Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Missing Data: Imputation

Htun Teza

C&B
Data Warehouse

Hierarchy of evidence

- To rank the relative strength of results obtained from medical research

- Randomized controlled trials (RCTs) ensure any variation of outcome is due to the intervention

1. Randomized allocation — by ensuring the case and control groups are comparable at the beginning of the study, they **reduce the risk of bias and confounding characteristics**
2. Sample size — by including sufficient numbers of participants, they **account for random error and chance**
3. Controlled conditions — by conducting in a standardized setting under the researcher's monitor, they **isolate the effects of treatment**

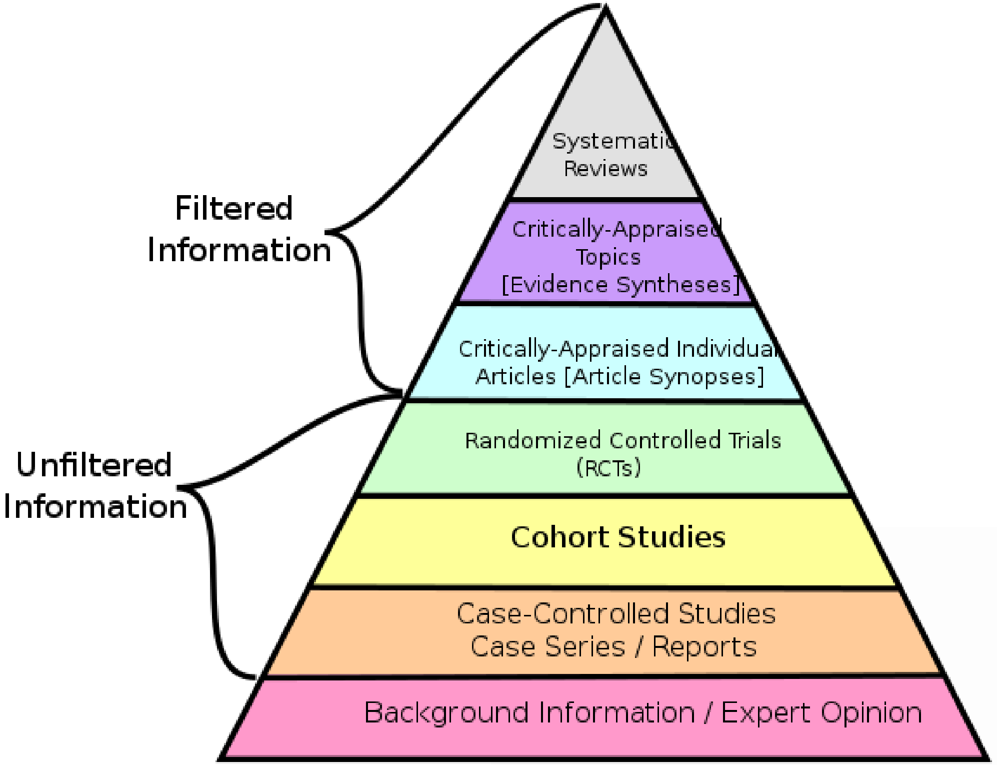

Filtered Information
Unfiltered Information
Systematic Reviews
Critically-Appraised Topics [Evidence Syntheses]
Critically-Appraised Individual Articles [Article Synopses]
Randomized Controlled Trials (RCTs)
Cohort Studies
Case-Controlled Studies
Case Series / Reports
Background Information / Expert Opinion

Diagram - SUNY Downstate EBM Tutorial (2015). Available at: library.downstate.edu. (Retrieved September 3, 2015; Cited September 14, 2023)

Randomized controlled trials

- Gold standard in medical research.
- Focus on interval validity
    - Stringent eligibility criteria leading to homogenous sample
    - Controlled treatment protocol
- Not necessarily external validity
    - General populations is diverse
    - Treatment regimens are complex
    - Observation period is limited

Real world data (RWD)

