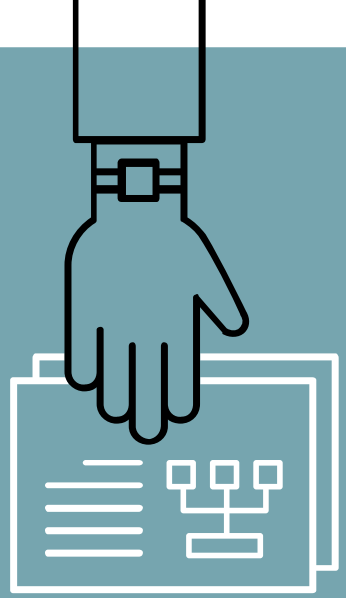
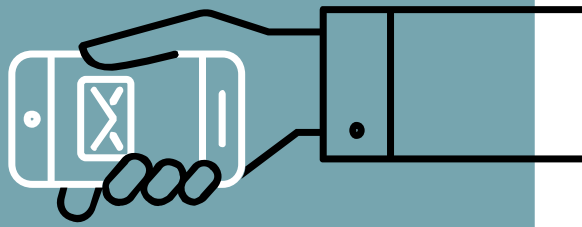
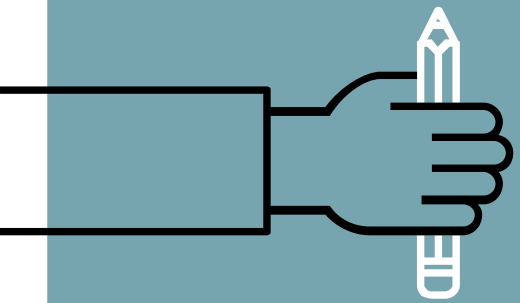
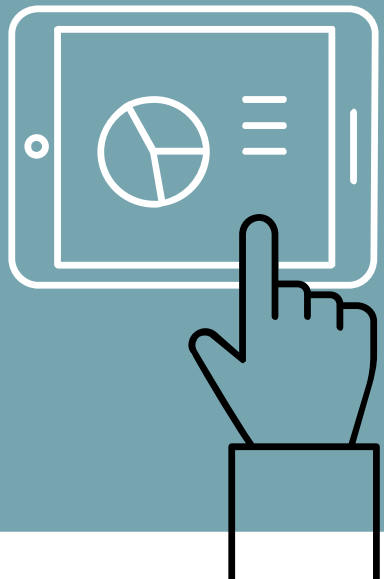




