Irace_eth_3	1.161231	.1424402	1.22	0.223	.9130776	1.476827
_Irace_eth_4	1.020481	.1012456	0.20	0.838	.8401444	1.239526
_Irace_eth_5	1.069938	.1441285	0.50	0.616	.821666	1.393226
_Ilanguage_2	1.39974	.0896627	5.25	0.000	1.234589	1.586984
_Ilanguage_3	1.449197	.186737	2.88	0.004	1.125759	1.865561
_Ifpl_2	.9989946	.1482087	-0.01	0.995	.7469314	1.33612
_cons	.0596171	.0069443	-24.21	0.000	.0474484	.0749066

Single Imputation: The Bad

- Estimates could be biased if the data are not MCAR
- Underestimates the standard errors
  - Thus, creates narrow confidence intervals
- Potential to distort correlations among variables

Ways to handle missing data

- Common approaches
  - Complete case analysis
  - Single imputation
  - **Missing indicator approach**

Missing Indicator: The Good

pat_id	age	sex
5094		Female
5711	48	Female
1013	26	Female
11608	59	Male
12443	61	Female
12895	34	Male
8530	54	Male
7310	59	Female
2637	50	Male
8913	40	Male
3205		Male
10722	61	

pat_id	age	age_missing	sex	sex_missing
5094	0	1	Female	Female
5711	48	0	Female	Female
1013	26	0	Female	Female
11608	59	0	Male	Male
12443	61	0	Female	Female
12895	34	0	Male	Male
8530	54	0	Male	Male
7310	59	0	Female	Female
2637	50	0	Male	Male
8913	40	0	Male	Male
3205	0	1	Male	Male
10722	61	0		missing

- Approach:
  - For continuous variables, replaces each missing value by a zero and adds a new dummy variable to indicate presence/absence of missing data.
  - For categorical, includes a missing category and will also include indicator

Missing Indicator: Example

```
. *Create missing age indicator
. generate age_missing=1

. replace age_missing=0 if age==.
(1,945 real changes made)

.
. *Set all missing ages to 0
. replace age=0 if age==.
(1,945 real changes made)

.
. *Create a new sex variable with missing categories
. generate sex_missing=sex
(1,310 missing values generated)

. replace sex_missing="missing" if sex==""
variable sex_missing was str6 now str7
(1,310 real changes made)
```

```
. *Run a logistic regression of cholesterol screening on only age and sex
. xi: logistic cholesterol_ehr age i.age_missing i.sex_missing
i.age_missing     _Iage_missi_0-1     (naturally coded; _Iage_missi_0 omitted)
i.sex_missing     _Isex_missi_1-3     (_Isex_missi_1 for sex~g==Female omitted)

Logistic regression                               Number of obs     =     12,797
                                                  LR chi2(4)        =    1589.05
                                                  Prob > chi2       =     0.0000
Log likelihood = -7776.3528                       Pseudo R2         =     0.0927
```

cholesterol_ehr	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.056503	.0018857	30.80	0.000	1.052814	1.060206
_Iage_missi_1	.0479172	.0044866	-32.45	0.000	.0398833	.0575694
_Isex_missi_2	1.472032	.0627336	9.07	0.000	1.354072	1.600269
_Isex_missi_3	1.623278	.1077169	7.30	0.000	1.425309	1.848743
_cons	1.037483	.0543941	0.70	0.483	.9361675	1.149764

DO THIS FOR ALL VARIABLES!!!

Missing Indicator: The Bad

- It has been shown that the method can yield severely biased regression estimates, even under MCAR and for low amounts of missing data
- Generally fails in observational data
- NOTE: In randomized trials, there is some evidence that this approach could yield unbiased estimates (Groenwold et al. 2012)
  - However, this method does not allow imputation of outcomes… only predictors.

Ways to handle missing data

- Common approaches
  - Complete case analysis
  - Single imputation
  - Missing indicator approach

- Some more appropriate ways
  - Maximum likelihood
  - Obtaining alternative sources of information
  - Weighting
  - **Multiple imputation**

Multiple imputation

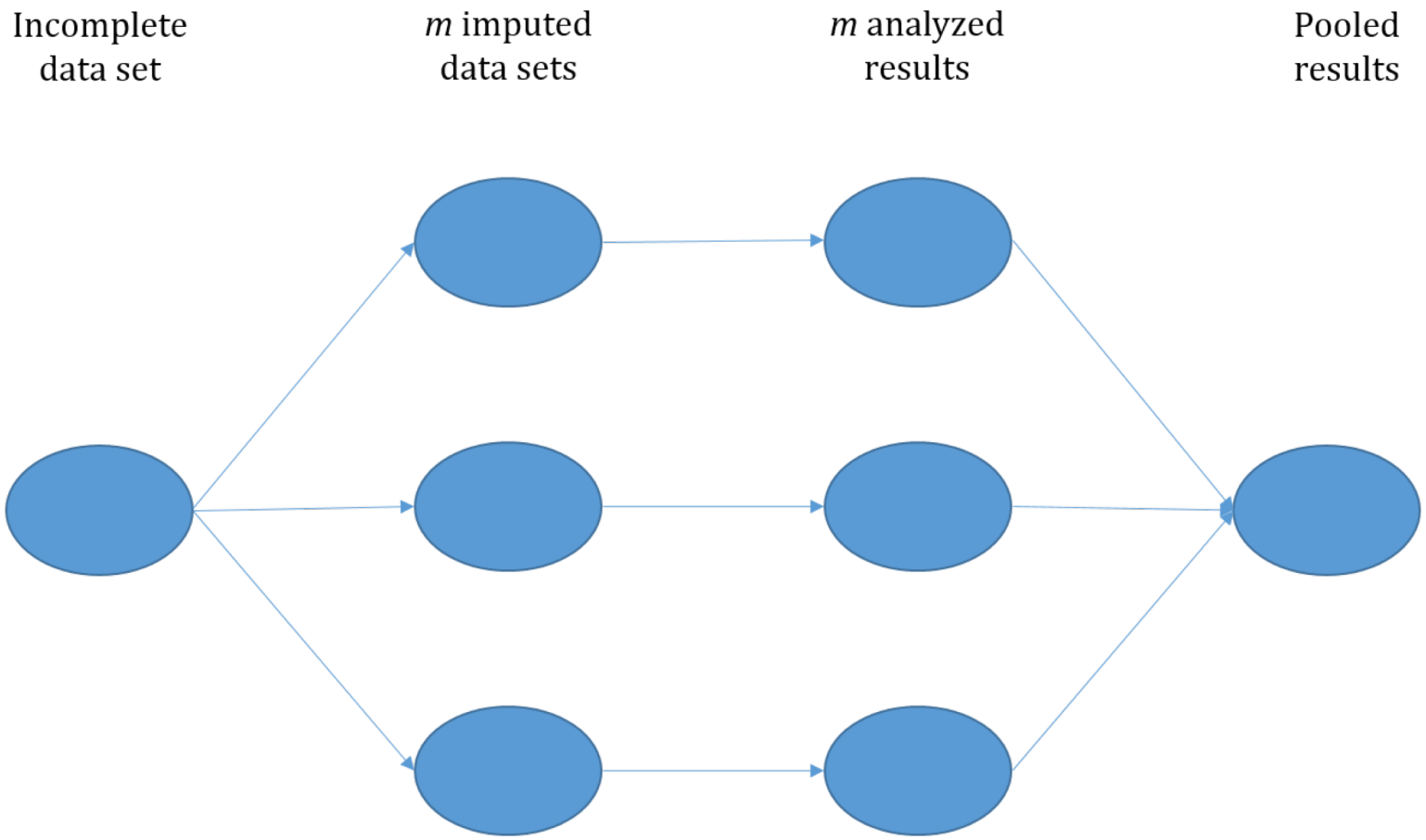
The goal of multiple imputation is to obtain statistically valid inferences from incomplete data.