- Use of data collected in routine clinical environment
- A longitudinal cohort through patient's interaction with the healthcare provider
  - Electronic medical records
    - Physical examination
    - OPD/IPD
    - Laboratory tests
  - Pharmacy dispensing data
  - Health claims data
- Complement clinical trials by generalizing the findings to general population

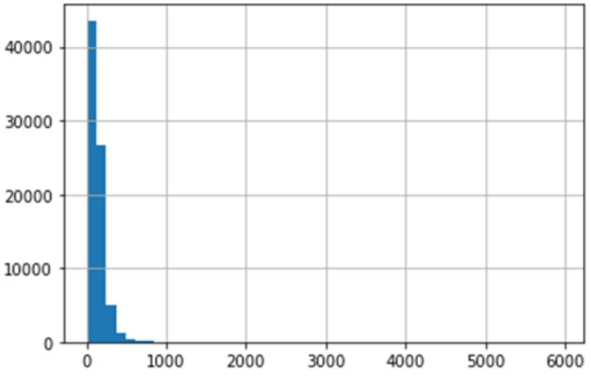
Limitations of RWD

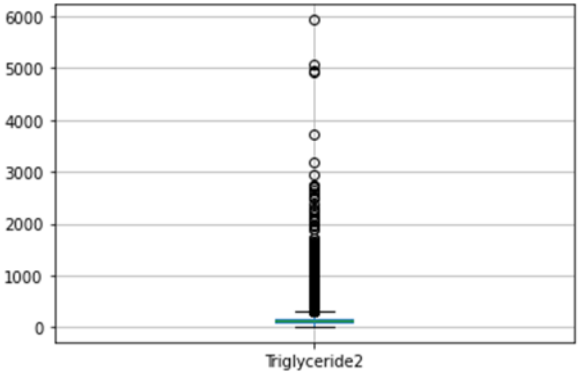
- Data is collected for healthcare process, not research; Missing Data is a prevalent problem.

1. Data entry errors
   - Outlier detection and data truncation as necessary
2. Clinical episodes are over multiple visits.
   - Measurements and physician consultations can be on different days.
3. Not all measurements of interest are made.
   - Interested biomarkers are not measured on every follow-up visit, and different diagnoses require different tests.

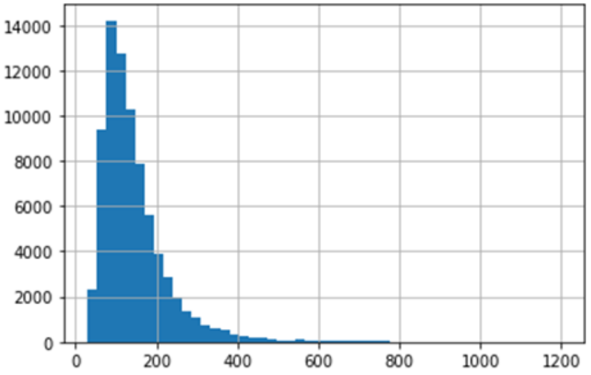
Data Truncation

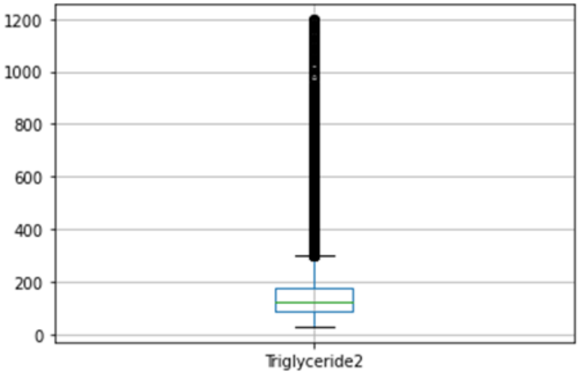
Raw Data	milligram per deciliter
Missing Data	44.6%
Mean (SD)	148.9 (114.9)
Median (IQR)	124 (90 – 175)
Range	15 - 5932





Truncated Data	30-1200 mg/dL
Missing Data	44.7%
Mean (SD)	146.9 (93.5)
Median (IQR)	124 (90 – 175)
Range	30 - 1197





- Since the cohort is from patients, the truncation range should be wider than normal.
- Decisions are made through expert opinion and discussions among the research team.