BST 270
Reproducible Data Science

Winter 2022
Session 3

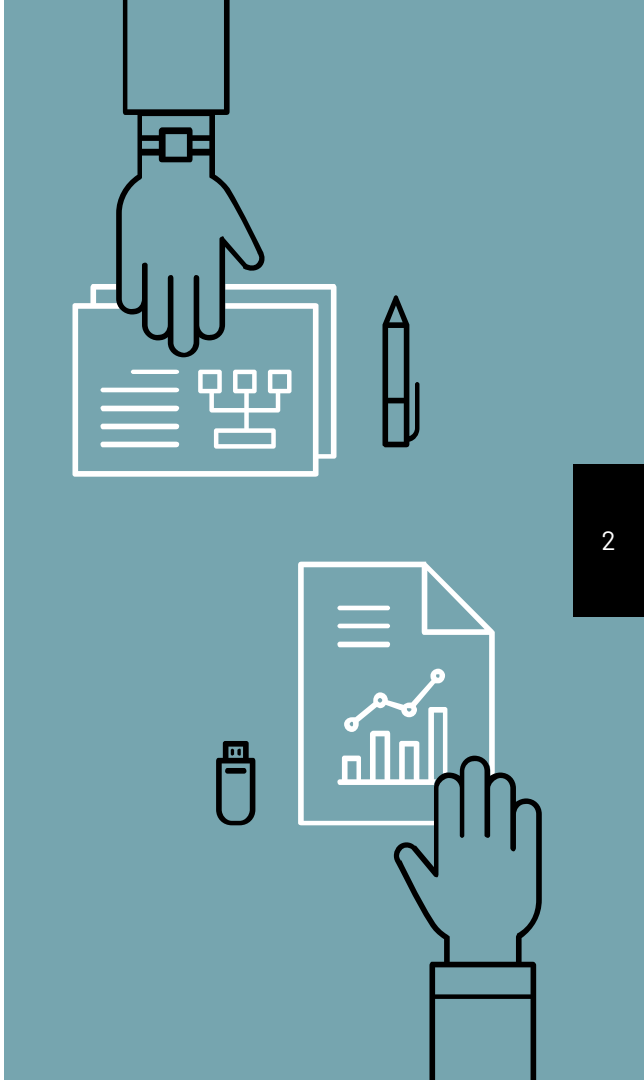




Module 3 Comments

Interesting that experiments that didn't work should be included, because that helps to really understand the process/analyses done. Sometimes I question what motivated authors to choose a certain analysis and guess that its because its what worked best, guiding their paper for better or worse

In the case study of Potti, incorrect results/concluding of a paper led to wrong treatments used in the clinical trials. This makes me think that a higher standard of reproducibility needs to be established for studies that potentially lead to clinical trials, e.g. the results have to be successfully reproduced by other researchers. This needs to be checked both before the publication of such papers and before the clinical trials.



Module 3 Comments

Some journals still have word limit on the Methods section, which makes reporting the detailed analysis procedure difficult. To encourage other researchers to reproduce the findings, journals should set a more generous word limit for methods section, to facilitate other researchers to re-run the analysis on their own.
After working so many years in an epidemiology department it is interesting to learn about differences in journal guidelines and how in biostatistics journals methods are so highly deemed, and how much detail and supplemental materials are required. I feel like most of the journals my research team published were bounded by word counts and number of supplemental materials which limited the amount of detail we could provide about our analysis.

Reproducibility in bioinformatics seems extremely important for us students to learn, given the many mistakes waiting to happen as walked through from Baggerly/Coombes dissection. Even as I reflect back on my own research experiences, many things could have been done better to conform to reproducibility guidelines.

Module 3 Comments

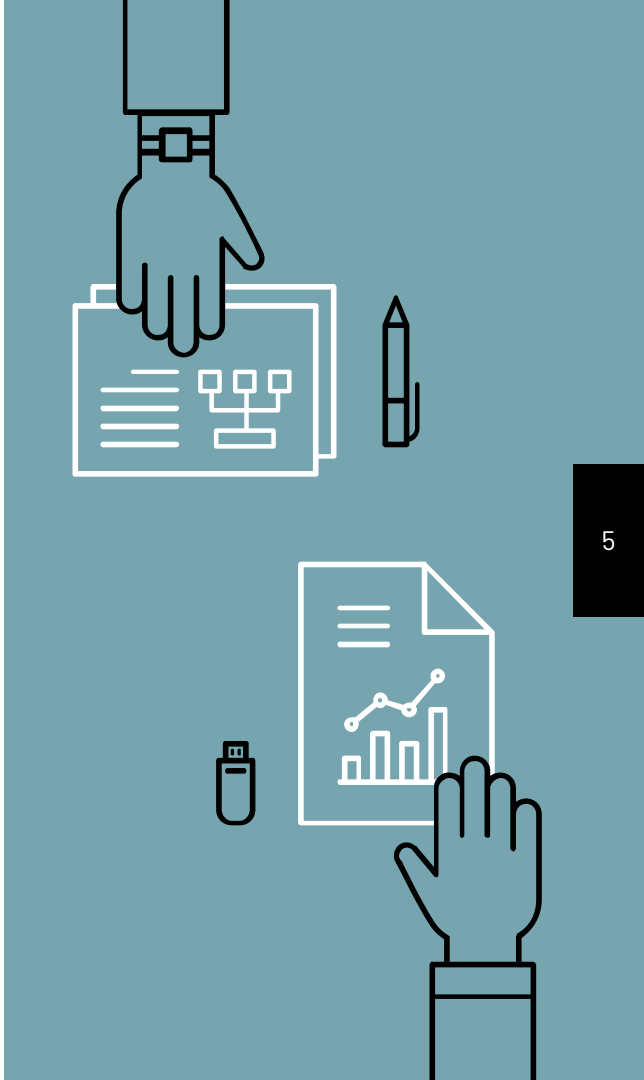
Having worked on creating public use datasets for clinical trials, I know that there are very few guidelines and best practices available from databases and funders like the NIH, especially when it comes to variable labeling, naming, and documentation. While the same guidelines may not work for all types of data (clinical trial data vs gene expression data, for example), it does seem like there are common best practices that would make developing and using public use data easier. For example, clearly marking in the name of a variable if it is raw data, if it has been obscured in some way to protect patient privacy, or if it is derived. Additionally, if the data has been used for manuscripts, the variables used in those manuscripts can be made available in the documentation to prevent those trying to reproduce published results from having to guess which variables to use.