-Van Buuren (2012)

Multiple imputation

- A statistical method for analyzing data sets with missing values
- Basic idea is to substitute a reasonable guess (imputation) for each missing value, multiple times
- Creates multiple “complete” data sets
- Analyze each data set separately and then combine (or pool) the results

Multiple imputation



Adapted from Van Buuren (2012)

Multiple imputation

- Joint model of all variables (traditional approach)
  - Multivariate normal distribution
  - Fit using observed cases
  - Used to predict the missing values
  - Sometimes use multivariate normal with categorical variables

Multivariate imputation by chained equations

- MICE
  - Sometimes called “fully conditional specification” or “sequential regression multiple imputation”
  - Specify multivariate imputation model on a variable-by-variable basis
  - Iteratively fits a model and imputes each variable
  - Model depends on the type of variable (binary, categorical, ordinal, continuous)
  - Raghunathan et al. (2001), van Buuren et al. (2006)

MICE

- Flexible approach to multiple imputation
- Unnecessary to assume that the variables share a common distribution
- Can more easily work with large data sets with complex data structures
- Models more accurately reflect the distribution of each variable

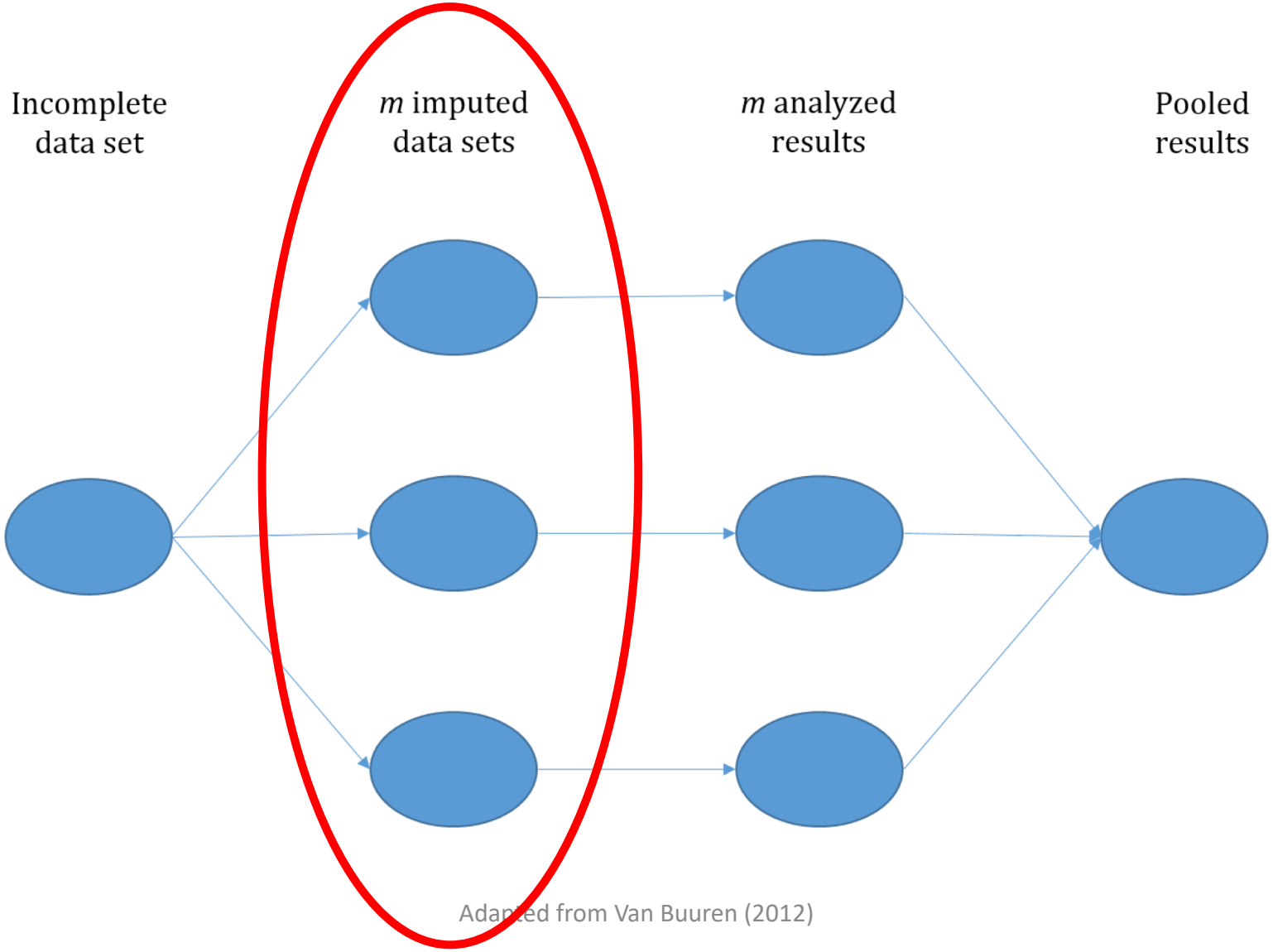
MICE

- Theoretically weaker than joint modelling
- Incompatibility of conditionals
  - No joint distribution exists for the specification of conditional distributions
- In simulation and in practice, the method seems to be robust when the conditions are not met

MICE algorithm

- MICE algorithm developed by Stef Van Buuren
- Implemented in his R software package `mice 3.13.0.`
- Other MICE software packages exist (including Stata and SAS) and vary somewhat in their exact implementation, but the general strategy is the same

Multiple imputation



Adapted from Van Buuren (2012)

MICE algorithm

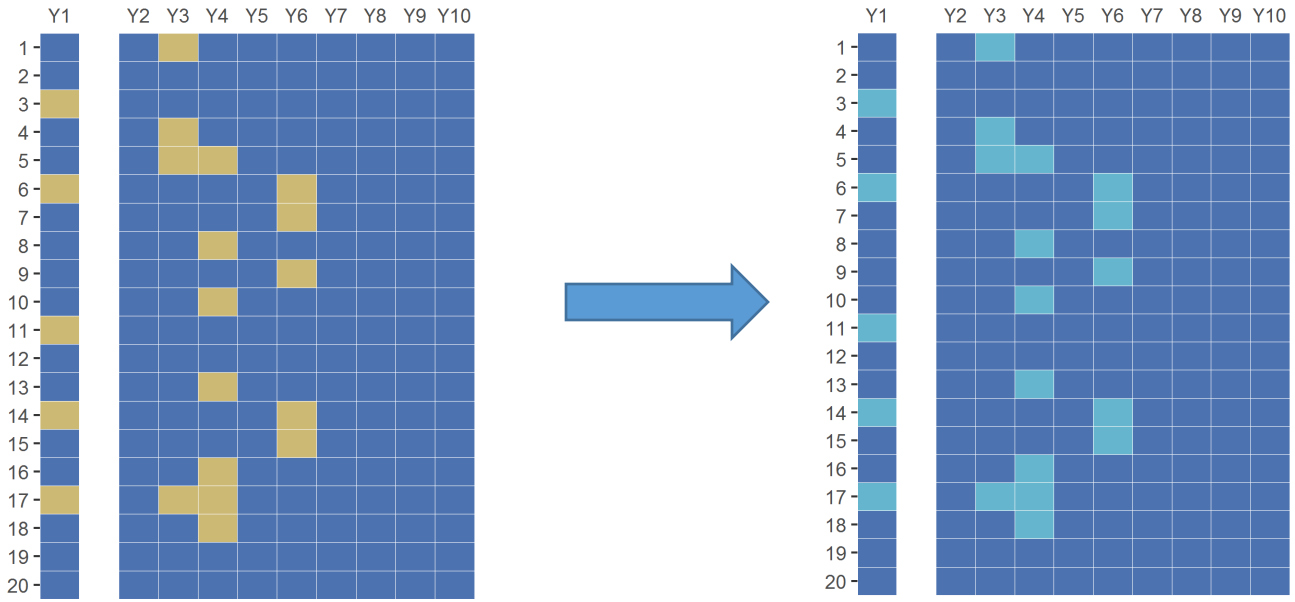
- Step 1
  - The analyst/researcher decides on an imputation model $P\left(Y_j^{mis} \middle| Y_j^{obs}, Y_{\_j}, R\right)$ for each variable Y_j with $j = 1, \ldots, p$.

  - Default choices for some variables type in `mice`

Method	Variable type
Predictive mean matching	Numeric
Logistic regression	Binary
Multinomial logit model	Nominal
Ordered logit model	Ordinal

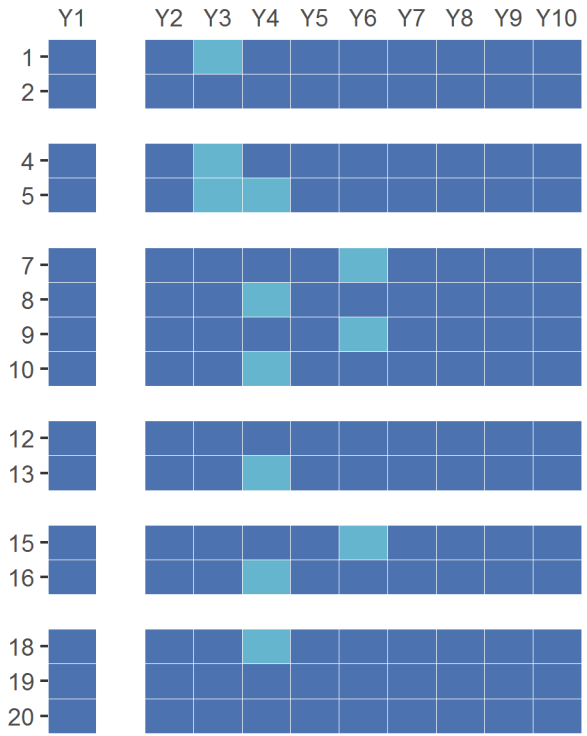
MICE algorithm

- Step 2
  - For each j, missing values are filled in with starting imputations Y_j^0 by a simple imputation using the observed values Y_j^{obs} (e.g. random sampling of observed value with replacement or mean substitution).
  - This initialization is repeated for $t = 1, \ldots, T$ and for $j = 1, \ldots, p$.