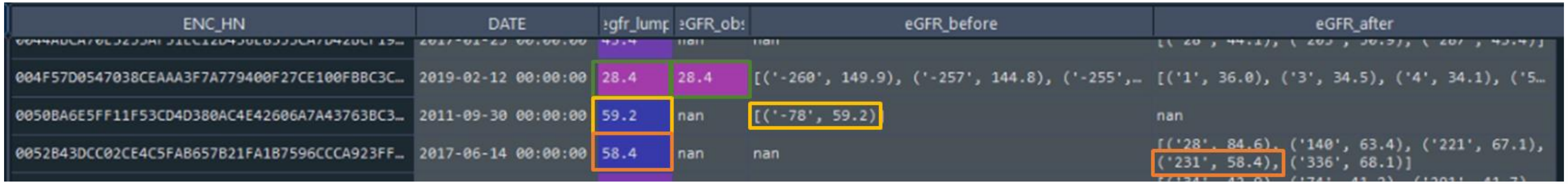
Lumping Data

Data Lumping helps reduce the sparsity of the data.

1. By preserving the record of interest and

2. Aggregating the rest of routine records for a determined time interval

Cross-sectional data



ENC_HN	DATE	:gfr_lump	:GFR_ob:	eGFR_before	eGFR_after
004F57D0547038CEAAA3F7A779400F27CE100FBBC3C…	2019-02-12 00:00:00	28.4	28.4	[('-260', 149.9), ('-257', 144.8), ('-255',…	[('1', 36.0), ('3', 34.5), ('4', 34.1), ('5…
0050BA6E5FF11F53CD4D380AC4E42606A7A43763BC3…	2011-09-30 00:00:00	59.2	nan	[('-78', 59.2)]	nan
0052B43DCC02CE4C5FAB657B21FA1B7596CCCA923FF…	2017-06-14 00:00:00	58.4	nan	nan	[('28', 84.6), ('140', 63.4), ('221', 67.1), ('231', 58.4), ('336', 68.1)]

Longitudinal data

For HT DW project, it reduces 71.2% missing values of 16.1 million observations to

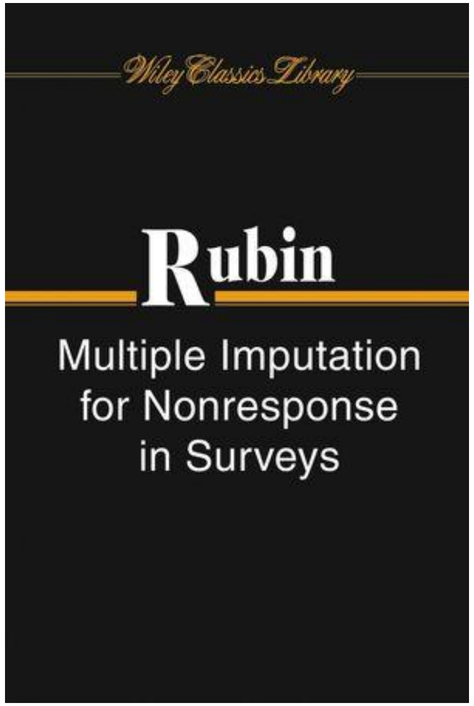
- 54.15% by lumping into 180 days interval (2.8 million observations)

- 49.43% by lumping into 365 days interval (1.6 million observations).

Missing Data

- Multiple Imputation by Chained Equation (MICE)
- Donald B. Rubin (1987)

1. Imputation – m datasets were imputed.
2. Analysis – each of m datasets were analyzed, resulting in m analyses.
3. Pooling – m results were consolidated into one result by specific pooling rules, most commonly Rubin’s rule.