John has a quote for everything – Yes. Yes he does :)

Module 3 Comments

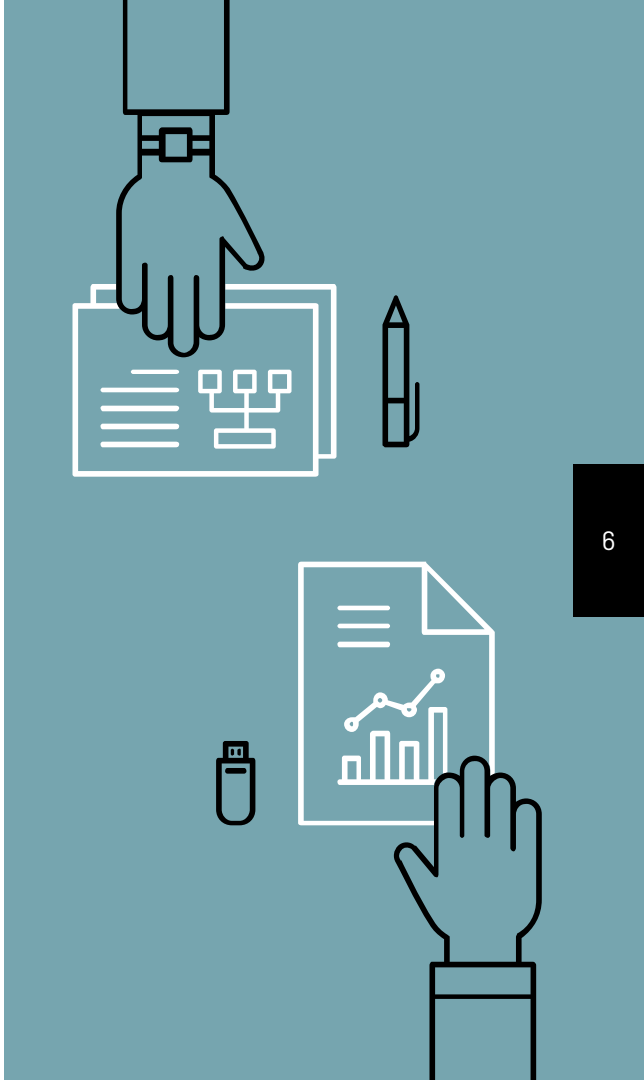
It's interesting that Ioannidis has been one of the proponents of data reproducibility, given that I know him as the author of one of the most infamously flawed early COVID papers (where he assumed that volunteers for COVID testing were representative of the general population and that there were something like 50 - 85 times as many cases as were reported). This is just an example that proper scientific work is not specific to individuals but that people can have varying degrees of scientific rigor, and that each piece of work needs to be judged on its own - not by who creates it

I like the idea of publishing all the decisions made along the way for the analysis process. It would be useful to know what modeling decisions people make that lead to failures and why certain modeling choices were made, but this information is rarely included in papers.



