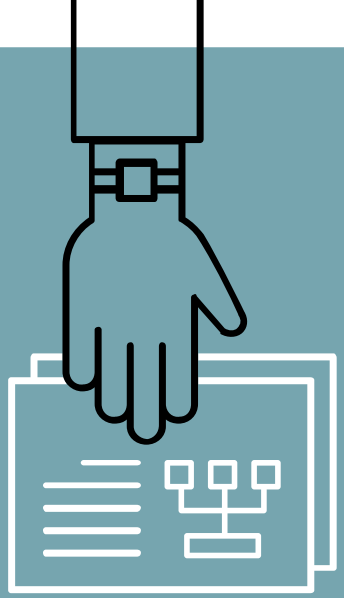
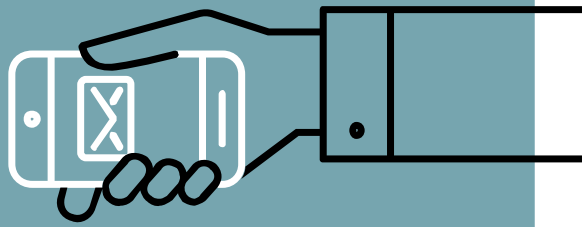
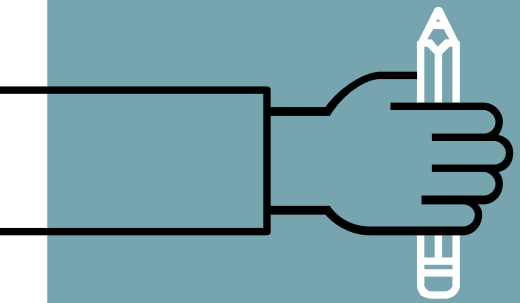
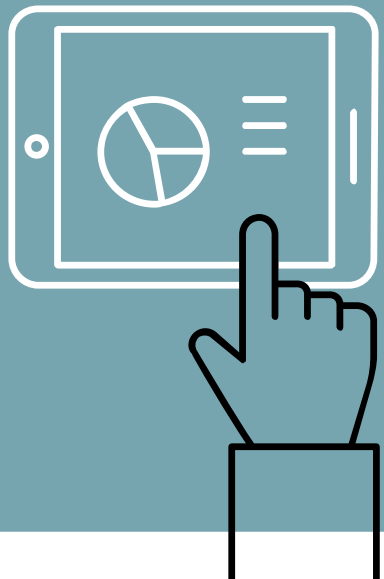




BST 270
Reproducible Data Science

Winter 2021
Session 3





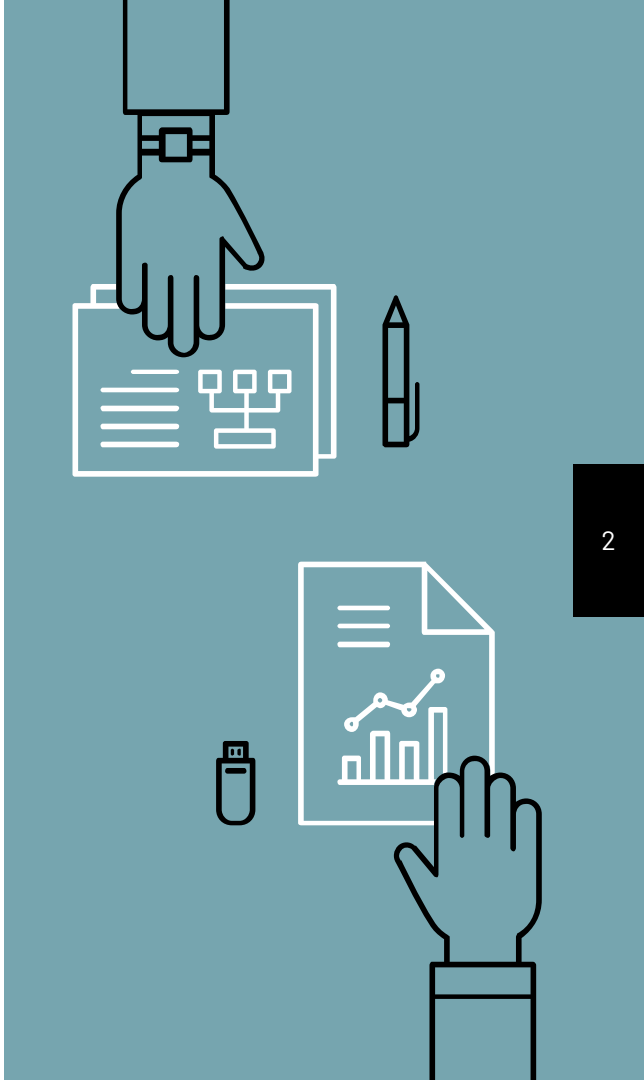
Module 3 Discussion

How do you make a study reproducible if you can't share the data due to HIPAA or other restrictions?