Single Imputation

- For a feature with missing data, an imputation model is specified.
- A regression model is fitted on complete observations,
  - with other features as independent variables or predictors and
  - the feature with missing data as dependent variable or outcome.
- The fitted model is used to predict for the observations with the missing value, called Regression Imputation.
- The set of predictors is selected through expert opinion and discussions among the research team.
- The choice of model depends on the data type of the missing feature.
  - Linear Regression for continuous features
  - Logistic Regression for dichotomous features

Multiple Imputation

- Similarly, multiple imputation models are set so one feature can be both predictor and outcome,
- Features are predicted sequentially from the least to increasing missing percentage, called "Chained Equations".

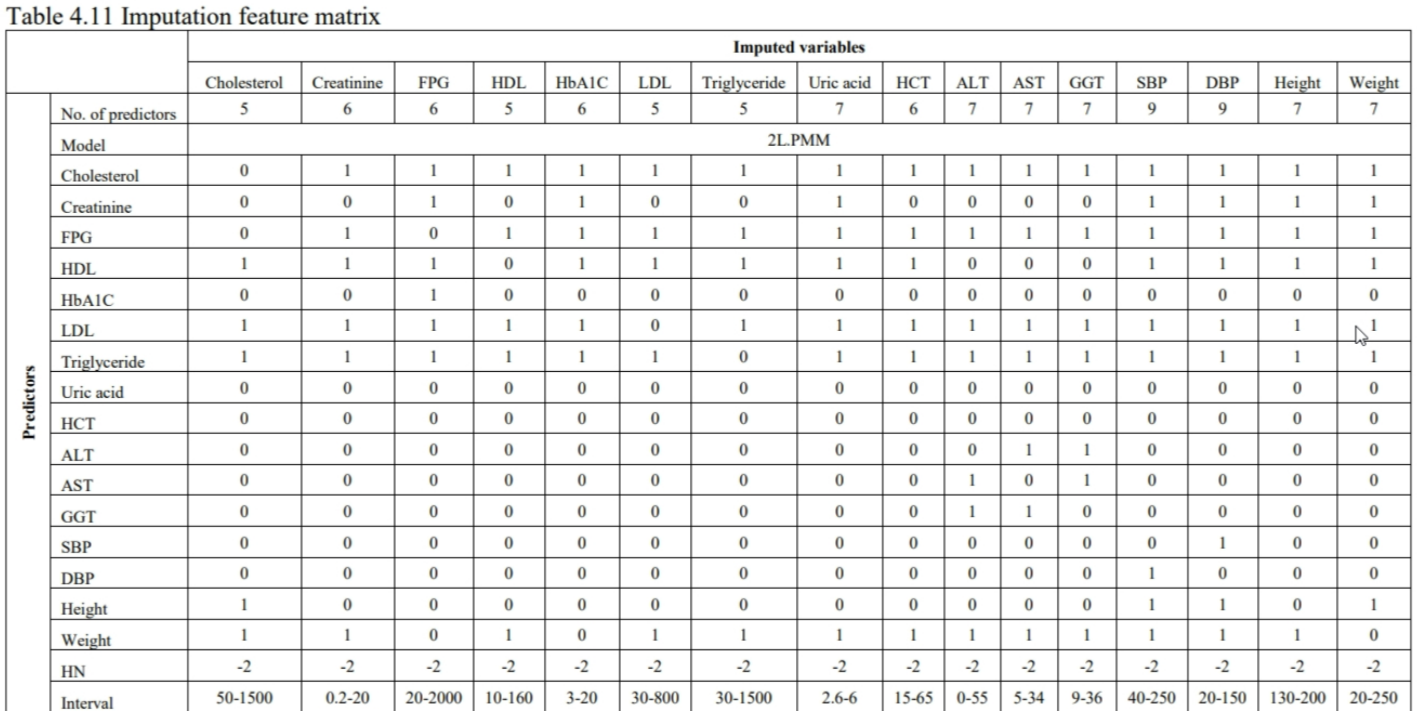


Table 4.11 Imputation feature matrix

		Imputed variables															
		Cholesterol	Creatinine	FPG	HDL	HbA1C	LDL	Triglyceride	Uric acid	HCT	ALT	AST	GGT	SBP	DBP	Height	Weight
Predictors	No. of predictors	5	6	6	5	6	5	5	7	6	7	7	7	9	9	7	7
	Model	2L.PMM															
	Cholesterol	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Creatinine	0	0	1	0	1	0	0	1	0	0	0	0	1	1	1	1
	FPG	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
	HDL	1	1	1	0	1	1	1	1	1	0	0	0	1	1	1	1
	HbA1C	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	LDL	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
	Triglyceride	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
	Uric acid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	HCT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ALT	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
	AST	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
	GGT	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
	SBP	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	DBP	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	Height	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1