Module Discussion

- **How do you reconcile issues of data privacy with the call to make the data publicly available?**
  Having full, publicly available data is crucial to being able to reproduce results; however, this is often at odds with patient's rights to privacy. For example, it is common practice to collapse categories of variables to obscure small samples and to obscure times and dates to protect patient identities in publicly released clinical trial data. What discussion has there been about the tension between data privacy and having full data available to ensure reproducibility?
  What steps should one take to ensure a process is reproducible/replicable if the data itself can not be shared (perhaps it is too sensitive).
- There has been a ton of discussion about this and recently a major push for open datasets. [Check out Leo Celi and PLOS Digital Health].

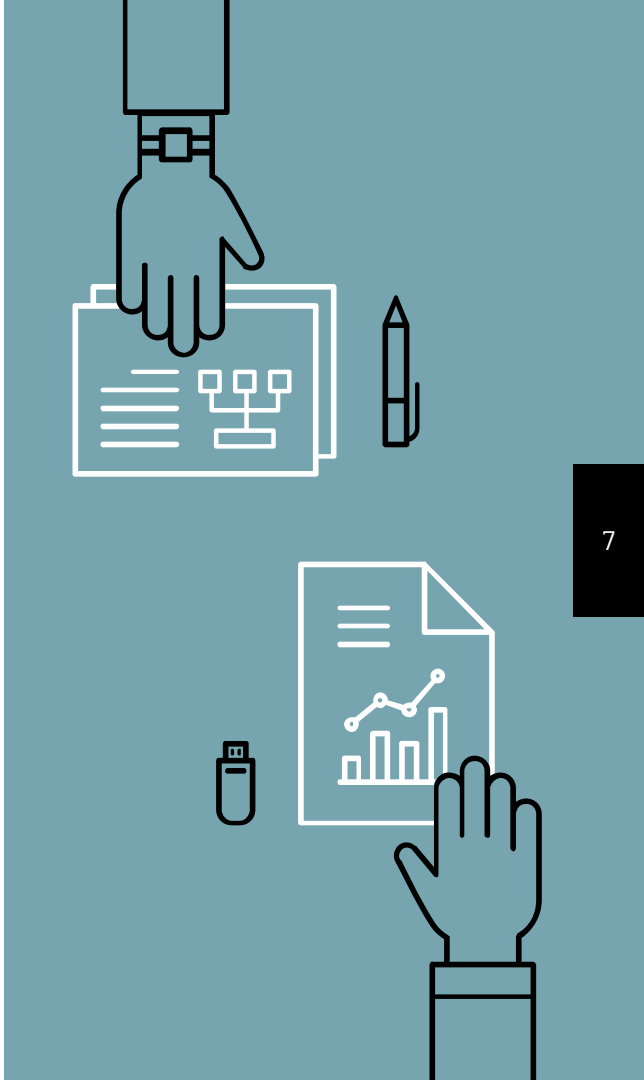


Module Discussion

Do authors generally report information on software and data for p-value calculation? – Depends on the journal but I have seen this more and more over the years.

Are smaller studies that collected their own data require to share their datasets for reproducibility purposes or does this generally apply to larger studies(i.e., genomics or national studies)? – I think it depends. Now, one way to be more competitive at receiving grant funding for the project is to make the data publicly available. The NIH has data sharing policies.

What are best practices if you're working in a collaborative setting on making sure that the raw unprocessed data remains available and up to date once a study is complete? It seems like there are a lot of nuances that could make this increasingly difficult such as people switching jobs or roles. – Thorough documentation and a clear protocol on who has access to the raw, unprocessed data.