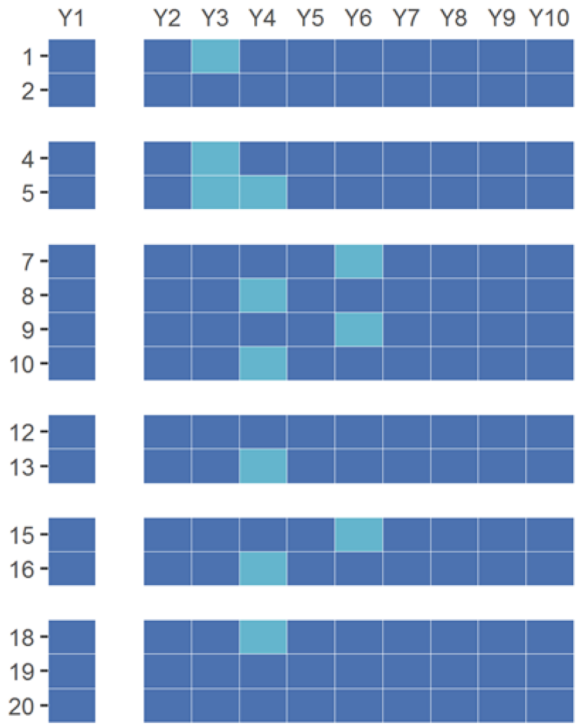
MICE algorithm

- Step 3
  - The values for the variable to be imputed, Y_j, are set back to missing.
  - The currently complete data except Y_j is defined $Y_{-j}^{t} = \left(Y_1^t, \ldots, Y_{j-1}^t, Y_{j+1}^{t-1}, \ldots, Y_p^{t-1}\right)$.



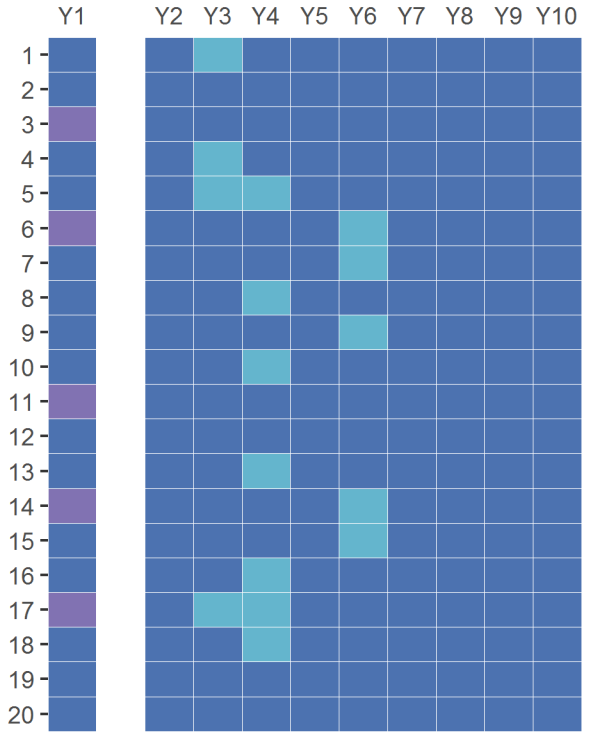
MICE algorithm

- Step 4
  - Draw $\phi_j^t \sim P\left(\phi_j^t \middle| Y_j^{obs}, Y_{-j}^t\right)$ where ϕ_j represents the unknown parameters of the imputation model.
  - The observed values Y_j^{obs} are regressed on the other variables in the imputation model Y_{-j}^t in order to obtain estimates of the regression model parameters ϕ_j^t.



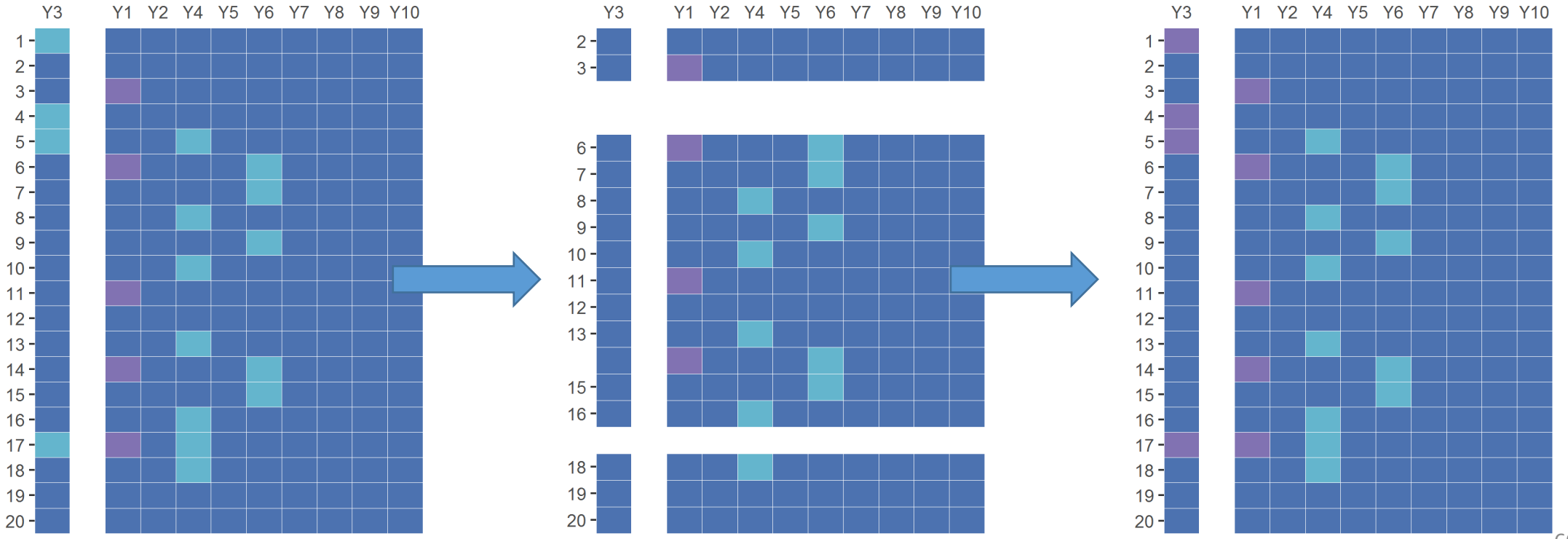
MICE algorithm

- Step 5
  - Draw imputations $Y_j^t \sim P\left(Y_j^{mis} \middle| Y_j^{obs}, Y_{-j}^t, \phi_j^t\right)$, the corresponding posterior predictive distribution of Y_j^t.
  - So missing Y_j^{mis} are replaced with predictions (imputations) from the regression model that was fit in step 4.