	Weight	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	0
	HN	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2
	Interval	50-1500	0.2-20	20-2000	10-160	3-20	30-800	30-1500	2.6-6	15-65	0-55	5-34	9-36	40-250	20-150	130-200	20-250

Looaresuwan P, Boonmanunt S, Siriyotha S, Lukkunaprasit T, et al. Retinopathy prediction in type 2 diabetes: Time-varying Cox proportional hazards and machine learning models. *Informatics in Medicine Unlocked*. (2023) 40(101285). 10.1016/j.imu.2023.101285

Multiple Imputation

- Initial guess
    - All missing values are replaced with simple imputation such as mean, median and mode.

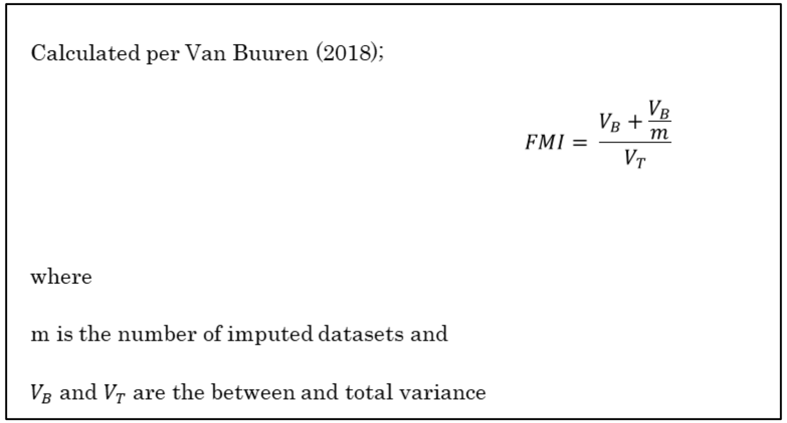
1. Fitting regression model
    - The first regression model is fitted.
2. Estimation
    - The fitted model is used to predict the values for the missing observation.
3. Replacing initial guess
    - The initial guess value for the missing observation is replaced with predicted value.

- The next regression model is continued with predicted value + initial guess.
- When all regression models were fitted, it is considered one iteration.
- This is repeated until predetermined number of iterations are reached.
- One imputed dataset is produced.

Hyperparameters

1. Number of datasets (m)
   - This imputation process is repeated to produce a predetermined number of datasets.
   - This allows the value to be imputed differently, reflecting the fact that there are multiple plausible values for each missing data points.
   - As the rule of thumb, fraction of missing information (FMI) is used.
2. Number of iterations (maxit)
   - The number of times the missing value is updated until one dataset is produced.
   - As the rule of thumb, mean of imputed value for each feature is plotted against the iteration number. The iteration without trending pattern is used.

- Typically, a small preliminary imputation (20 m, 20 maxit) is done to determine the final set of parameters.