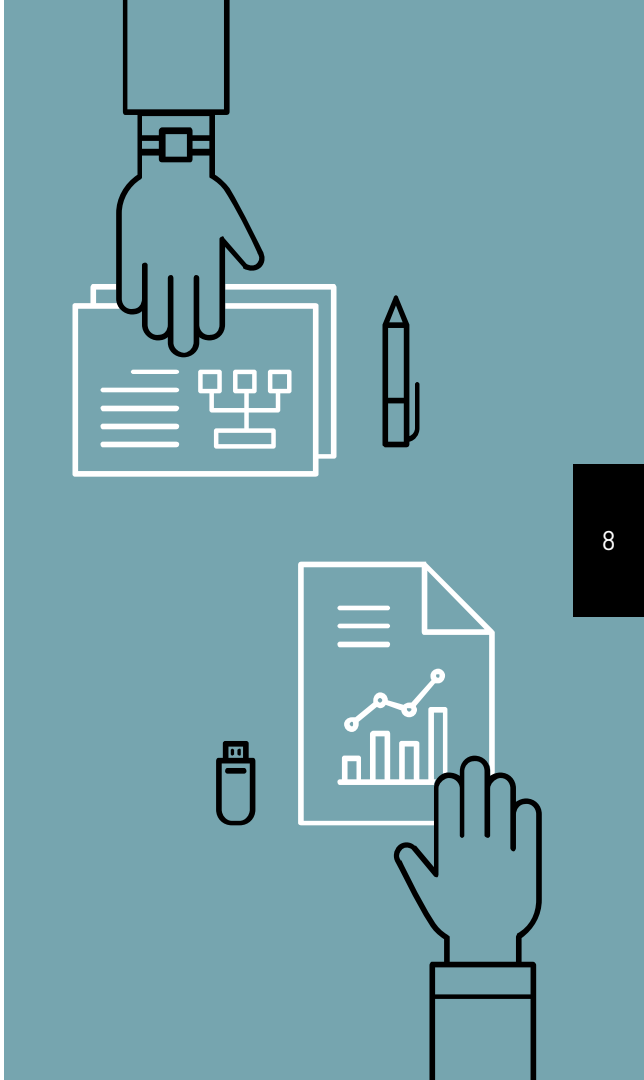


Module Discussion

How do you determine whether your methods section allows for reproducibility without being too convoluted? – Have others read it. Reviewers do catch a lot of things and most revisions are making something clearer to the reader.

How often do journals dismiss inconsistencies as "scientific differences of opinion"? – Great question! I have no idea. A lot of journals now have submissions called “Comments” that are written by other scientists in the field who have read a particular, usually somewhat controversial, paper, and are providing their viewpoint on the validity/scientific importance of the paper.

The Potti scandal really stressed the importance of reproducible data science: clinical trials were being conducted based on falsified results, and money was wasted, people's health was potentially harmed. I was wondering if there are any laws enforced to help prevent academic misconduct?
Yes - https://www-science-org.ezp-prod1.hul.harvard.edu/content/article/duke-university-settles-research-misconduct-lawsuit-1125-million



