BST 270 Reproducible Data Science
Winter 2023,
M-F January 09-13, T-TH January 17-19, 9:00am-12:00pm

Instructor	Viola Fanfani, vfanfani@hsph.harvard.edu. Address: HSPH, Building 2, Room 412
Instructor	John Quackenbush, johnq@hsph.harvard.edu. Address: HSPH, Building 2, Room 421
Guest lecture	Jonas Fischer
Office hours	by appointment

Purpose of this course

Reproducible research has become increasingly important in the biomedical sciences. The science community has recognized reproducibility is a growing challenge in basic, clinical and population sciences. Experimental design, data provenance, analytic methods and tools, and reporting science play a critical role in the biomedical research ecosystem to ensure scientic rigor, robustness and transparency. Statistical and computational methods and tools are fundamental for making scientic results reproducible.

Course Description and Structure

The central theme of the course will be to meet these scientific needs of reproducible science through training in reproducible research. The topics covered in this course include the fundamentals of reproducible science, case studies in reproducible research, data provenance, statistical methods for reproducible science, and computational tools for reproducible science. This is a blended course where students are introduced to course content online through videos and reading assignments, and then shown how to use the tools and methods described in the videos in class to conduct reproducible research.

In-class lectures will be interactive and focus on applying the tools and methods introduced in the videos to reproduce a recently published scientic journal article:

[1] Boehm, J. K., Williams, D. R., Rimm, E. B., Ry , C., & Kubzansky, L. D. (2013). Relation between Optimism and Lipids in Midlife. The American Journal of Cardiology, 111(10), 1425-1431.
http://doi.org/10.1016/j.amjcard.2013.01.292

In 1995, MIDUS survey data were collected from a total of 7,108 participants. The baseline sample was comprised of individuals from four subsamples: (1) a national RDD (random digit dialing) sample (n=3,487); (2) oversamples from five metropolitan areas in the U.S. (n=757); (3) siblings of individuals from the RDD sample (n=950); and (4) a national RDD sample of twin pairs (n=1,914). All eligible participants were non-institutionalized, English-speaking adults in the contiguous United States, aged 25 to 74. All respondents were invited to participate in a phone interview of approximately 30 minutes in length and complete 2 self-administered questionnaires (SAQs), each of approximately 45 pages in length. In addition, the twin subsample was administered a short screener to assess zygosity and other twin-specic information. With funding provided by the National Institute on Aging, a longitudinal follow-up of MIDUS I began in 2004. Every attempt was made to contact all original respondents and invite them to participate in a second wave of data collection. Of the 7,108 participants in MIDUS I, 4,963 were successfully contacted to participate in another phone interview of about 30 minutes in length. MIDUS II also included two self-administered questionnaires (SAQs), each of about 55 pages in length, which were mailed to participants. The overall response rate for the SAQs was 81%. Over 1,000 journal articles have been written using MIDUS I and II data since 1995.

We will attempt to reproduce the findings of [1] and critique the reproducibility of the article as a class. This particular article focuses only on MIDUS II data, including biomarker data, and investigates the relationship between optimism and lipids. You can download the MIDUS II data and supporting codebook and other documents here. You can download the biomarker data here.

Course Objectives

Upon successful completion of this course, you should be able to:

- Describe the fundamentals and importance of reproducible research
- Assess the reproducibility of other research
- Create a fully reproducible research project
- Develop new methods and tools for reproducible research

Credits

This is a 2.5 credit, Pass/Fail course. A minimum total score of 70% is needed to pass this course.

Course Materials

Course videos, electronic copies of course readings, rubrics, guidelines, notes/slides, useful website links and datasets will be posted on the edX course website, course Canvas site and course GitHub repository. Students will need to create an edX account and sign-up for the course.

Optional Reading

- Christopher Gandrud (2015), Reproducible Research with R and RStudio, 2nd Ed.
- Kitzes, Turek, Deniz (2017), The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences, 1st Ed.

Grading

Homework, In-class Participation and Attendance (60%)

Homework assignments will consist of viewing online course videos (edX videos) and submitting 2 questions or comments for use in the discussion during class.

