Reproducibility at the Biometrical Journal: Simple tips for writing and publishing clear code to ensure reproducible results

Roman Hornung

Institute for Medical Information Processing, **B**iometry and **E**pidemiology, LMU Munich

Statistical Computing 2024, Schloss Reisensburg, Günzburg

July 30, 2024

Reproducible Research at the Biometrical Journal

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- Serving as a **Reproducible Research Editor** at the Biometrical Journal (**BiomJ**) since May 2022, along with Fabian Scheipl and Lien Le.
- Michael Kammer will join us soon and potentially another person.

Reproducible Research at the Biometrical Journal

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- At the BiomJ authors are **required to submit the code and data** to all their real data analyses and simulations.
- This will (eventually) be **published as a supplement** to the paper on the journal homepage.
- This should (ideally) **ensure the reproducibility** of the results.

Reproducible Research at the Biometrical Journal

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- After the review process of the paper is successfully finished, the paper is only **conditionally accepted**.
- The submitted **code then undergoes** a **review process** to verify/improve...
  - ... the **reproducibility**.
  - ... the code's adherence to certain **structural requirements**, enhancing its manageability for independent individuals.

Reproducible Research at the Biometrical Journal

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- **More than one round of review** is often required before the code meets our publication standards.
- The finally **published code supplement** will receive one of several **badges** related to the degree of reproducibility achieved.
- If the **code review process reveals** that the analyses
  - ... were not performed as claimed or
  - ... do not support the results presented in the paper

  it **can be still rejected** (rare occasion).

Reproducibility Badges at the Biometrical Journal

- **Fully reproducible**: All results presented in the paper are reproducible.
- **Partially reproducible**: This has several reasons/undercategories, for example:
  - data confidentiality issues
  - computational complexity
  - use of proprietary software not available to code reviewers
- **Not reproducible**: For various reasons the results shown in the paper cannot be reproduced — can be given after two unsuccessful review rounds, but should be avoided.

Why Reproducible Results?

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- **Validation of Results:**
  - Code serves as **empirical proof**, validating the results and claims made in studies.
  - **Lack** of accessible code **equates to unproven claims** (adapted from Fabian Scheipl).
- **Clarity and Detail:**
  - Code encapsulates every detail of the analysis flow, **improving** the paper's **readability** by omitting exhaustive descriptions.
  - Ensures **no detail is omitted** or overlooked.
- **Encourages Reuse and Extension:**
  - **Others can use**, modify, and expand upon the published methods, simulation designs, code etc.
  - **Accompanying data availability** enables further research by peers.

Why Reproducible Results?

- **Enhances Integrity:**
  - Publishing code mandates tidiness, aiding in **early error detection**.
  - Results are **more trustworthy** and less suspectible to manipulation.
- **Promotes Open Science:**
  - Sharing code aligns with Open Science principles, **potentially improving funding prospects**.

Preliminary Remarks About the Presented Tips

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- The upcoming **tips** are **based on** the Biometrical Journal's reproducible research **guidelines** and **my experiences** in reproducing results from various papers.
- Specifically, my tips are **based on common challenges encountered** during my work for the Biometrical Journal.
- A **fundamental principle** in code structuring and commenting is: **Keep it simple!**

Commenting – README file

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- Effective **commenting** is **essential for understanding**.
- Incorporating a **README file is essential** for providing instructions for the reproduction.
- **Describe all contents** of the code supplement:
  - Provide a short list of the files and folder names.
  - Shortly explain the purpose/content of each.
- Provide clear, concise **instructions for reproducing** all results:
  - Describe **which scripts need to be run** in which order to produce which results.
  - Explain whether any **manual edits** are required before running the scripts — this should be avoided (if possible).

