Data and Code for Reproducible Research

Lessons Learned from the NLM Reproducibility Workshop

Maryam Zaringhalam, PhD
Lisa Federer, PhD, MLIS
Office of Strategic Initiatives
National Library of Medicine

THOUGHT EXPERIMENT

What can we learn by trying to reproduce published research?

The goal was to provide participants with…

01

Knowledge of tools for reproducible research and NLM data resources for bioinformatics

02

An understanding of how to incorporate these tools into their own research practices

03

A path towards a deliverable, in the form of an executable notebook and/or publication

… while also helping us think about …

How might a curriculum around reproducibility take shape?

How are researchers approaching reproducibility?

What is some low-hanging fruit to promote reproducible research practices?

Structure: NLM Reproducibility Workshop

- Three-day workshop for 25 intramural NIH researchers
- Worked in 5 teams to reproduce a bioinformatics paper, with underlying data available in NLM-hosted repositories
- Day 1
  - Primer on open science and reproducibility
  - Three 30-minute tutorials on
    - Executable notebooks (Jupyter)
    - Version control (Git and Github)
    - Containerization (Docker)
- Days 2-3: Teams work in groups, code-a-thon style

TAKEAWAYS

No papers were successfully reproduced

Reproducibility is not trivial

- Missing underlying data
- Missing software and tools
- Inadequate descriptions of software and tools
- Workflows inadequately described or difficult to follow

Need better minimum standards for peer review

- Underlying raw data are made readily available
- All software and tools must detail the appropriate version
- Underlying analysis tools are made readily available

Still many different ways to interpret reproducibility

- Raw versus processed data
- Re-using scripts versus re-engineering them
- Re-creating the computing environment versus using an environment that’s “close enough”
- Re-generating the figures versus re-generating the general conclusions

Clarity and community consensus around expectations for reproducibility could go a long way

Communication for open science

- Some teams reached out to corresponding authors for data or with questions about methods
- Authors responded within hours, suggesting that lack of reproducibility, in many cases, isn’t the result of bad faith!

QUESTIONS?

Maryam Zaringhalam, PhD

NLM Data Science & Open Science Officer
Maryam.Zaringhalam@nih.gov

Lisa Federer, PhD, MLIS

NLM Data Science & Open Science Librarian
Lisa.Federer@nih.gov

Office of Strategic Initiatives

National Library of Medicine
National Institutes of Health