MICE algorithm

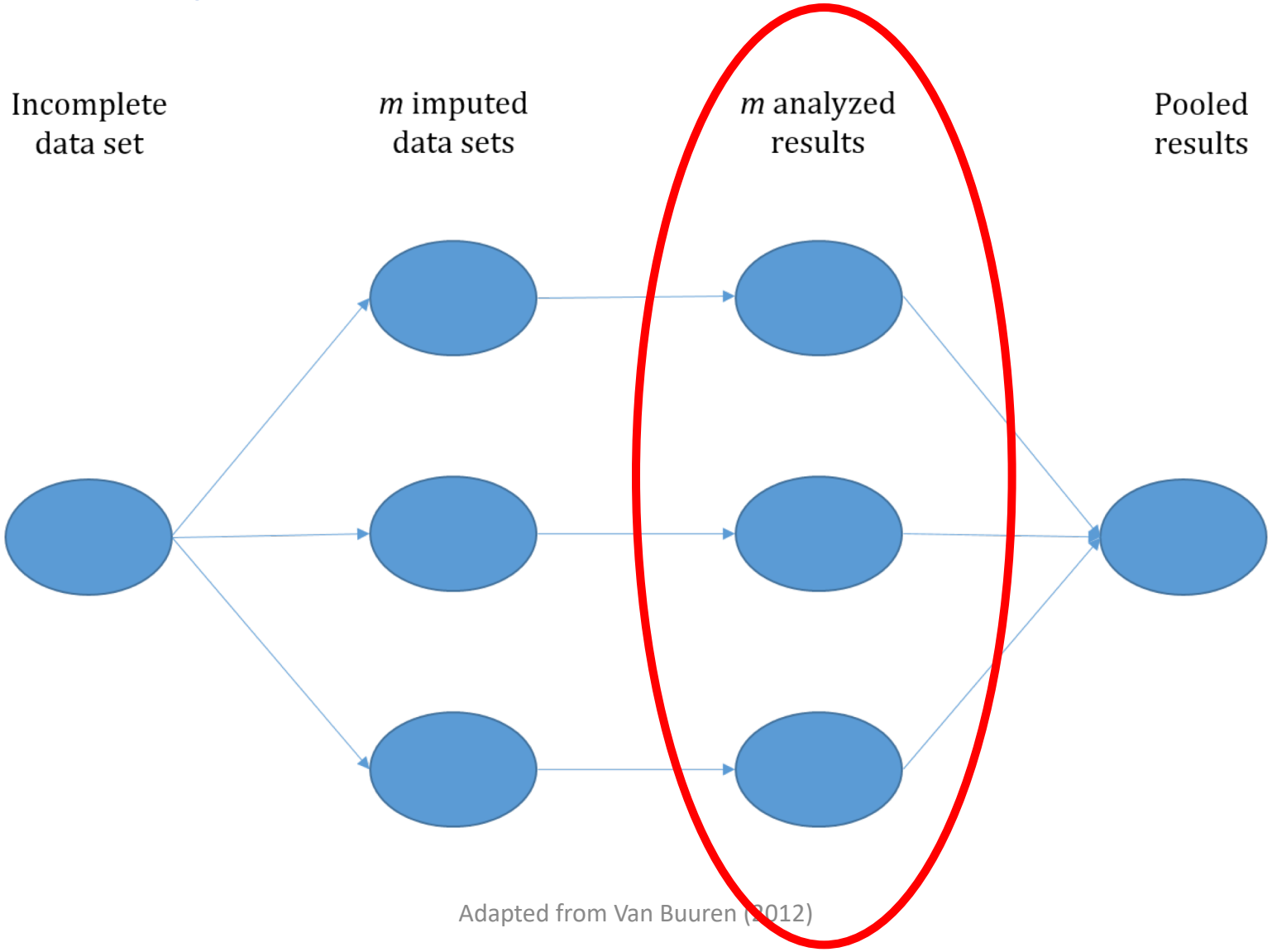
- Step 6
  - End and repeat (3–5) for the next j.
  - End and repeat for the next t.



MICE algorithm

- Steps 1 through 6 are repeated to create m imputed data sets.
  - After one cycle, all of the missing values have been replaced with predictions from regression models that reflect relationships observed in the data
  - The researcher decides how many cycles to perform so that the results have converged or stabilized (generally 10 to 20) which will produce one imputed data set.
  - This whole process is repeated *m* times to produce *m* imputed data sets.

Multiple imputation

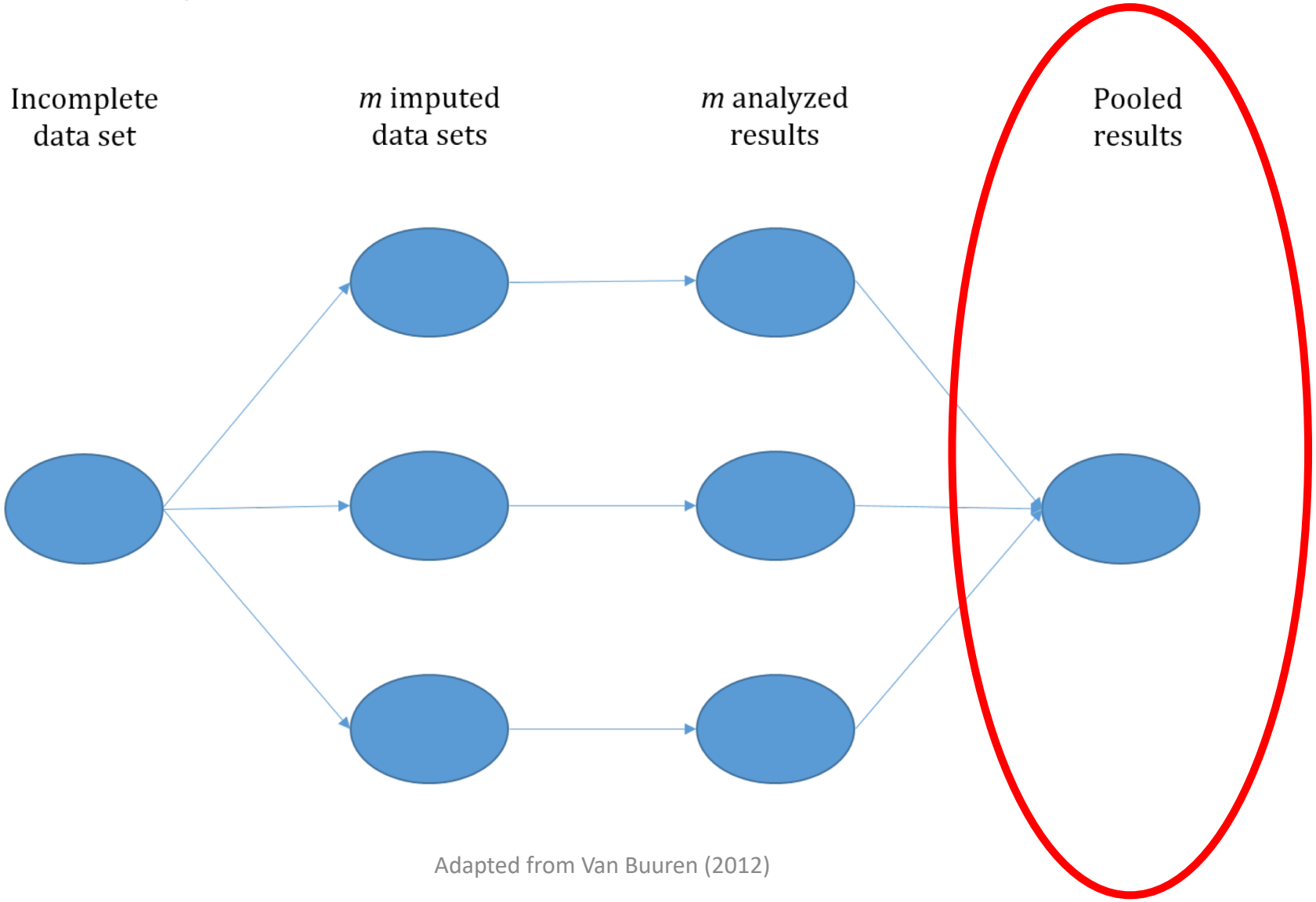


Adapted from Van Buuren (2012)

Analyzing m imputed data sets

- Treat each data set as its own and perform your analysis of interest
  - E.g. Run a logistic regression on each of the m data sets
- Software packages have helpful tools for analyzing the m data sets
- Tend to be designed to neatly handle models
- For certain statistics, it is likely you will have to do it “by hand”

Multiple imputation



Adapted from Van Buuren (2012)

Pooling the results

- Rubin's rules (1987)
  - Provide the method to pool *m* parameter estimates, $\hat{Q}_1, \ldots, \hat{Q}_m$ into a single estimate $\bar{Q}$ and to estimate its variance-covariance matrix
  - Accounts for both within- and between- imputation variance
  - Software is available to help, but designed for models
  - `mice` has some functionality to help "by hand"

Pooling the results

- $\hat{Q}_l$ is the complete-data estimate of the scalar quantity of interest (e.g. regression coefficient, kappa statistic) from the lth imputed data set

- U_l is the complete-data variance of $\hat{Q}_l$

- The overall estimate is the average of the estimates from the m complete data sets

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m} \hat{Q}_l$$

Pooling the results

- The combined within-imputation variance $\bar{U}$ is equal the average of the complete data variances

$$\bar{U} = \frac{1}{m}\sum_{l=1}^{m} U_l$$

- The between-imputation variability reflects the uncertainty due to missing information, i.e. the variance between (among) the *m* complete data estimates

$$B = \frac{1}{m-1}\sum_{l=1}^{m}\left(\hat{Q}_l - \bar{Q}\right)^2$$

Pooling the results

- The total variance of $\bar{Q}$ is given by

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B$$

- An additional term B/m is included to reflect the additional variance since $\bar{Q}$ is estimated for finite m

Pooling the results

- For multi-parameter inference, approaches are available such as Wald Test, likelihood ratio test, and χ^2-test

- For single parameter or scalar inference, like kappa statistics and others examined in this thesis, Wald-type significance tests and confidence intervals can be calculated in the usual way