Commenting – README file

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- **Extensive referencing to** settings or **details** of the analysis is **not enough** and **might be** even **hindering**; those reproducing might not be acquainted with all paper details.
- **Instead**, clearly delineate which scripts produce which figures and tables, **referencing result names** (e.g., **Table 1, Figure 1**).
- Software is constantly updated. ⇒ **Detail** the **versions of the software packages** and the system used, e.g., via the output of `sessionInfo()` in R — essential, but often forgotten.
- **Describe real data** used, their origin and any licensing information. Provide links to sources explaining the data.

Commenting – Code

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- **Comment chunks** of code; it's not necessary to comment every line of code, can even be confusing.
- Comment **user-defined functions**, including their input and output parameters (like in an R package).
- **Mark the locations** in the code where **figures** and **tables** from the manuscript are produced by referencing their names (e.g., Table 1, Figure 1).
- **Remove** unnecessary **comment artifacts** that are not intended for publication.
- Ensure that all comments are in **English**.

Code Structure & Organization

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- Organize files using a **clear folder structure** such as: code, data, results.
- **Save** final **results in** a **readable format**, e.g. figures as PDFs and tables as Excel files (or data.frames in R).
- Ensure output **figures match their appearance** in the paper; e.g. do not save different parts of the same figure in different files.
- Ensure output **tables match their appearance** in the paper: same column and row ordering, same column and row names.

Code Structure & Organization

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- Include commands that **automatically save** the output **figures** and **tables** in the results folder (potentially with subfolders) using unique, descriptive file names.
- **Keep** the **output** table and figure **files** in the results folder. This makes it easier to check reproducibility.
- **Split long R scripts** into separate files according to functionality.
- However, **avoid copy-pasting** code: Do not include multiple copies of the same simulation script changing only some parameters/methods.

Code Structure & Organization

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- Instead, **reused code** is contained **in functions/macros** that are repeatedly called in loops through simulation settings/methods etc.
- For clarity, **store functions** used in the analysis **in separate scripts**, sourced in the analysis scripts.
- Save all iterations of a simulation in a **single (.Rda) file**.
- Save R code in files with **extension .R or .r**. Never use txt files, PDFs, Word files etc.
- The code should be in **ASCII files**, preferably in UTF-8 encoding.

Code Structure & Organization

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- If **self-developed R packages** are included, save these as **package bundles** (i.e., .tar.gz) rather than .zip files.
- If self-developed R packages are provided on **GitHub** it is crucial to specify the **commit hash** (e.g., 1a2b3c4d5e6f7g8h9i0j1k2l3m4n5o6p7q8r9s0t) because the package may change in the future.
- If possible/meaningful, include a **master script** that calls other scripts in your supplement to produce all results.
- If using Windows, ensure that you use the **correct capitalization** of folders in paths because other systems are case-sensitive.

Coding Guidelines & Naming Conventions

Reproducibility at the BiomJ

Roman Hornung

Background

Commenting

Structure & Form

Achieving Reproducibility

Code & Data Publication

Advanced Tips

Future Challenges & Open Questions

Take Home Messages

- **Avoid absolute paths**; **use relative paths** such as "./code/results/figure1.pdf" instead of "C:/Users/Sepp/Arbeit/code/results/figure1.pdf".
- Choose **descriptive file names** like "simulation.R" over ambiguous ones like "rs_rfg_3.R".
- Use **underscores instead of spaces in** file and folder **names**.
- All object names and file names should in **English**.

Coding Guidelines & Naming Conventions

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- Ensure **code** is **well-spaced and properly indented** for ease of reading.
- **Avoid** excessive line lengths (max. 80-100 characters), as **long lines** make the code difficult to read.
- **Do not** include commands that **manipulate the users' setup**, e.g. by installing packages automatically; worst case: Deleting the users workspace using: `rm(list=ls())`
- Code should be as **platform-independent** as possible, e.g. avoid Windows-specific commands like `windows()`.

Coding Guidelines & Naming Conventions

- **Delete out-commented lines** that are no longer used.
- Similarly, delete **code** that is no longer used.