- Journals may request you create a synthetic data set that mimics the data used and provide that and your code
- Some journals don't make you do this yet

What sanity checks do you do on publicly available datasets before analysis?

- Exploratory data analysis with plots, summary statistics, etc.
- Reading the documentation about how, when and by whom the data was collected/organized and how it has transformed at all (cleaning/wrangling)



Module 3 Discussion

How do you make sensitive data (for example, medical records) publicly available?

- The MIMIC Critical Care Database created and maintained by researchers at MIT is a great example
- They alter the data a bit to make it unidentifiable
  - Change an individual's date of birth to be a date in the future (e.g. 2100)
  - Change the dates of ICU admission
  - Add noise/perturbations to the measurements in a way that doesn't alter the conclusions reached from analyses
- Takes a ton of work and expertise to do this
- Check out other publicly available data here

Module 3 Discussion

In the Ioannidis study the lack of accessibility to the original data was a major hindrance to the reproducibility of the 18 papers evaluated. Is there a standard public website where we typically publish our raw data, or is this on a paper-to-paper basis?

- It is on a paper-to-paper basis, but Harvard affiliates are encouraged to publish their data on Harvard Dataverse
  - You don't have to be a Harvard affiliate to publish data on the platform
  - You can also upload code and metadata
- No "standard" at the moment

Module 3 Discussion

Seeing how long it took to get to the bottom of the Nevins and Potti paper makes me wonder how many other fabricated or faulty results are still out there. It's reassuring that there are pushes for more reproducible research, but I'm sure there are still gaps in the guidelines. Maybe there should be more independent institutions dedicated to looking at reproducibility?

Module 3 Discussion

Seeing how long it took to get to the bottom of the Nevins and Potti paper makes me wonder how many other fabricated or faulty results are still out there. It's reassuring that there are pushes for more reproducible research, but I'm sure there are still gaps in the guidelines. Maybe there should be more independent institutions dedicated to looking at reproducibility?

- It's scary how many papers have been retracted
  - There is a **database** - an entire database! - of retracted papers
  - There is a list of authors with the most retracted papers. The current record is 169
    - How are these authors still allowed to publish??
- Yes, there should definitely be more independent institutions focusing on reproducibility

Top 10 retracted authors

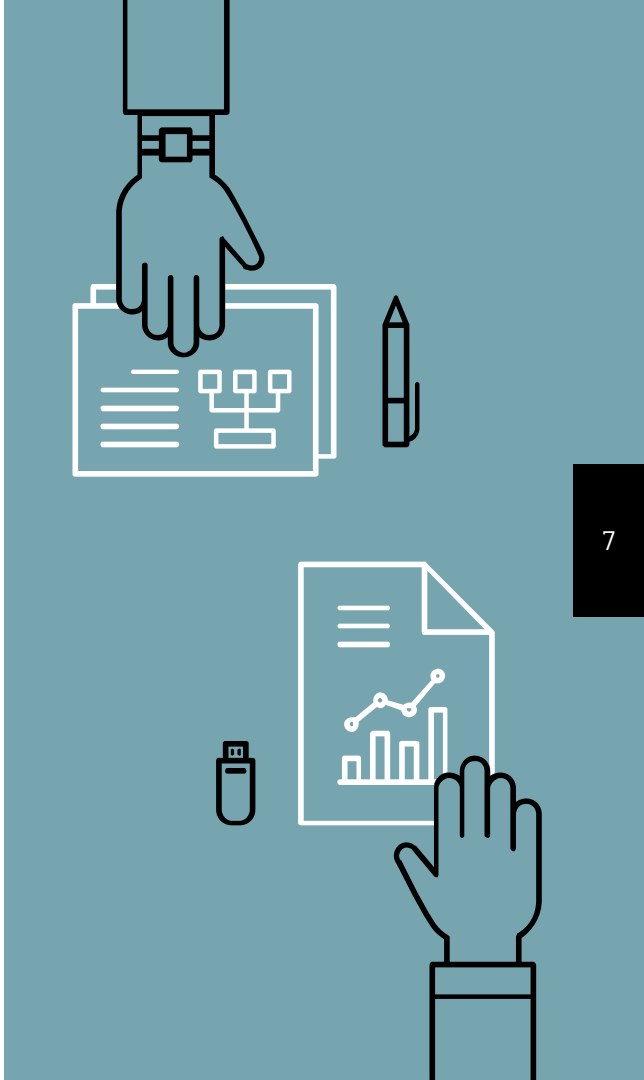
Yoshitaka Fujii, Japan	169
Joachim Boldt, Germany	96
Diederik Stapel, Netherlands	58
Chen-yuan Peter Chen, Taiwan	43
Yoshihiro Sato, Japan	43
Hua Zhong, China	41
Shigeaki Kato, Japan	39
James Hunton, United States	36
Hyung-in Moon, South Korea	35
Jan Hendrik Schön, United States	32