Pooling the results

- Since the total variance of T is not known, $\bar{Q}$ follows a t-distribution rather than normal

- So univariate tests are based on the approximation

$$\frac{Q - \bar{Q}}{\sqrt{T}} \sim t_v$$

where t_v is Student's t-distribution with v degrees of freedom

- For degrees of freedom see Van Buuren (2012), Schafer (1997), or Barnard and Rubin (1999)

Pooling the results

- The $100(1-\alpha)\%$ confidence interval for Q is calculated as

$$\bar{Q} \pm t_{v,1-\alpha/2}\sqrt{T}$$

Cholesterol Screening Example Revisited

pat_id	age	sex	race_eth	language	fpl	cholesterol_eligibilit	cholesterol_ehr	cholesterol_total
5094		Female	Non-Hispanic, White	English	<=138% FPL	1	1	152.9439156
5711	48	Female	Non-Hispanic, White	English	<=138% FPL	1	0	
1013	26	Female	Non-Hispanic, Black	English	<=138% FPL	1	0	
11608	59	Male	Non-Hispanic, White	English	<=138% FPL	1	1	193.5783138
12443	61	Female	Hispanic	Spanish	<=138% FPL	1	0	
12895	34	Male	Hispanic	English		1	1	226.7803733
8530	54	Male	Non-Hispanic, White	Other	<=138% FPL	1	1	244.2103728
7310	59	Female	Non-Hispanic, White	English	<=138% FPL	1	0	
2637	50	Male	Non-Hispanic, White	English		1	0	

- Research Question: What patient characteristics are associated with receipt of cholesterol screening (as determined by the EHR)?
- Model:

$$logit(p) = \beta_0 + \beta_1 Age + \beta_2 I(sex) + \cdots + \beta_8 I(FPL)$$

Cholesterol Screening Example Revisited

```
. *Need to convert string variables into numeric
. encode sex, generate(sex_n)

. generate sex_binary=sex_n-1
(1,310 missing values generated)

. encode race_eth, generate(race_eth_n)

. encode fpl, generate(fpl_n)

. generate fpl_binary=fpl_n-1
(2,754 missing values generated)

. encode language, generate(language_n)

.
. *Declare the storage style
. mi set wide

.
. *Register Variables
. mi register imputed age sex_binary race_eth_n fpl_binary

. mi register regular language_n

.
. *Perform m=10 imputations
. mi impute chained (regress) age (logit) sex_binary fpl_binary (mlogit) race_eth_n = language_n, ///
> add (10) rseed(2021)
```

Cholesterol Screening Example Revisited

```
. *Perform multiply-imputed logistic regression
. mi estimate: logistic cholesterol_ehr age i.sex_binary i.race_eth_n i.language_n i.fpl_binary

Multiple-imputation estimates                   Imputations       =         10
Logistic regression                             Number of obs     =     12,797
                                                Average RVI       =     0.0719
                                                Largest FMI       =     0.1690
DF adjustment:   Large sample                   DF:     min       =     334.55
                                                        avg       =  10,566.90
                                                        max       =  41,450.13
Model F test:       Equal FMI                   F(   8,12151.3)   =     115.52
Within VCE type:          OIM                   Prob > F          =     0.0000
```

cholesterol_ehr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0446522	.0017193	25.97	0.000	.0412702	.0480343
1.sex_binary	.3178366	.0424621	7.49	0.000	.2344158	.4012574
race_eth_n						
Non-Hispanic, Black	.1341047	.1098992	1.22	0.222	-.0813426	.3495521
Non-Hispanic, Other	.1680072	.1192138	1.41	0.159	-.0656915	.4017059
Non-Hispanic, White	.0441576	.096146	0.46	0.646	-.1443106	.2326257
language_n						
Other	.3282324	.0643025	5.10	0.000	.2021982	.4542666
Spanish	.3803784	.1265602	3.01	0.003	.132315	.6284417
1.fpl_binary	-.1015944	.1425816	-0.71	0.477	-.3819017	.178713
_cons	-2.48986	.1132523	-21.99	0.000	-2.712004	-2.267716

Multiple Imputation Considerations

- Review the steps to specify the imputation model
- Provide some considerations for variable selection in the imputation models
- Assess convergence of the imputations models
- Assess the multiply imputed data

Specify the imputation model

- Decisions and set up prior to running the MICE algorithm
- Van Buuren describes this as the most challenging step
- The model should
  - Account for the process that created the missing data,
  - Preserve the relations in the data, and
  - Preserve the uncertainty about these relations.
- To help, he provides 7 ordered considerations to take

Specify the imputation model

- Step 1
  - Decide if the missing at random (MAR) assumption is reasonable
- Step 2
  - Decide on the form of the imputation model
    - Software defaults are usually best

Specify the imputation model

- Step 3
  - Decide the set of predictors to include in the imputation model
- Step 4
  - Decide whether to impute variables that are function of other (incomplete) variables.
- Step 5
  - Decide the order in which variables should be imputed

Specify the imputation model

- Step 6
  - Decide the number of iterations
    - 10 to 20 are recommended

- Step 7
  - Decide m, the number of multiply imputed data sets.
    - Rule of thumb from more recent authors
      - The number of imputations should be similar to the percentage of cases that are incomplete (at least 5)
      - For example, if a data set had 26% incomplete cases, then choose m = 30

Variable selection

- The general advice cited is to include as many variables as possible
  - Tends to make MAR assumption more reasonable
  - Reasonable to include all variables for small to medium data sets (20 to 30 variables)
  - For large data sets, the advice is to select a subset of 15 to 25
    - Avoid multicollinearity issues
    - Avoid computational issues

Additional advice for variable selection

- Include variables
  - In the model of scientific interest
  - Related to occurrence of missing data (i.e. related to non-response)
  - Where distributions differ between response and non-response
  - Correlated with the target variable

Additional advice for variable selection

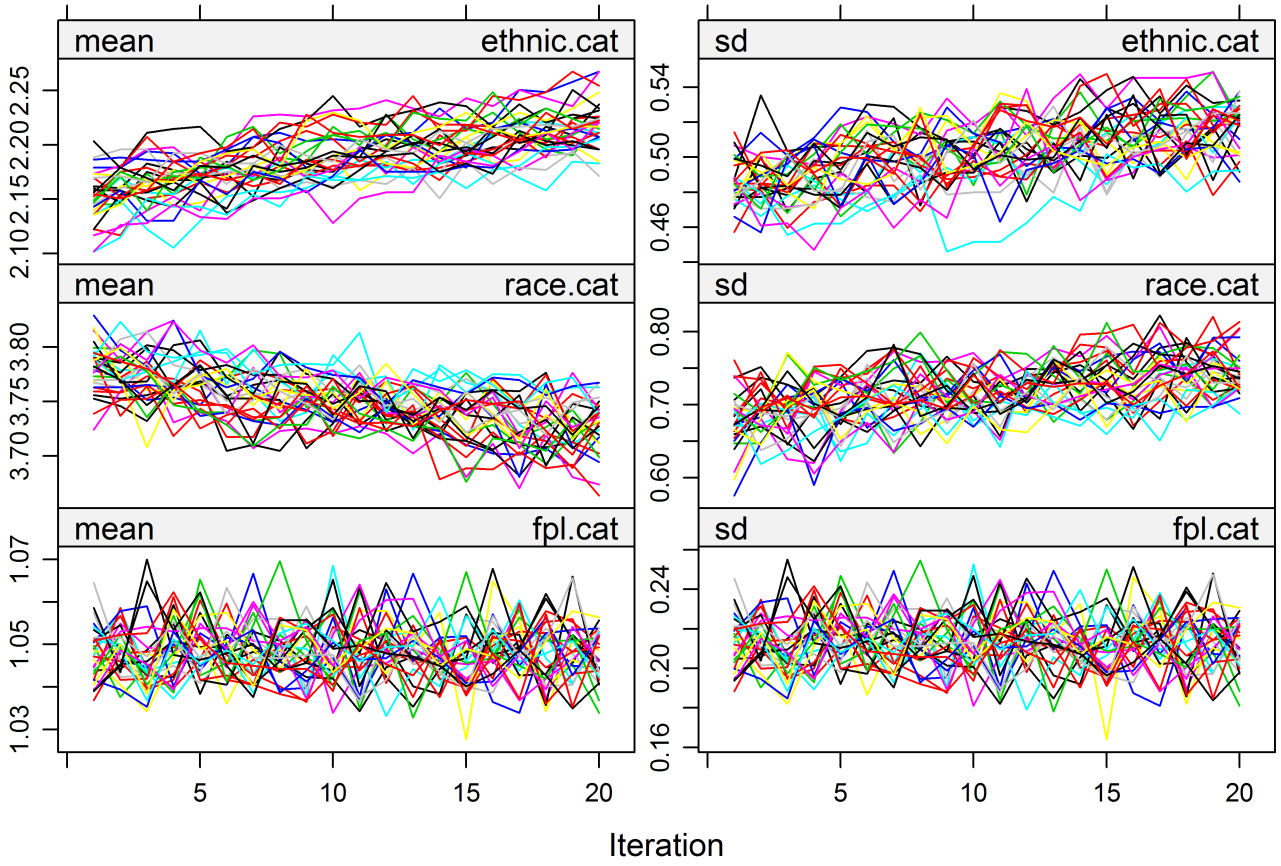
- Remove variables
  - With too many missing values
  - If missing on same cases as target variable