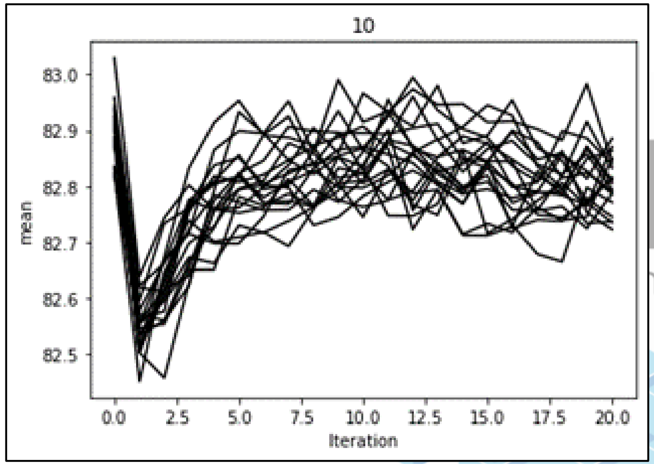
Calculated per Van Buuren (2018);

$$FMI = \frac{V_B + \frac{V_B}{m}}{V_T}$$

where

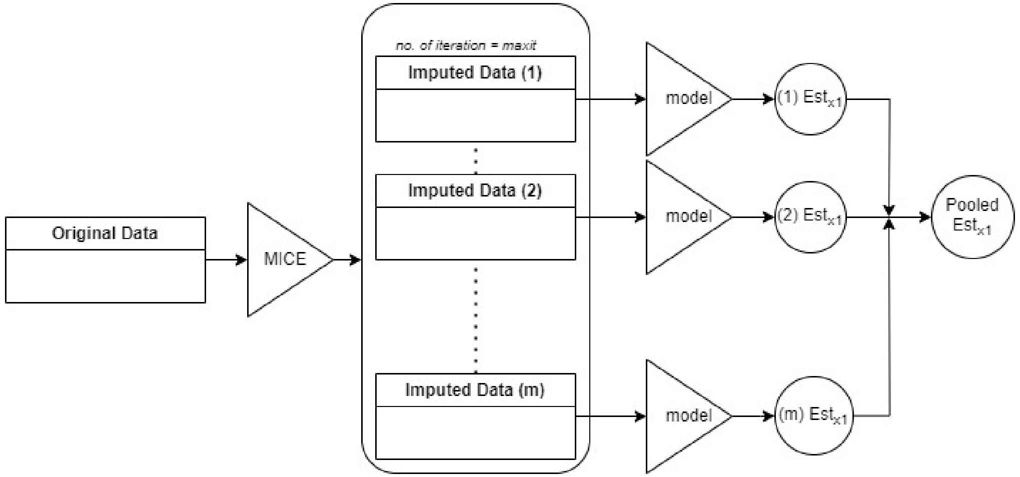
m is the number of imputed datasets and

V_B and V_T are the between and total variance



Analysis and Pooling

- For m dataset, m analyses are to be done, resulting in m estimates.
- The estimates are aggregated to create a pooled estimate.
- Rubin’s rule is commonly used.