Module 3 Discussion

Do you recommend Sweave, Jupyter or knitr in our efforts of documenting what we have done?

- I recommend using what you prefer - anything that gets you to document what you do is great - or whatever your PI/team uses

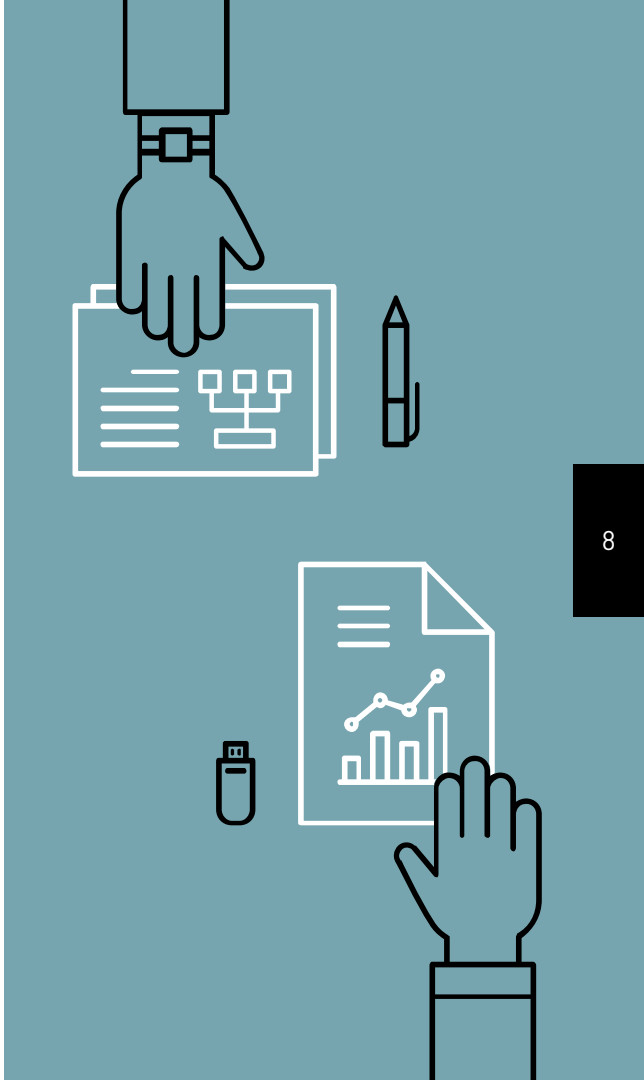
- I personally haven't used Sweave (it looked confusing to me 10 years ago) so I use Jupyter and knitr for documenting code I write



Module 3 Discussion

What are some steps to ensure a reproducible paper when you're working with several different people (and therefore not doing all the work yourself?)