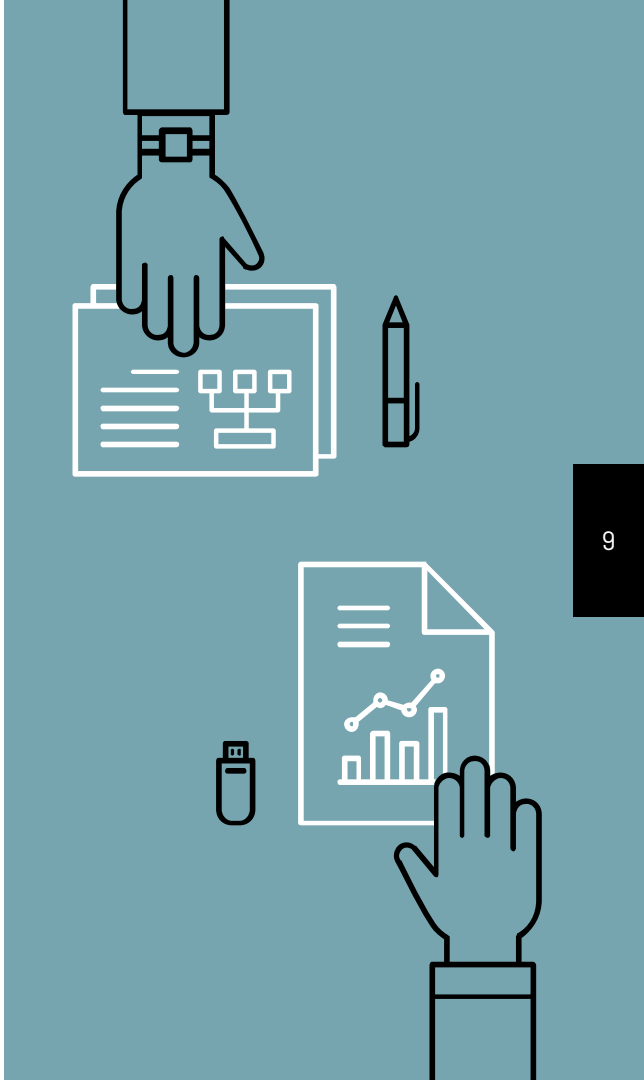
Module Discussion

Out of the four areas of focus for reproducibility: Authentication isn't very clear. Could you please explain what is meant by the line: "methods to ensure the identity and validity of key biological/chemical resources used". - Example: sex annotation. See these NIH guidelines.

I'm interested in learning more about what sort of incentives exist for performing extensive "forensic"-type analyses, such as the Baggerly and Coombes case discussed in the videos. – Not a lot to my knowledge. Doing something like this can help you realize how to make your own work reproducible and rigorous; if you do find misconduct it would help your academic career.

Do journals have requirements/recommendations for software beyond that it be open source? Such as code style, version control preferences, etc.? – At this point, I don't think so. It may change in the future, but software is constantly be updated/changing and varies by field so this would be very difficult.

Is it common to look at the reproducibility checklists journals provide prior to starting a project as a basis to follow? – Common? Probably not. But hopefully it's becoming more common with checklists like these.

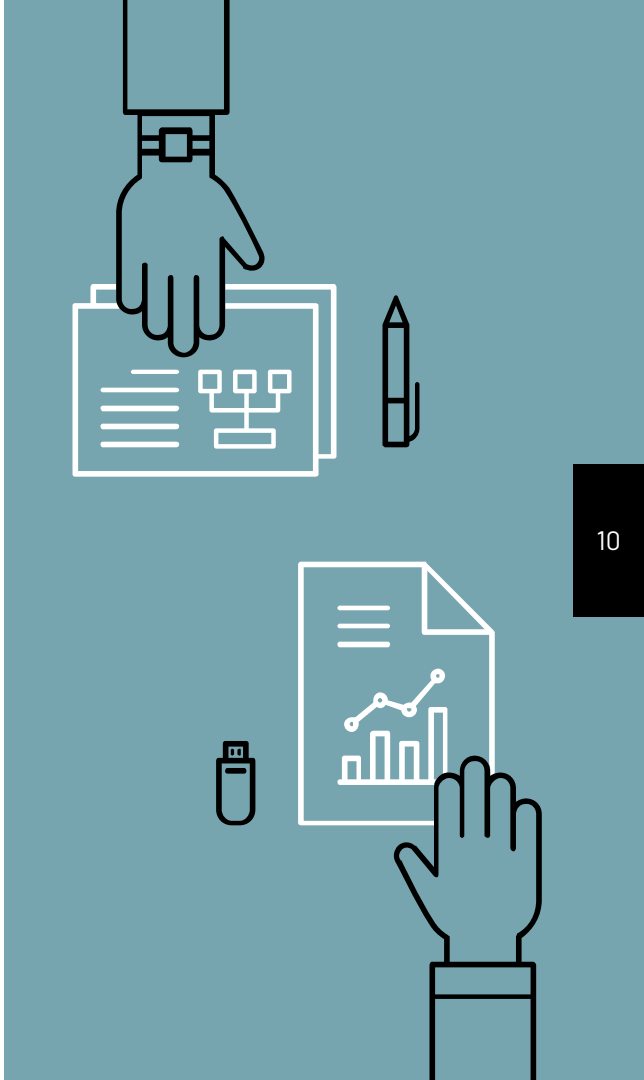


Module Discussion

Helpful articulation of the importance of discussing plans for resource sharing while writing grant applications. Do you have recommendations for a mini-course or workshop in our department or school on writing grants that include all of these elements? – Yes. We used to have a grant writing course in the department. I think your oral exam paper is required to be written in an F31 grant format.