Assessing convergence

- Models must be reviewed for convergence
- Advice is to plot parameters against the iteration number
- `mice` makes this easy
- Streams should be (1) well mixed, and (2) show no signs of trend

Assessing convergence

- Example of
  - Good mixing
  - Possible issue with trend



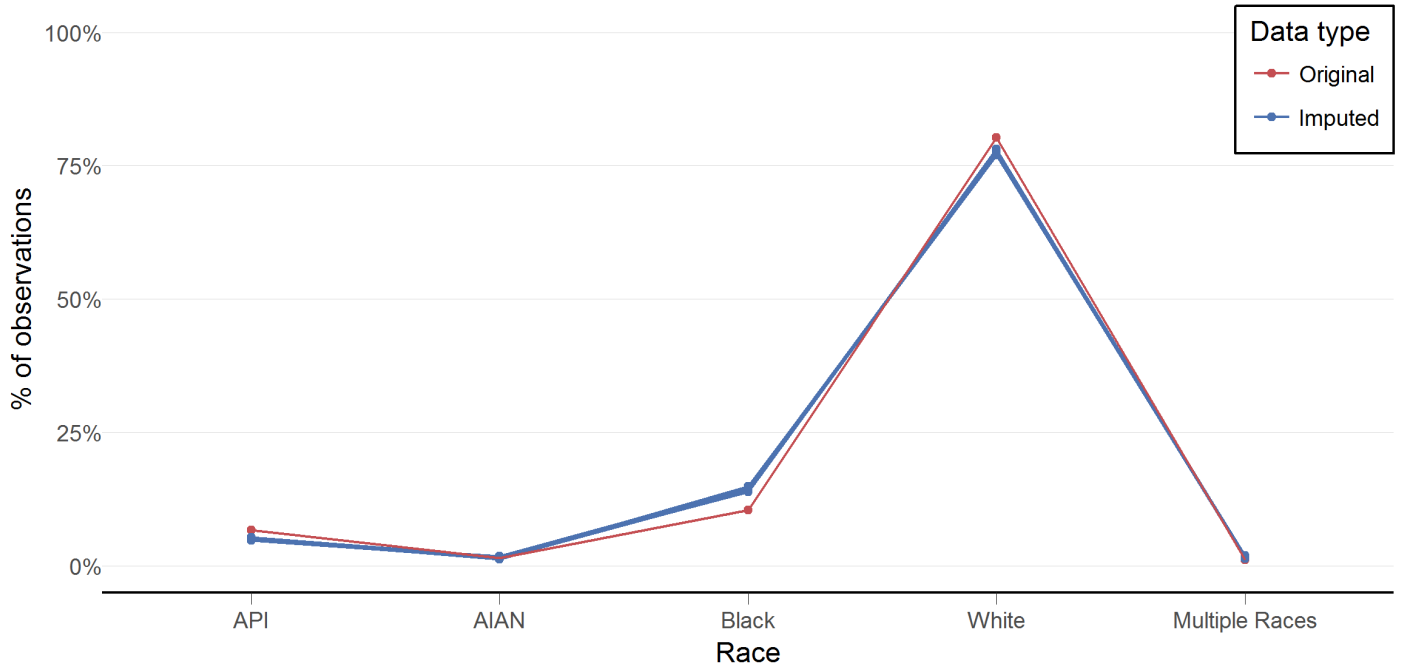
Assessing the imputations

- Numerical diagnostics exist
  - Most examples use continuous variables and regression models.

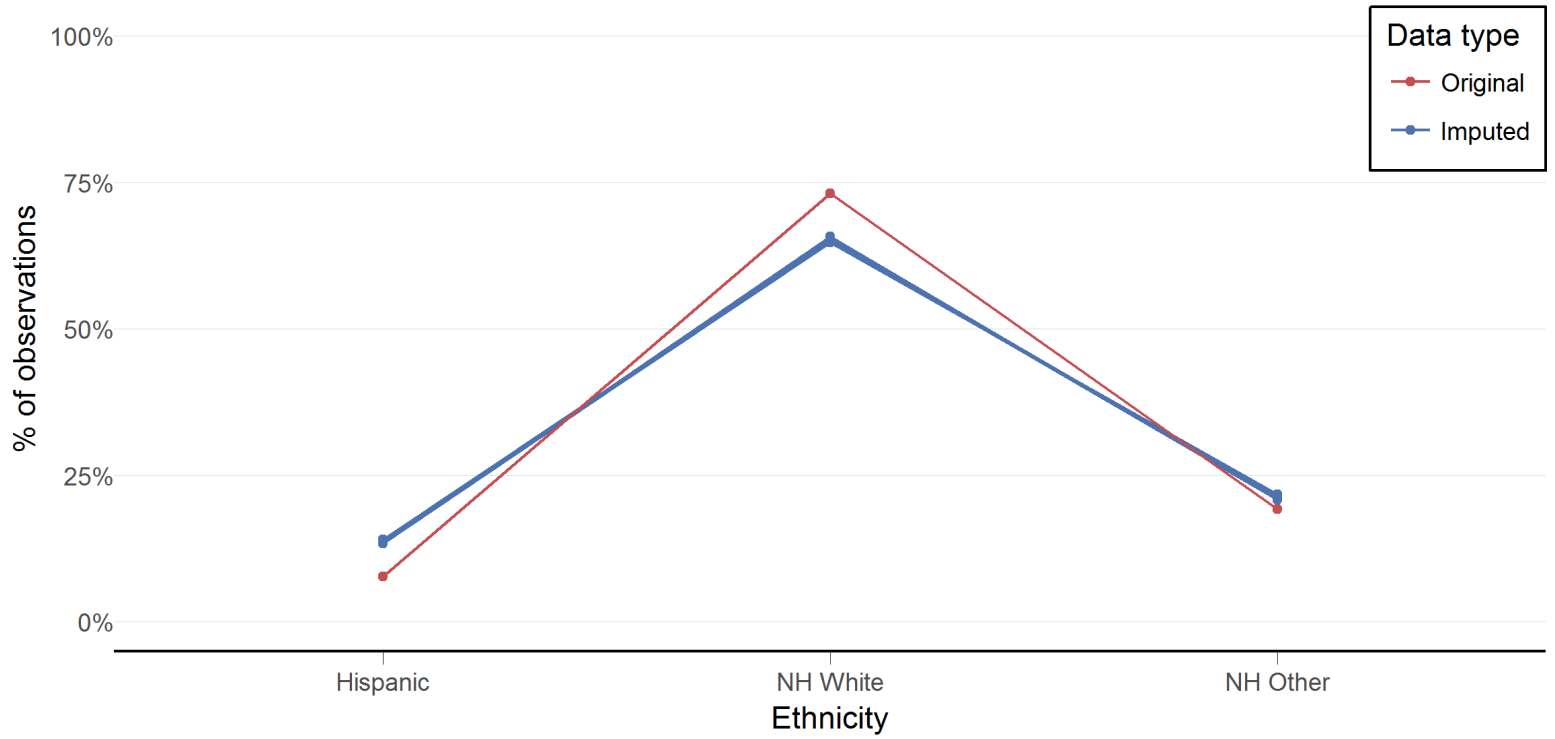
- Data visualization is also recommended to assure that imputed data is
  - Close to the original data,
  - Plausible
    - Within an appropriate range
  - Make common sense
    - e.g. pregnant fathers