- If someone writes code, have at least one other collaborator test the code and check for any bugs
  - The more eyes to check it the better
- Have meetings consistently throughout the project and take notes about what was discussed in each meeting
  - Discuss any major decisions as a team to think through any potential issues or biases - having a diverse team is great for this (diverse in background and expertise)
- Have the entire team review the final project before publishing or sharing anything



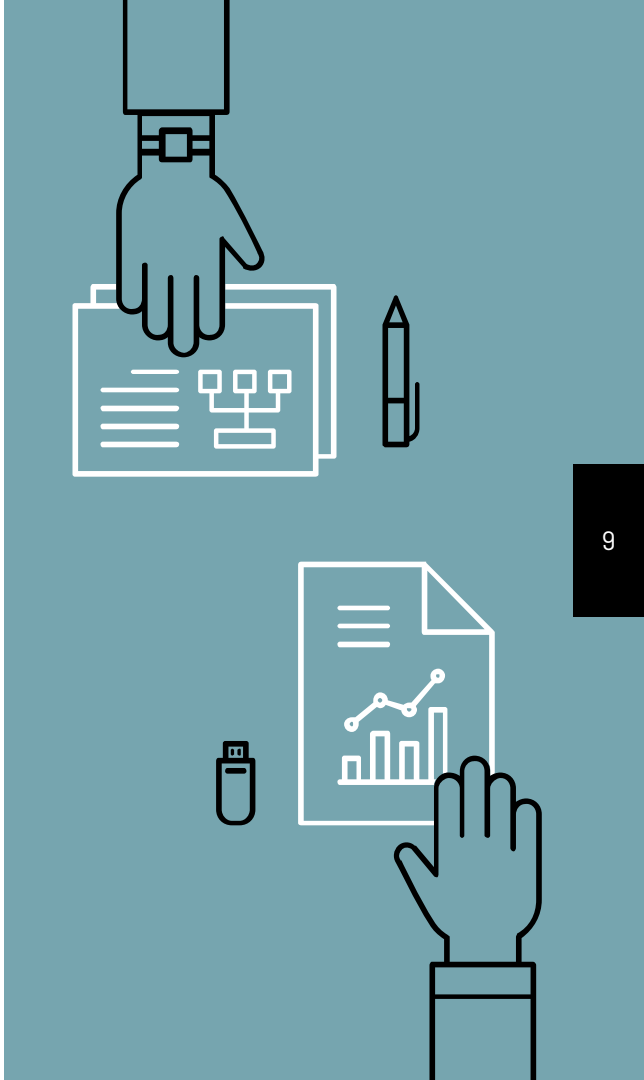
Module 3 Discussion

What is forensic bioinformatics?

- Bioinformatics for reproducing or validating a study
  - What Ioannidis et al did for those 18 other papers

What do reproducibility and replicability look like for other non-computational (or at least not as computational) fields?

- For wet lab fields, thorough documentation of how samples were created and tested
- This paper has a nice introduction about reproducibility and replicability in the social sciences
- Almost all scientific papers use some sort of data to support their conclusions, so the reproducibility workflow we are learning about applies
- For something like history or humanities research, the reproducibility lies in the references