Achieving Reproducibility

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- Ensure **every figure and table** presented in the paper is **reproducible**.
- For analyses involving randomness, **set the seed** of the random number generator to ensure reproducibility.
- **Repeat** your simulation with **different seeds** to confirm that the Monte Carlo error is sufficiently small, ensuring the validity of the key findings (e.g., method rankings).
- **Conduct reproducibility checks** by running individual scripts; **ideally** have a **project partner** reproduce results to uncover potential errors.

Purposes of Intermediate Results

Reproducibility at the BiomJ

Roman Hornung

Background

Commenting

Structure & Form

Achieving Reproducibility

Code & Data Publication

Advanced Tips

Future Challenges & Open Questions

Take Home Messages

- **Intermediate results** are the **"raw" results** of the analysis (e.g., results of individual simulation iterations).
- These results allow for **reproducibility spot checks** without repeating the entire analysis, **crucial** for computationally expensive analyses.
- They **also allow** you to make **changes or additions at a later time**, such as editing or adding figures.

Generation and Use of Intermediate Results

Reproducibility at the BiomJ

Roman Hornung

Background

Commenting

Structure & Form

Achieving Reproducibility

Code & Data Publication

Advanced Tips

Future Challenges & Open Questions

Take Home Messages

- Save the **results of each simulation iteration**.
- Make sure that these **unaggregated results** are **reproducible**:
  - Ideal: Use a **different seed for each iteration**, where the collection of used unique seeds is sampled reproducibly from a large number of possible seeds.
  - Also possible: Set one seed in the beginning and then loop through the iterations.

Generation and Use of Intermediate Results

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- Use **one script for calculations** and intermediate results storage.
- Use **another one to evaluate** the intermediate results, producing final results (figures, tables).
- All intermediate result files **must be producible** by code available in the supplement for the results to be reproducible.

Code & Data Publication

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- **Preferred**: Ideally, the code should be published as a **supplement on the journal's homepage** alongside the paper.
- **GitHub**: An **acceptable** alternative. Crucially, **specify** the **commit hash** in the paper since repositories are subject to changes.
- For supplements with **large intermediate results**, use platforms like **figshare**.
- Strive to **make data available**. **If unfeasible**, provide **pseudo data**.
- Combine all files into a **single, organized folder**, regardless of the publication method chosen.

Advanced Tips

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- Using the output of `sessionInfo()` **requires manual installation** of the old R package versions.

  — **Advances solutions** like Docker, `renv`, `checkpoint`, and `packrat` streamline R environment setup by providing all necessary package versions.

- **Code review among colleagues** can identify issues in the code and improve code quality.

- Weekly or monthly **internal coding clubs** can help share knowledge about good coding practices.

- **Online coding communities** can also help here, in particular regarding more intricate issues.

Future Challenges & Open Questions

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- How can the code review process be **further accelerated**?
- Are the **guidelines available** sufficient, or perhaps even overwhelming for less computationally savvy people?
- How could the **quality** of initial submissions be made **more consistent**?
- We usually review code produced in the last few years. Are the results **reproducible in the long run**?

Take Home Messages

Reproducibility at the BiomJ
Roman Hornung
Background
Commenting
Structure & Form
Achieving Reproducibility
Code & Data Publication
Advanced Tips
Future Challenges & Open Questions
Take Home Messages

- Availability of **code** ...
  - ... serves as **empirical proof of results**.
  - ... **facilitates reuse** and further development.
  - ... potentially **improves funding opportunities**.
- **Clarity and organization**: clear folder structures, README with clear instructions, (if possible) few files, clear alignment of code with results in the paper.
- **Intermediate results** enable easier reproducibility and later modifications.
- **Final reproducibility spot checks** to catch potential errors **before publication**.
- If possible, **publish code** as a supplement **with the paper** to ensure long-term accessibility.

Thank you for your attention!

Reproducibility at the BiomJ

Roman Hornung

Background

Commenting

Structure & Form

Achieving Reproducibility

Code & Data Publication

Advanced Tips

Future Challenges & Open Questions

Take Home Messages