Assessing the imputations

- Visually compare original data values to imputed values



Assessing the imputations



Missing data Methods for Uncommon Statistics

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2]\,[n(\sum y^2) - (\sum y)^2]}}$$

$$kappa(\kappa) = \frac{P_o - P_e}{1 - P_e}$$

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

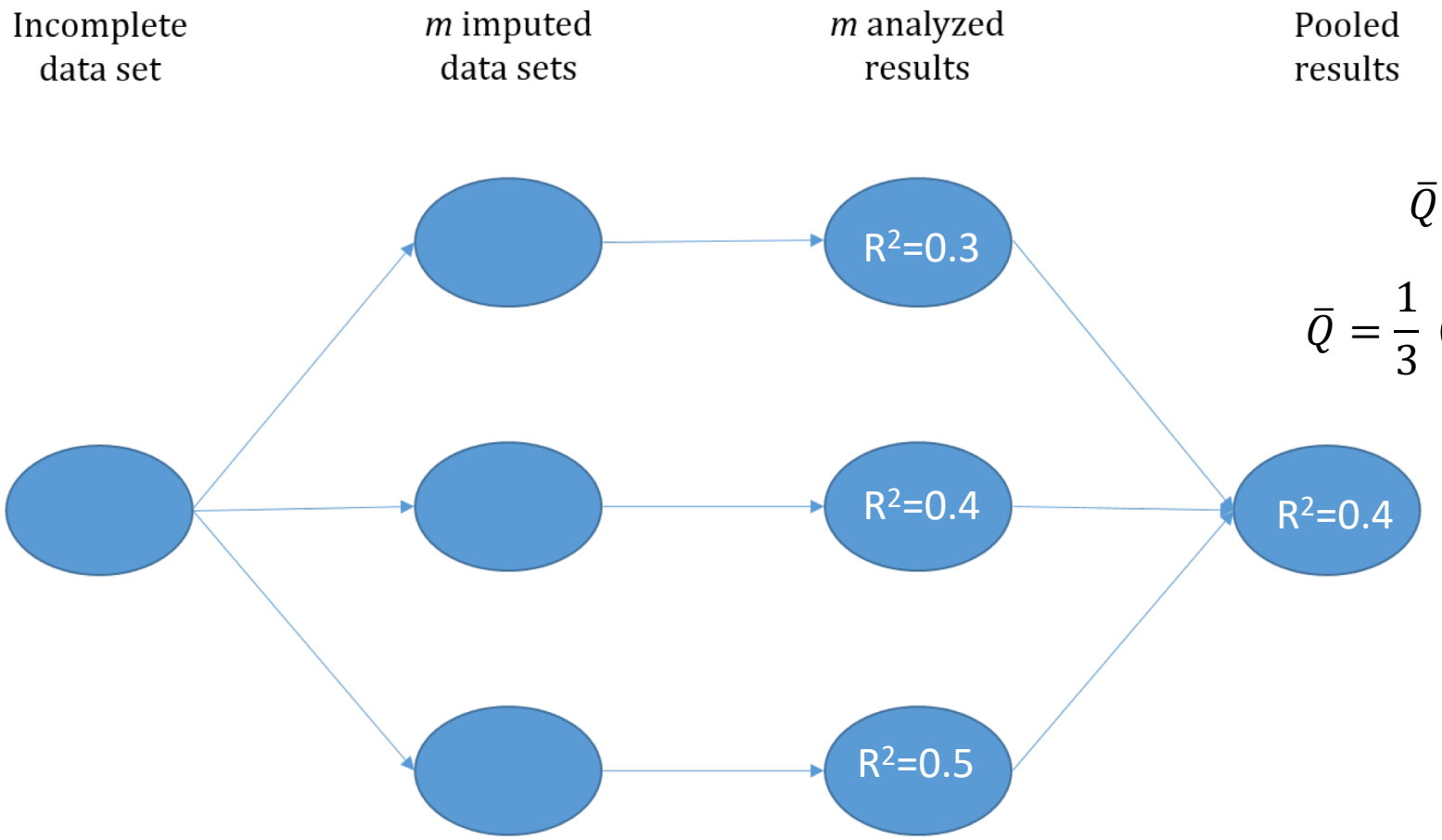
Cholesterol Screening Example Revisited

pat_id	age	sex	race_eth	language	fpl	cholesterol_eligibilit	cholesterol_ehr	cholesterol_total
5094		Female	Non-Hispanic, White	English	<=138% FPL	1	1	152.9439156
5711	48	Female	Non-Hispanic, White	English	<=138% FPL	1	0	
1013	26	Female	Non-Hispanic, Black	English	<=138% FPL	1	0	
11608	59	Male	Non-Hispanic, White	English	<=138% FPL	1	1	193.5783138
12443	61	Female	Hispanic	Spanish	<=138% FPL	1	0	
12895	34	Male	Hispanic	English		1	1	226.7803733
8530	54	Male	Non-Hispanic, White	Other	<=138% FPL	1	1	244.2103728
7310	59	Female	Non-Hispanic, White	English	<=138% FPL	1	0	
2637	50	Male	Non-Hispanic, White	English		1	0	

- Research Question: Among patients with a cholesterol screening, what is the R-squared value for the multiple linear regression of total cholesterol on patient-level covariates?
- Model:

$$E(Total\ chol|X) = \beta_0 + \beta_1 Age + \beta_2 I(sex) + \cdots + \beta_8 I(FPL)$$

Isn’t it this simple?



$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m} Q_l$$

$$\bar{Q} = \frac{1}{3}\,(0.3 + 0.4 + 0.5)$$

$$\bar{Q} = 0.4$$

The Importance of Normality and Transformations

- REMEMBER: Rubin's rules for combining these multiply imputed estimates are based on asymptotic theory.

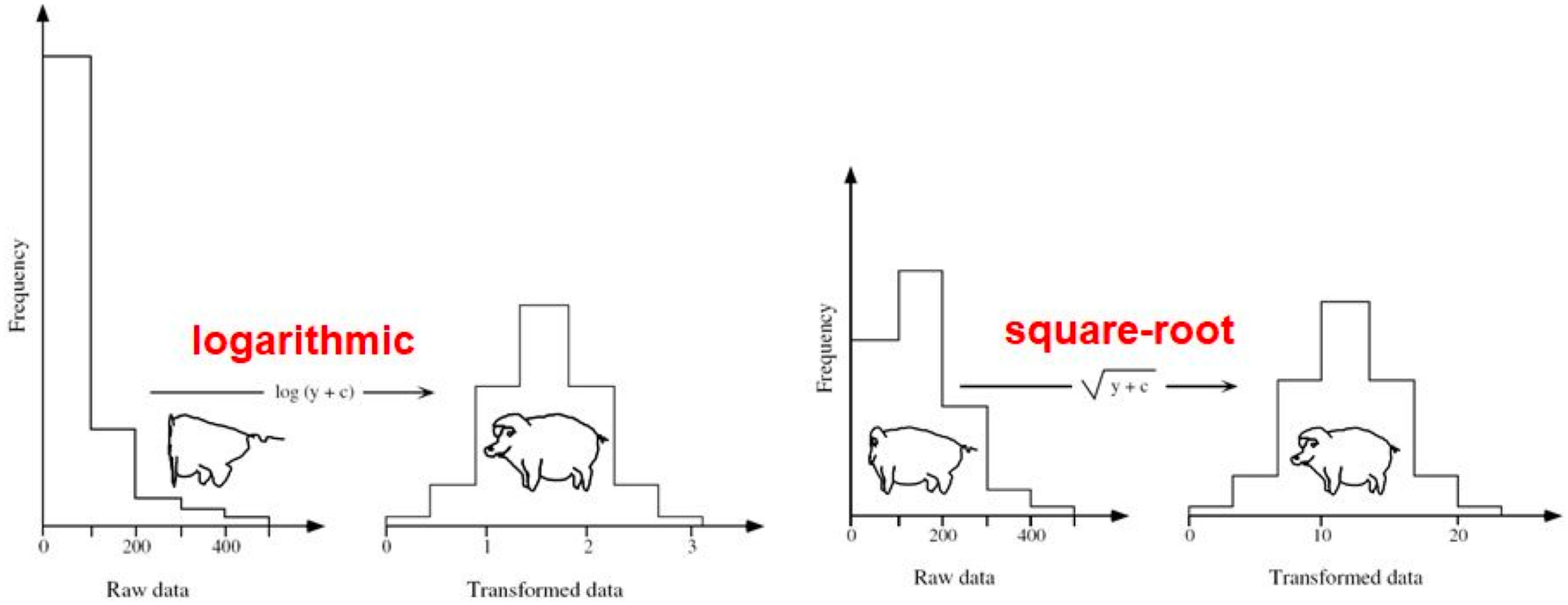


Fig. Source: https://www.davidzeleny.net/anadat-r/doku.php/en:data_preparation

(Un)Common Transformations

Statistic	Transformation	Equation to transform	Equation to back transform
Means, standard deviations, regression coefficients, proportions and linear predictors	None	N/A Use Rubin's Rules as is.	N/A Use Rubin's Rules as is.
Correlation	Fisher *z*	$z = \frac{1}{2}\ln\frac{1+x}{1-x}$	$q = \frac{e^{2z-1}}{e^{2z+1}}$
Odds ratio, relative risk, hazard ratio	Logarithm	$z = \ln(x)$	$q = e^{z}$
Explained variance (R^2)	Fisher z on root	See Harel (2009)	See Harel (2009)
Survival Probabilities	Complementary log-log	See Marshal (2009)	See Marshal (2009)
Survival Distribution	Logarithm	$z = \ln(x)$	$q = e^{z}$