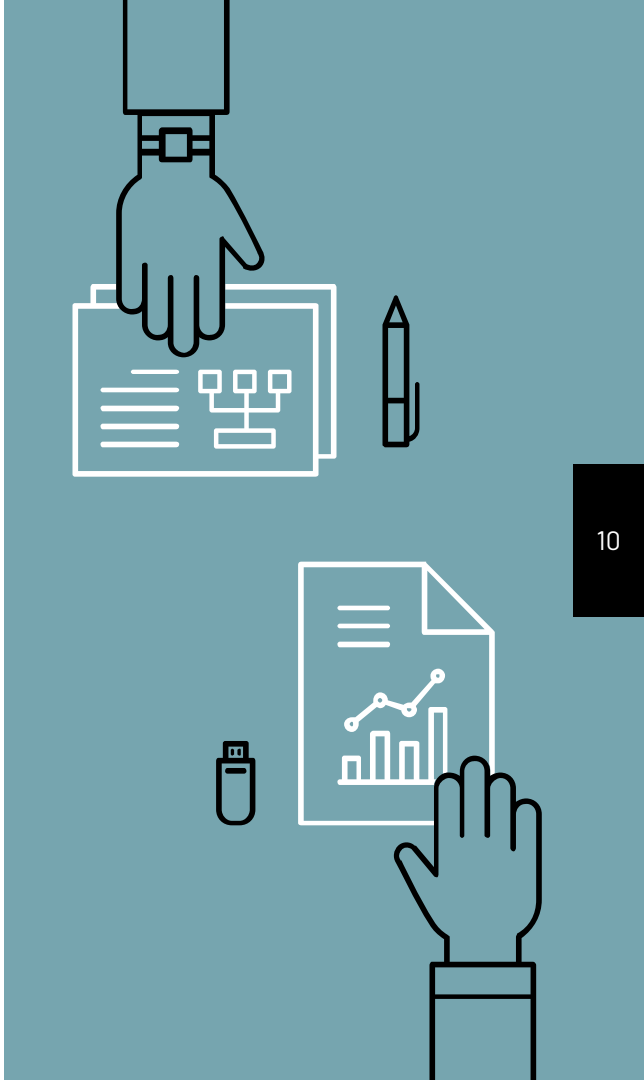
Module 3 Discussion

Obviously it is primarily the responsibility of the original research team to ensure their study is scientifically/statistically reliable, but what about the peer reviewers and the journal itself? For instance, I am surprised that the Potti papers were able to pass the peer review process. Also, Dr. Quackenbush provided some examples where it sounded like the journal was notified of issues but chalked it up to difference in scientific opinion. I want to emphasis that I still believe that researchers should be responsible, but I am just curious if there are any repercussions for the peer reviewers or journals?

- Unfortunately, none that I'm aware of. Journals may take a hit to their reputation and reviewers may not be asked to review again or quite as often.
- Journals do take rebuttals/responses more seriously now.

Why would some journal not accept the principles and guidelines that NIH suggested for rigor and reproducibility?

- More work for everyone involved and investment of resources.

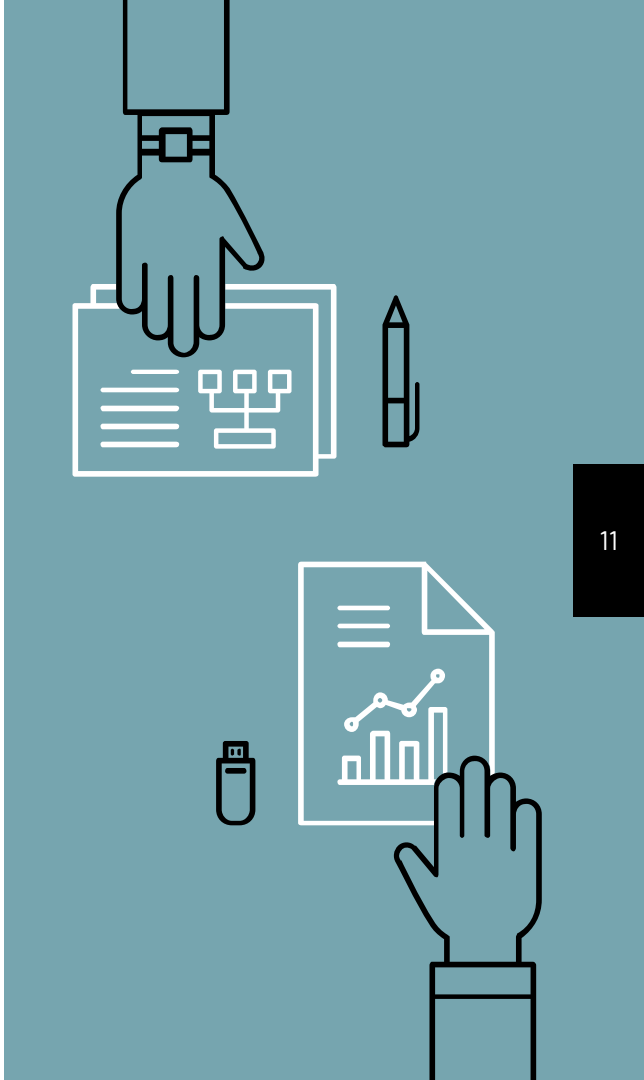


Module 3 Discussion

In video 3.3, when Professor Quackenbush mentions that before moving onto clinical trials, we need a clean verification study, what does this mean exactly? Does this mean we need to have a second paper verify and reproduce our papers before we should move into clinical trials?

- It means someone not part of the team who did the study and analyses should try to reproduce the findings using the same data and methods, and a clinical trial should only move forward if everything can be reproduced.

- "Independent" here means another capable researcher that had nothing to do with the original work (so they aren't biased).

