UNIVERSITY OF LINCOLN

Missing data: Imputation or deletion?

Missing data?

1. How often do we encounter missing data?

2. How do we deal with it?

Deletion

Excluding from the analysis any cases with data missing on any variables involved in the analysis.

Removing the data from the dataset can lead to a reduction in size and raises concerns for biasing the dataset.

Listwise Deletion

Removes the rows that have missing data → we consider only those rows where we have complete data.

Listwise deletion is the default method for dealing with missing data in most statistical software packages.

Listwise Deletion

When to Use:

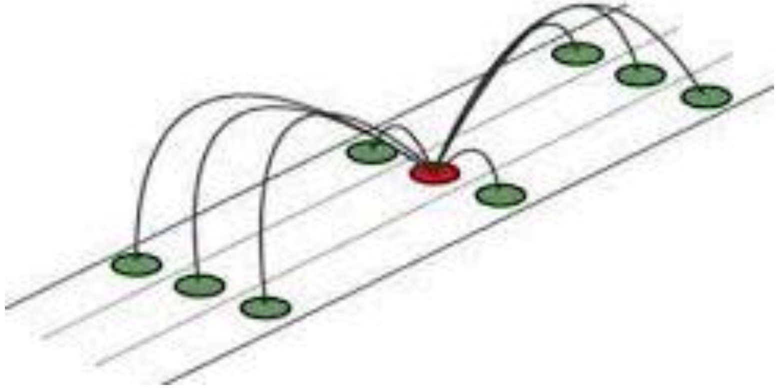
- ❑ Data is MAR(Missing At Random).
- ❑ Good for Mixed, Numerical, and Categorical data.
- ❑ Missing data is not more than 5% – 6% of the dataset.
- ❑ Data doesn't contain much information and will not bias the dataset.

Listwise Deletion: Limitations

- ❑ Deleted data can be informative.
- ❑ The missing data can cause a bias in the dataset and can lead to a faulty analysis by the model.
- ❑ Can lead to the deletion of a large part of the data.
- ❑ A huge amount of missing data can cause distortions in the variable distribution.
- ❑ Can create a bias in the dataset, if a large amount of a particular type of variable is deleted from it.

Imputation

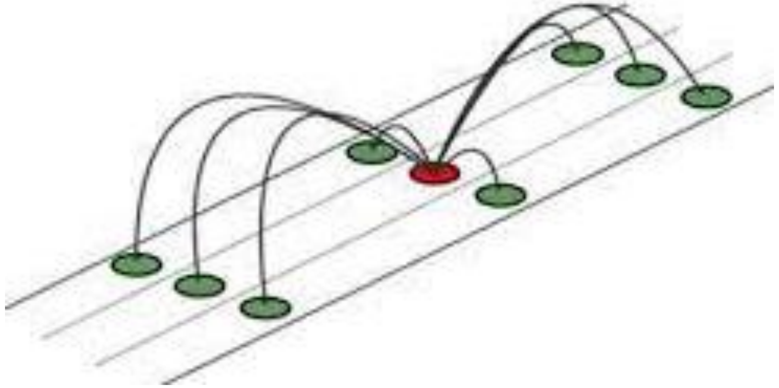
Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset.



Imputation

The imputation method to be used depends on:

- ❑ Type of data: numerical, categorical
- ❑ The analysis
- ❑ The rate of missing data 6-8%, no more than 10%.
- ❑ If the rate is very small (2-3%) , any method could be used.



Arbitrary Value Imputation

Assign missing values a new value (e.g. 99999999, “Missing” or “Not defined”).
Groups missing values into a category on its own.

Assumptions:

- Data is not Missing At Random.
- The missing data is imputed with an arbitrary value that is not part of the dataset or Mean/Median/Mode of data.

Arbitrary Value Imputation

Advantages

- Easy to implement.
- We can use it in production.
- It retains the importance of “missing values” if it exists.

Disadvantages

- Can distort original variable distribution.
- Arbitrary values can create outliers.
- Extra caution required in selecting the Arbitrary value.

Mean imputation

Replace missing values with the mean of the sample.