Sources: Harel (2009), Marshall et al. (2009), Agresti 1990

Cholesterol Screening Example Revisited

pat_id	age	sex	race_eth	language	fpl	cholesterol_eligibilit	cholesterol_ehr	cholesterol_total
5094		Female	Non-Hispanic, White	English	<=138% FPL	1	1	152.9439156
5711	48	Female	Non-Hispanic, White	English	<=138% FPL	1	0	
1013	26	Female	Non-Hispanic, Black	English	<=138% FPL	1	0	
11608	59	Male	Non-Hispanic, White	English	<=138% FPL	1	1	193.5783138
12443	61	Female	Hispanic	Spanish	<=138% FPL	1	0	
12895	34	Male	Hispanic	English		1	1	226.7803733
8530	54	Male	Non-Hispanic, White	Other	<=138% FPL	1	1	244.2103728
7310	59	Female	Non-Hispanic, White	English	<=138% FPL	1	0	
2637	50	Male	Non-Hispanic, White	English		1	0	

- Research Question: Among patients with a cholesterol screening, what is the R-squared value for the multiple linear regression of total cholesterol on patient-level covariates?
- Model:

$$E(Total\ chol|X) = \beta_0 + \beta_1 Age + \beta_2 I(sex) + \cdots + \beta_8 I(FPL)$$

Cholesterol Screening Example Revisited

```
**************************
*R-Squared example
**************************

*install mibeta to extraact R-square from MI results (Run code, click on link, and click install)
findit mibeta

*Declare the storage style
mi set wide

*Register Variables
mi register imputed age sex_binary race_eth_n fpl_binary
mi register regular language_n

*Perform m=10 imputations
mi impute chained (regress) age (logit) sex_binary fpl_binary (mlogit) race_eth_n = language_n, ///
add (10) rseed(2021)

*Perform multiply-imputed linear regression (m=10 imputations)
mi estimate: regress cholesterol_total age i.sex_binary i.race_eth_n i.language_n i.fpl_binary

*Estimate multiply-imputed R-square
mibeta cholesterol_total age i.sex_binary i.race_eth_n i.language_n i.fpl_binary
```

Cholesterol Screening Example Revisited

```
. *Estimate multiply-imputed R-square
. mibeta cholesterol_total age i.sex_binary i.race_eth_n i.language_n i.fpl_binary

Standardized coefficients and R-squared
Summary statistics over 10 imputations
```

	mean	min	p25	median	p75	max
age	.1113764	.0984	.1033196	.1146423	.1178938	.125
1.sex_binary	.1375369	.124	.1321125	.139071	.1445235	.147
race_eth_n						
Non-Hispa..	.0483349	.0396	.04213	.0480534	.0548731	.0593
Non-Hispa..	.0536932	.0422	.0465503	.0565708	.0588298	.0606
Non-Hispa..	.0716214	.0581	.0609914	.0724271	.0802926	.0865
language_n						
Other	.0001269	-.00491	-.0005947	.0005713	.0016826	.00292
Spanish	.0098492	.00234	.0056396	.0098354	.0147312	.0177
1.fpl_binary	-.0208468	-.0317	-.0257714	-.019695	-.015078	-.00987
R-square	.036749	.0318	.0354574	.0369935	.0380154	.0422
Adj R-square	.0352112	.0303	.0339175	.0354561	.0364796	.0406

See CSP2021MissingData_stata_code for all the code used in these stata examples

Flu Screening Agreement using Kappa Example

- Compare two data sources
  - EHR
  - Medicaid claims
- Independent single rating for each subject
    - Yes, received screening/procedure
    - No, did not receive screening/procedure

Measuring agreement

- The 2×2 table

		Claims data		
		Yes	**No**	**Total**
EHR data	**Yes**	a	b	m_1
	No	c	d	m_0
	Total	n_1	n_0	n

Measuring agreement

$$\textbf{\textit{Proportion of observed agreement}} = p_O = \frac{a}{n} + \frac{d}{n}$$

		Claims Data		
		Yes	**No**	**Total**
EHR data	**Yes**	**a**	b	m_1
	No	c	**d**	m_0
	Total	n_1	n_0	n

Measuring agreement

- Proportion of observed agreement neglects that some agreement may be due to *chance*

- Kappa statistics were introduced as a measure of “true” agreement
  - Accounts for the agreement *expected due to chance*

Measuring agreement

$$\textit{Proportion of expected agreement} = p_E = \frac{n_1}{n} * \frac{m_1}{n} + \frac{n_0}{n} * \frac{m_0}{n}$$

		Claims data		
		Yes	**No**	**Total**
EHR data	**Yes**	a	b	$\mathbf{m_1}$
	No	c	d	$\mathbf{m_0}$
	Total	$\mathbf{n_1}$	$\mathbf{n_0}$	n

Measuring agreement

$$\hat{\kappa} = \frac{p_O - p_E}{1 - p_E}$$

- The proportion of agreement *after* chance agreement is removed
- Chance adjusted agreement
- How well do EHR and Medicaid data agree beyond pure chance

Flu Screening Agreement using Kappa Example

- See slides named: **mi-in-r-using-mice.html**

Final Thoughts

- How much missing data is too much? (Marino et al. 2021)
  - In general, the lower percentage of missing data, the better. Missing percentages of ≤5% are thought to be trivial
  - One study found that an analysis is likely to be biased if 10% or more of the data are missing
  - Another suggests that if more than 40% of data on important variables are missing, the results should be considered hypothesis-generating rather than confirmatory
  - **RECOMMENDATION: Use of methods such as multiple imputation to reduce bias and improve efficiency at any proportion of missing data**

Final Thoughts

- Real world analysis is complicated and you will encounter missing data.
- Understanding the missing data mechanism (MCAR/MAR/MNAR) is important to guide the analyst in understanding data limitations, and identifying the appropriate analytic approach to employ.
- Methods that addressing missing data are ever evolving. With lots of options, multiple imputation (e.g. using MICE) is a flexible and powerful tool to use in practice.

Resources and References

- King, G., J. Honaker, A. Joseph, and K. Scheve. 2001. “Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation.” *American Political Science Review* 95 (1): 49–69.
- Groenwold RH, White IR, Donders AR, Carpenter JR, Altman DG, Moons KG. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. Cmaj. 2012 Aug 7;184(11):1265-9.
- Little, R. J. A., and D. B. Rubin. 2002. Statistical Analysis with Missing Data. 2nd ed. New York: John Wiley & Sons.

- SAS References: https://support.sas.com/documentation/onlinedoc/stat/141/mi.pdf
- https://support.sas.com/resources/papers/proceedings15/2081-2015.pdf

Resources and References

- Harel, O. 2009. “The Estimation of R^2 and Adjusted R^2 in Incomplete Data Sets Using Multiple Imputation.” Journal of Applied Statistics 36 (10): 1109–18.
- Marshall, A., L. J. Billingham, and S. Bryan. 2009. “Can We Afford to Ignore Missing Data in Cost-Effectiveness Analyses?” European Journal of Health Economics 10 (1): 1–3.
- Agresti, A. 1990. Categorical Data Analysis. New York: John Wiley & Sons.
- Marino, M, Lucas J, Latour E, and Heintzman J. 2021 "Missing data in primary care research: importance, implications and approaches." Family Practice.

Resources and References

- Stata References:
  - https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/
  - https://stats.idre.ucla.edu/stata/faq/how-can-i-estimate-r-squared-for-a-model-estimated-with-multiply-imputed-data/

Resources and References

- Buuren, S.V. (2018). Flexible Imputation of Missing Data (2nd ed.). CRC Press. https://doi.org/10.1201/9780429492259 (available free online)
- McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Methodology in the social sciences. Missing data: A gentle introduction.* Guilford Press.

- R Packages
  - Mice
  - Naniar
  - Vim
  - Finalfit

References

- Buuren, S.V. (2018). Flexible Imputation of Missing Data (2nd ed.). CRC Press. https://doi.org/10.1201/9780429492259 (available free online)
- Stef van Buuren and Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, volume 45, issue 3
- Why You Probably Need More Imputations Than You Think. Paul Allison
- Bodner, Todd. (2008). What Improves with Increased Missing Data Imputations?. Structural Equation Modeling. 15. 651-675. 10.1080/10705510802339072.

Workshop Materials are Available Online

- Workshop materials include:
  - Workshop Slides
  - Practice Data set
  - R code and HTML slides
  - Stata code
- Materials will be found at:
  - ASA CSP workshop website
  - Emile Latour's GITHUB:
    - https://github.com/emilelatour/CSP-2021-missing-data