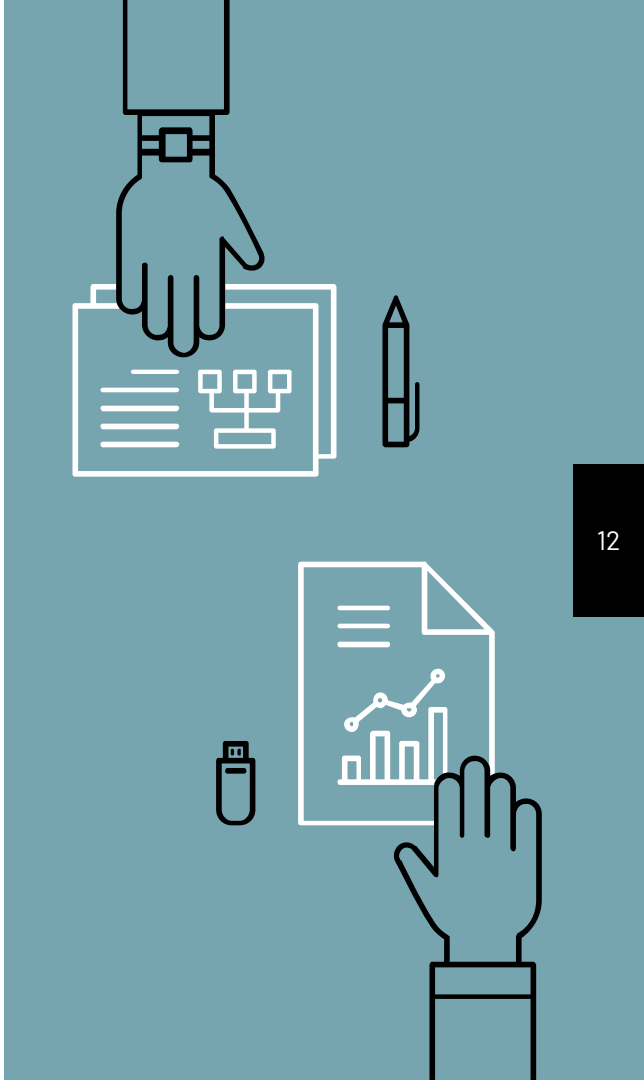


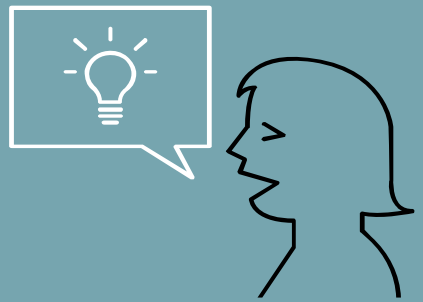
Module 3 Discussion

Can we discuss ways to validate non-genetic data? For example, survey data.

We understand that people may have the chance to make mistakes on reproducibility issues. But I wonder whether there are criteria to determine that such issues are caused deliberately, or only by a proportion (but not all) of the research personnel listed as one of the authors.

- One of the biggest indicators of deliberate issues is fabricating data
- It's more difficult for those outside of the research team to pinpoint who is responsible, but some papers do require a statement in every paper describing which authors were responsible for which parts of the paper.