Throughout the course we as a class will be attempting to reproduce a recently published journal article in class. Students are expected to participate in class activities that will involve completing tasks related to reproducing the journal article. The reproducibility of the article will be discussed and critiqued in class.

Class attendance is mandatory. Students will be excused from class in the event of a family emergency, medical issue, religious observance, or other extenuating circumstance, and should contact the instructor to inform them of their absence. A maximum of one absence is allowed without penalty to the class attendance grade and overall grade. A 20% deduction in class attendance grade will be taken for every additional missed class.

Individual project (40%)

Each student will be given a dataset and prompt and will be expected to answer all questions with code or text and have the project be completely reproducible. The project must be submitted in the form of a Jupyter notebook or RMarkdown file and corresponding converted/knitted pdf, with commented code and text interspersed. All project documents must also be uploaded to a GitHub repository each student will create within the reproducible data science organization here. The repository must also include a README file describing the contents of the repository (files, data) and how to reproduce all results. Students should keep in mind the file and folder structure we will cover in class and make the reproducible process as automated as possible

Schedule

Day 1 - Jan 9

In class:

- Introduction to course

1. Module 1: Introduction to Reproducible Science (Total time: 10:55 min)
    1.1. Welcome to Reproducible Science (5:14)
    1.2. Intro to the People/Faculty (2:39)
    1.3. Intro to the Modules (3:02)

At home:

- Create GitHub account (if you don't already have one)
- Send teaching staff your GitHub username
- Clone course repository to laptop/computer

2. Module 2: Fundamentals of Reproducible Science (Total time: 40:40 min)
    2.1. Why this matters (13:35)
    2.2. Definitions and concepts (5:58)
    2.3. Factors affecting reproducibility (3:30)
    2.4. Experimental design (4:25)
    2.5. Organization (3:27)
    2.6. Understanding yourself later (1:52)
    2.7. Ensuring others can continue your work (5:01)
    2.8. Providing workflows and results (2:52)

Day 2 - Jan 10

In class:

- Brief discussion of Module 2 videos
- Discuss in-class project
- Create data dictionary for in-class project
- Wrangle data for in-class project

At home:

3. Module 3: Case Studies of Reproducible Science (Total time: 61:29 min)
    3.1. Introduction to case studies (2:51)
    3.2. Potti 2006 (7:40)
    3.3. Baggerly and Coombes (17:22)
    3.4. Ioannidis 2009 (10:56)
    3.5. Case studies wrap up (5:18)
    3.6. Introduction to reproducible reporting science (1:51)
    3.7. Journals and reproducible research (9:12)
    3.8. NIH guidance on reproducible research (6:19)

Day 3 - Jan 11

In class:

- Brief discussion of Module 3 videos
- Reproduce Figure 1 for [1]
- Reproduce Table 1 for [1]

At home:

4. Module 4: Data Provenance (Total time: 75:44 min)
   4.1. Introduction to data provenance (2:40)
   4.2. Data provenance concepts (4:37)
   4.3. Experimental design before the computer (4:53)
   4.4. Experimental design before on computer (6:39)
   4.5. Tools and standards (3:09)
   4.6. Work ows (5:19)
   4.7. Journals and reporting (4:17)
   4.8. Mechanisms for reporting (5:26)
   4.9. Data-type-specific repositories (5:05)
   4.10. General repositories (5:13)
   4.11. Code (2.14)
   4.12. Documentation (3:25)
   4.13. Data privacy and security (6:56)
   4.14. Privacy (5:50)
   4.15. Security (6:01)

Day 4 - Jan 12

Guest instructor: Jonas Fischer **In class:**

- Brief discussion of Module 4 videos
- Reproduce table 2 of [1]
- Reproduce table 3 of [1]

At home:

5. Module 5: Statistical Methods for Reproducible Science (98:21 min)
   5.1. Introduction to cross study validation of prediction models (10:58)
   5.2. Statistical methods introduction (1:19)
   5.3. Coefficient of determination (8:12)
   5.4. Brier Score (7:25)
   5.5. Area under the curve (AUC) (9:19)
   5.6. Concordance in survival analysis (16:18)
   5.7. Motivation for cross validation (10:53)
   5.8. Cross validation (6:33)
   5.9. Bootstrap (6:56)
   5.10. Simulations (11:05)
   5.11. Clustering (9:23)

Day 5 - Jan 13

In class:

- Brief discussion of Module 5 videos
- Reproduce Table 4 of [1]
- Reproduce Table 5 of [1]

At home:

6. Module 6: Computational Tools for Reproducible (Total time: 69:45) Science
    - 6.1. Computational tools introduction (2:50)
    - 6.2. Contributor introductions (2:20)
    - 6.3. Introduction to editors (2:42)
    - 6.4. Introduction to R and Rstudio (2:58)
    - 6.5. Introduction to Python (0:46)
    - 6.6. Introduction to Git and GitHub (1:08)
    - 6.7. Downloading and installing Git (2:54)
    - 6.8. How to create a repository (2:36)
    - 6.9. GitHub interface (2:01)
    - 6.10. A conversation with Eric Surface (11:46)
    - 6.11. A conversation with Merce
    - 6.12. (3:09)
    - 6.13. A conversation with Merce Crosas 2 (5:42)
    - 6.14. A conversation with Merce Crosas 3 (2:33)
    - 6.15. A conversation with Merce Crosas 4 (0:47)
    - 6.16. A conversation with Merce Crosas 5 (6:23)
    - 6.17. A conversation with Merce Crosas 6 (3:20)
    - 6.18. A conversation with Merce Crosas 7 (5:19)
    - 6.19. A conversation with Merce Crosas 8 (3:48)
    - 6.20. A conversation with Merce Crosas 9 (2:07)
    - 6.21. A conversation with Merce Crosas 10 (4.36)

Day 6 - Jan 17

In class:

- Brief discussion of Module 6 - part 1 videos
- Reproducibility critique of [1]
- Begin working on individual project

At home:

6. Module 6: Computational Tools for Reproducible Science (Total time: 89:16 min)
   6.1. A conversation with Sonia Barbosa (12:28)
   6.2. Dataverse: Sonia Barbosa P1 (18:58)
   6.3. Dataverse: Sonia Barbosa P2 (14:06)
   6.4. API access (3:03)
   6.5. Introduction to dynamic report genera- tion (1:07)
   6.6. A conversation with Christopher Gan- drud (8:31)
   6.7. knitr and Rmarkdown with Christopher Gan- drud (17:37)
   6.8. Notebooks: R notebook (3:13)
   6.9. Notebooks: Compiling in an R File (0:30)
   6.10. Notebooks: Jupyter (2:34)
   6.11. Make files (7:09)

Day 7 - Jan 18

In class:

- Brief discussion of Module 6 - part 2 videos
- Continue working on individual project

At home:

6. Module 6: Computational Tools for Reproducible Science (Total time: 52:33 min)
   6.1. A conversation with Simon Adar 1 (4:54)
   6.2. A conversation with Simon Adar 2 (1:47)
   6.3. A conversation with Simon Adar 3 (7:09)
   6.4. A conversation with Simon Adar 4 (3:41)
   6.5. A conversation with Simon Adar 5 (4:06)
   6.6. A conversation with Simon Adar 6 (5:58)
   6.7. A conversation with Simon Adar 7 (0:49) 6
   6.8. A conversation with Simon Adar 8 (1:05)
   6.9. A conversation with Simon Adar 9 (4:15)
   6.10. A conversation with Simon Adar 10 (10:43)
   6.11. Code Ocean conclusion (0:48)
   6.12. Creating a new algorithm in Code Ocean (3:40)
   6.13. Conclusion (1:30)
   6.14. Course conclusion (2:08)

Day 8 - Jan 19

In class:

- Brief discussion of Module 6 - part 3 videos
- Course Summary
- Finish individual project

At home:

- Submit course evaluation