A simple and appealing method devised

Advantages:

- ❑ The mean is not affected
- ❑ Cases are not lost from the analysis

Disadvantages:

- ❑ The standard error of that variable will be underestimated.
- ❑ The underestimation increases the more missing data there are.
- ❑ Too-small standard errors lead to too-small p-values, so now you're reporting results that should not be there.

Multiple imputation

Aims of Imputation

- ❑ Avoid excluding form the analysis large amount of data
- ❑ Unbiased parameter estimates in the final analysis (regression coefficients, group means, odds ratios, etc.)
- ❑ Accurate standard errors of those parameter estimates thus accurate p-values in the analysis
- ❑ Adequate power to find meaningful parameter values significant

Multiple imputation

Assumptions:
The data are missing at random (MAR)

Advantages:
Resulting estimates (e.g., regression coefficients and standard errors) will be unbiased with no loss of power.

Multiple imputation

- Impute missing values in continuous, binary, ordinal, categorical, count variables.
- Uses univariable and multivariable methods to estimate parameters.
- Depending on the nature of the missing variable, linear, logistic, multinominal logit etc models can be fitted.

Multiple imputation

- Fits the specified model (e.g. multinominal logit model here) on each of the imputation datasets (five) and then combines the results into one MI inference.

- The advice for years has been that 5-10 imputations are adequate.

- Have as many imputations as the percentage of missing data Bodner (2008) .

Investigating conveyance to hospital from care homes

Outcome: Conveyance

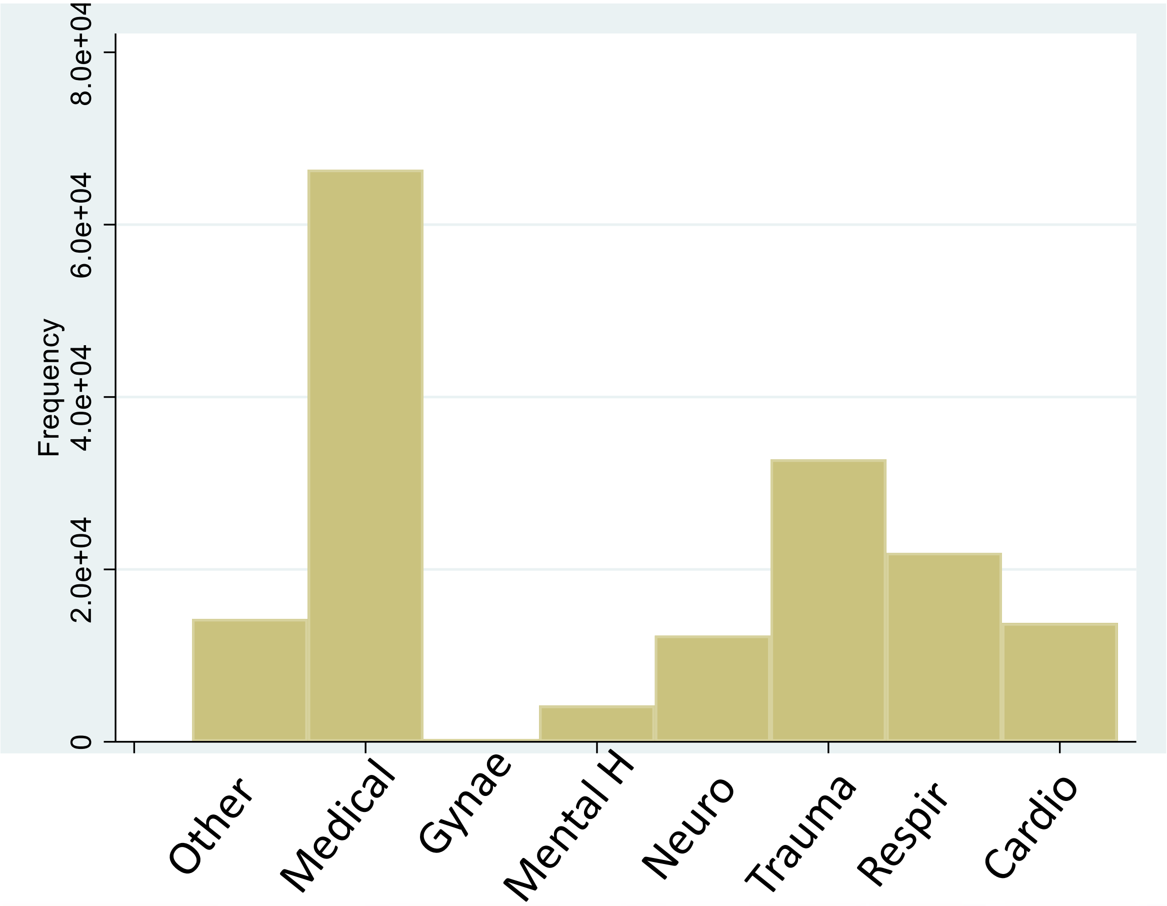
One of the main predictors: Condition Category

