MISS: Benchmarking strategies for handling missing data in statistical tests and models

Supervisor: Fabian Woller

Biomedical datasets often contain substantial amounts of missing data, which may arise from a variety of underlying mechanisms. Common approaches to address missingness include complete-case analysis, pairwise deletion, and imputation, with imputation constituting a particularly active area of research. Because biomedical data typically comprise heterogeneous variable types, their analysis often relies on a combination of statistical hypothesis testing and predictive modeling techniques, such as regression-based models and tree-based methods.

The goal of this project is to determine which missing data handling strategy performs most effectively under different missingness mechanisms in mixed-type datasets. To achieve this, we will employ the recently developed Python tool MissMecha to introduce controlled missingness patterns into publicly available biomedical datasets. Subsequently, we will assess the performance of various methods across different statistical analyses to identify the most robust approach for handling missing data.

The steps of this project can roughly be summarized as follows:

- Find suitable (publicly available) datasets
- Implement the simulation of different missingness patterns using MissMecha
- Implement common statistical tests/models in combination with complete-case analysis/pairwise removal/different imputation methods, possibly with the help of the statistics package NApy
- Evaluate the respective outcomes with respect to ground truth results on the actual input datasets

Requirements:

- Programming skills in Python
- Basic knowledge of common statistical tests (e.g. T-test, ANOVA) and statistical models (regression, tree-based predictors)

Literature:

- Youran Zhou, et al. MissMecha: An All-in-One Python Package for Studying Missing Data Mechanisms. https://doi.org/10.48550/arXiv.2508.04740
- Fabian Woller, et al. NApy: efficient statistics in Python for large-scale heterogeneous data with enhanced support for missing data. https://doi.org/10.1093/gigascience/giaf140