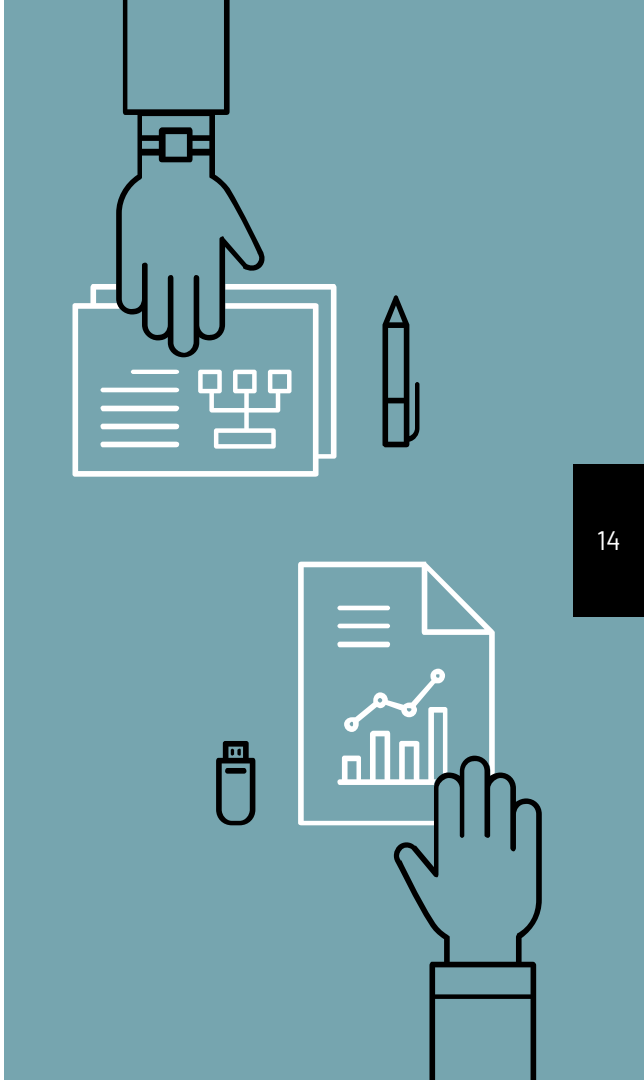
In-Class Project

We will attempt to reproduce the results from and critique the reproducibility of:

[1] Boehm, J. K., Williams, D. R., Rimm, E. B., Ryff, C., \& Kubzansky, L. D. (2013). Relation between Optimism and Lipids in Midlife. The American Journal of Cardiology, 111(10), 1425-1431.

Specific Tasks:

- Create a data dictionary
- Wrangle data
- Recreate Figure 1
- Recreate Tables 1-5
- Critique reproducibility

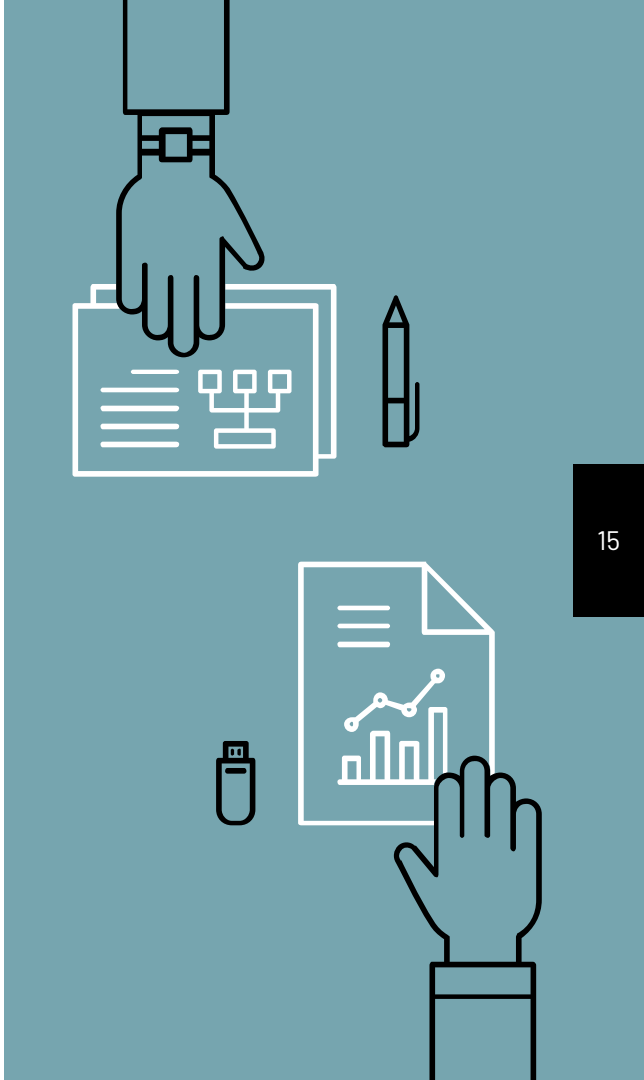


We will attempt to reproduce the results from and critique the reproducibility of:

[1] Boehm, J. K., Williams, D. R., Rimm, E. B., Ryff, C., \& Kubzansky, L. D. (2013). Relation between Optimism and Lipids in Midlife. The American Journal of Cardiology, 111(10), 1425-1431.

Specific Tasks:

- Create a data dictionary
- Wrangle data
- Recreate Figure 1
- Recreate Tables 1-5
- Critique reproducibility



Homework

- Watch Module 4 videos
- Submit Module 4 discussion points