- Other: Fall – no injury; No apparent problem
- Medical: Allergies, Sepsis, Abdominal problem
- Gynaecological
- Mental Health
- Neurological
- Trauma
- Respiratory
- Cardiovascular

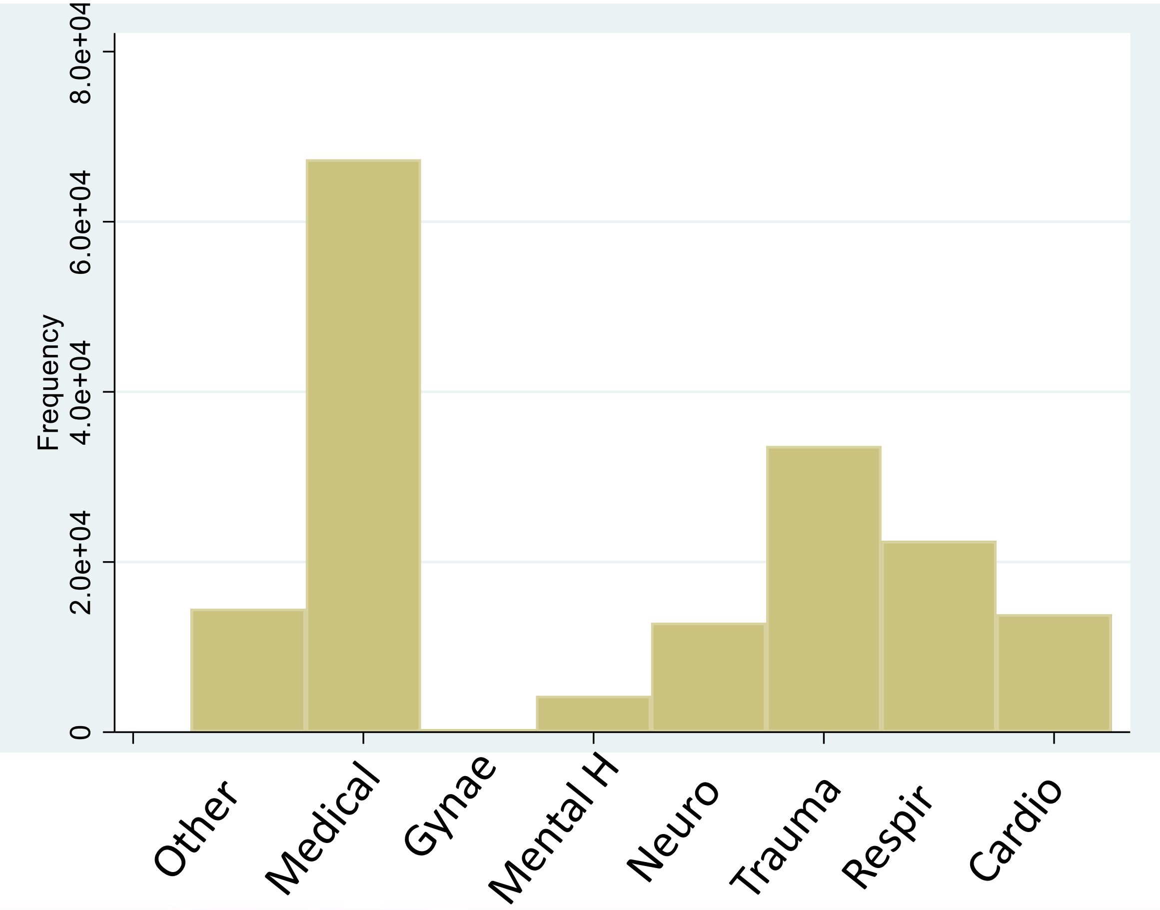
Multiple imputation

- ❑ There were 4,572 (2.74% missing data points for condition category)
- ❑ Multiple imputation using 5 iterations was applied
- ❑ Multinominal logit model was used with the following predictors: age, gender, call category, NEWS score.

Distribution following MI



Distribution before MI



Multiple Imputation (MI)			Listwise Deletion		
Conveyed (Not conveyed)	**RRR**	**95% CI**	**Conveyed (Not Conveyed)**		
Sex (Female)	1	-	**Sex (Female)**	1	-
Male**	1.07	1.03, 1.10	Male**	1.07	1.03, 1.10
Transgender	2.20	0.98, 4.87	Transgender	2.50	1.06, 5.77
Age (under 60)	1	-	**Age (under 60)**	1	-
60-69	1.05	0.96, 1.14	60-69	1.05	0.96, 1.14
70-79*	1.09	1.03, 1.17	70-79*	1.10	1.02, 1.18
80-89**	1.10	1.03, 1.17	80-89**	1.11	1.05, 1.19
90-99	0.98	0.92, 1.04	90-99	0.99	0.93, 1.06
100 and over**	0.61	0.54, 0.70	100 and over**	0.62	0.55, 0.71
Deprivation (Low)	1	-	**Deprivation (Low)**	1	-
High**	1.06	1.03, 1.09	High**	1.06	1.02, 1.09
Rurality (Rural)	1	-	**Rurality (Rural)**	1	-
Urban	1.01	0.98, 1.05	Urban	1.02	0.99, 1.06
Impression Group (Other)	1	-	**Impression Group (Other)**	1	-
Medical**	8.93	8.46, 9.42	Medical**	9.18	8.69, 9.69
Gynae**	23.84	15.37, 36.99	Gynae**	23.57	15.19, 36.57