$$\theta_{Pooled} = \frac{1}{m}\left(\sum_{i=1}^{m} \theta_i\right)$$

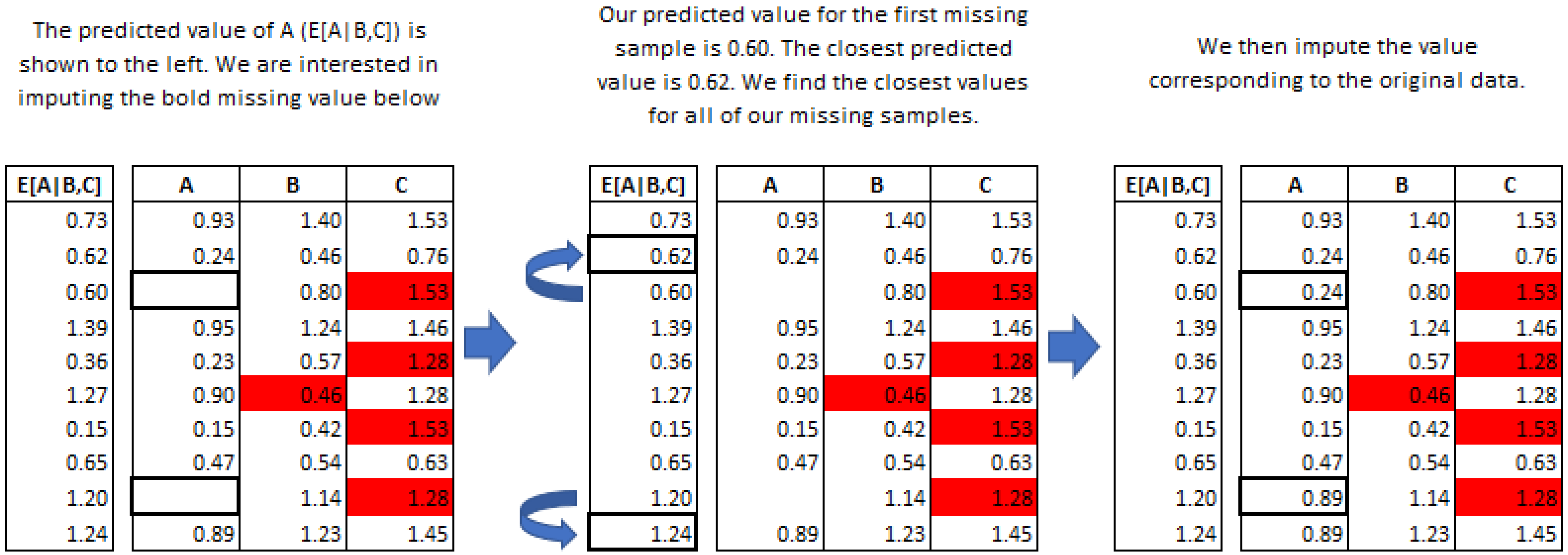
$$SE_{Pooled} = \sqrt{V_{Total}} = \sqrt{V_{within} + (1 + \frac{1}{m})\, V_{Between}}$$

MICE adaptations

- Replacing regression model with
  - Mixed effects models to impute for correlated data
  - Machine learning models such as decision trees, random forests and neural networks
- Regression based imputation has limitations related to regression problems itself.
  - Assumption of linearity
  - Assumption of distribution
  - Model misspecification

Predictive mean matching

- Regression model is used to create a representative estimate for all observations in the dataset.
- The closest estimate with the missing observation is considered the donor.
- The missing value is replaced with the observed value of the donor, or mean observed value of the donors.
- It allows the imputed values to be estimated within the range of observed value.



The predicted value of A (E[A|B,C]) is shown to the left. We are interested in imputing the bold missing value below

E[A|B,C]	A	B	C
0.73	0.93	1.40	1.53
0.62	0.24	0.46	0.76
0.60		0.80	1.53
1.39	0.95	1.24	1.46
0.36	0.23	0.57	1.28
1.27	0.90	0.46	1.28
0.15	0.15	0.42	1.53
0.65	0.47	0.54	0.63
1.20		1.14	1.28
1.24	0.89	1.23	1.45

Our predicted value for the first missing sample is 0.60. The closest predicted value is 0.62. We find the closest values for all of our missing samples.

E[A|B,C]	A	B	C
0.73	0.93	1.40	1.53
0.62	0.24	0.46	0.76
0.60		0.80	1.53
1.39	0.95	1.24	1.46
0.36	0.23	0.57	1.28
1.27	0.90	0.46	1.28
0.15	0.15	0.42	1.53
0.65	0.47	0.54	0.63
1.20		1.14	1.28
1.24	0.89	1.23	1.45

We then impute the value corresponding to the original data.

E[A|B,C]	A	B	C
0.73	0.93	1.40	1.53
0.62	0.24	0.46	0.76
0.60	0.24	0.80	1.53
1.39	0.95	1.24	1.46
0.36	0.23	0.57	1.28
1.27	0.90	0.46	1.28
0.15	0.15	0.42	1.53
0.65	0.47	0.54	0.63
1.20	0.89	1.14	1.28
1.24	0.89	1.23	1.45

Diagram – Wilson SV, Cebere B, Myatt J. miceforest: Fast, Memory Efficient Imputation with LightGBM (2022). Available at: https://github.com/AnotherSamWilson/miceforest. (Retrieved September 18, 2023; Cited September 18, 2023)