Some notes on missing data

Emily Malcolm-White

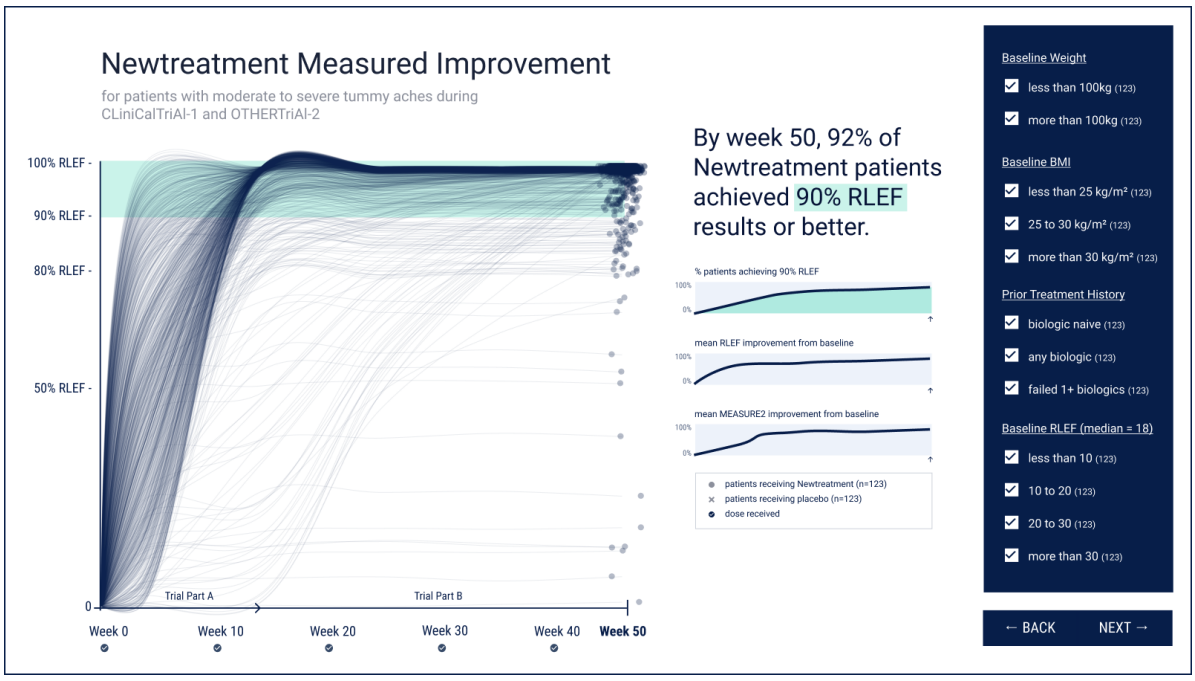
Understand the Nature of Missing Data

If the data isn't missing completely at random (MCAR), you might systematically lose information from a particular subgroup. For example, if respondents with lower incomes are less likely to answer a survey question about income, the overall average income could be overestimated.

- Missing values in key variables can distort summary statistics (mean, median, variance) and lead to misinterpretations of the underlying distribution.
- In time series plots, missing data points can create gaps or misleading trends. For example, if sales data is missing for several months, a line chart might falsely suggest a sharp drop or spike.
- Visualizations like heatmaps or choropleth maps may display blank areas or irregular patterns if underlying data is missing. This might lead to incorrect conclusions about regional performance or density.
- In plots comparing groups, if one group has more missing data than another, the comparisons might be skewed.

Some Examples:

- In a clinical trial, if follow-up data for patients who experienced adverse effects is missing, the overall effectiveness of a treatment may be overestimated, potentially leading to incorrect conclusions about its safety.



- Missing information on key demographics (age, income, education) can lead to misleading subgroup comparisons. For instance, if younger respondents are underrepresented due to missing age data, conclusions about youth preferences might be flawed.
- If key transaction data is missing for segments of customers (like those using alternative payment methods), it may bias analyses of purchasing habits, affecting marketing strategies and revenue predictions.

How to Handle Missing Data

Deletion (if appropriate)

- **Listwise deletion** (Complete-case analysis): Remove rows with missing values.
  - `r drop_na()`
  - Best for missing at random data and when missing values are minimal.
  - Not ideal if a large portion of data is lost.
- **Pairwise deletion**: Use available data in analyses without deleting entire rows.
  - `r drop_na(var1, var2)`

Imputation (Filling in Missing Values)

- **Mean/Median/Mode Imputation:** Replace missing values with the mean, median, or mode of the column.
  - Best for numerical data with few missing values.
- **Forward/Backward Fill** (Time Series Data): Fill missing values using previous or next observed values.
- **K-Nearest Neighbors (KNN) Imputation:** Estimate missing values using similar observations. (Take STAT 218: Statistical Learning)

Use Models to Predict Missing Values

- **Regression-based imputation:** Predict missing values using other variables.(Take STAT 211: Regression or ECON 211: Regression)
- **Machine learning approaches:** Use decision trees or random forests to fill in missing values.(Take STAT 218: Statistical Learning and/or CSCI 451: Machine Learning)

Best Practices

- Perform exploratory data analysis (EDA) to detect patterns in missingness.
- Consult subject matter experts to determine if missing data should be filled, ignored, or used to infer patterns.
- Check missing data percentage: If a variable has a small percentage of missing values (<5%) at random, simple imputation or row deletion might be acceptable.
- When visualizing data, consider marking missing values distinctly or annotating where data gaps exist so that the audience understands the data limitations.

Further Reading

- R for Data Science Chapter on Missing Data