Mental Health**	3.25	2.93, 3.60	Mental Health**	3.30	2.97, 3.66
Neurological**	9.06	8.42, 9.75	Neurological**	10.26	9.51, 11.07
Trauma**	9.50	8.97, 10.05	Trauma**	10.17	9.59, 10.77
Respiratory**	6.81	6.35, 7.30	Respiratory**	7.30	6.79, 7.84
Cardiovascular**	11.29	10.43, 12.22	Cardiovascular**	11.51	10.62, 12.47
Call Category (1)	1	**-**	**Call Category (1)**	1	**-**
2**	1.48	1.39, 1.57	2**	1.51	1.42, 1.60
3**	1.22	1.14, 1.30	3**	1.23	1.15, 1.32
4**	13.28	11.48, 15.35	4**	15.96	13.63, 18.68
5	1.05	0.79, 1.41	5	1.05	0.78, 1.42
HCP**	15.37	13.41, 17.62	HCP**	19.42	16.66, 22.63
First NEWS2**	1.23	1.22, 1.24	**First NEWS2****	1.22	1.21, 1.23

*** p<0.05; ** p<0.001**

Conclusions

- Imputation methods can be useful and can help researchers avoid excluding valuable data from the analysis.

- Different imputation methods can be used in different scenarios.

- Multiple imputation is a more robust method which avoids bias due to distortions in the variable distribution.

- As a good practice the results should be compared with and without the newly generated values.

References

Allison, P. (2000). Multiple Imputation for Missing Data: A Cautionary Tale, Sociological Methods and Research, 28, 301-309.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling, 15*(4), 651–675.

Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ2007;335:136.

Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj. 2009 Jun 29;338.