Validation of data analysis code

Oliver Sander and Andy Stein <link>

Planning

Getting Started

- What is the intended use of the modeling/analysis results?
- What is the list of scripts, inputs, and outputs to validate?
- What is the analysis workflow, i.e. what is the input and what is the purpose of each output?

Risk assessment

- What are the consequences of mistakes?
- What is the likelihood of mistakes?
- Where do the authors and the reviewers expect errors to be most likely?

Select validation approach (depending on risk tolerance) Ensure alignment between all parties

- Unit testing
- Spot checks (e.g. ad hoc tests)
- Review code and outputs
- Re-running code step by step and inspecting intermediate results
- Independently reproduce key results and conclusions

Checklist

- **Overall**
  - Reproducibility: Is there a README file that provides an overview for running the code?
  - Readability: Does the code follow best practices and a style guide. Is it appropriately commented?
- **Importing Data**
  - Are you reading the right folder, file, and version?
  - Are missing values handled correctly?
  - Could columns be converted to an incorrect type (e.g. factor/character)
  - What other assumptions are there about the input data
- **Modifying Data**
  - Are there hard-coded constants/flags/labels that could fail if the data changes?
- **Results**
  - Did the analysis script run without errors and all warnings are understood and accepted?
  - Are the results sensible at first sight?
  - Can some numbers (e.g. number of subjects, clearance estimate, etc.) be checked against other outputs (e.g. statistical reports, NCA, etc.)?
  - Are all outputs from the last run? Are there outdated outputs lying around with a risk of being used?