What are some factors that prevent reproducibility standards from being applied at all journals? i.e. why do we still hear of instances of journals continuing the publication process for papers that have been shown to have non-reproducible results? – Time, money, need to get papers to publish.

In Video, 3.6.2, John describes a checklist to ensure that the journal articles are reproducible. One of the points in the checklist is, "Require authors to report replication strategy". How is replication strategy different from the methodology section of the paper. What extra information needs to be provided in replication strategies? - The replication strategy includes the steps you have taken to make your work reproducible.

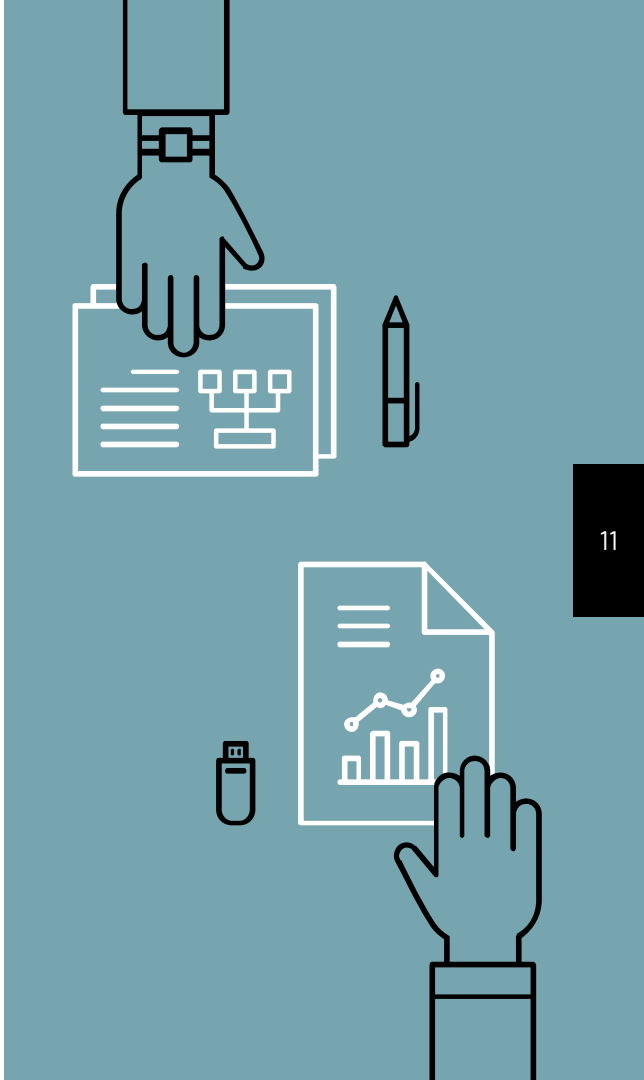


Module Discussion

Is pseudocode ever provided in supplements for reproducibility purposes? I feel like that's a good compromise between interpretability and conciseness. Yes – a lot of papers with new methods or algorithms provide pseudocode

It seems like quite a few of the common errors in manuscripts could be solved by training researchers in best practices in data management (variable naming conventions, data documentation, correct maintenance of records of data changes and provenance, etc). Do you see data management training becoming a standard part of statistics training? I do! I think it should be/will be added to training courses for biomedical data. At the moment, some PIs/advisors mentor students about this, but it is far from consistent and not all students receive this guidance.

Does the fact that so many studies are difficult or impossible to reproduce suggest that the veracity of the underlying science is called into question, or does it mean that the authors simply didn't take enough effort and time to make their work reproducible or does or did they make mistakes? Is there a way to tell if we can't reproduce the work?
How can we distinguish the failure of keeping reproducibility from academic misconduct?
How to efficiently verify that the data we use from others are reliable?



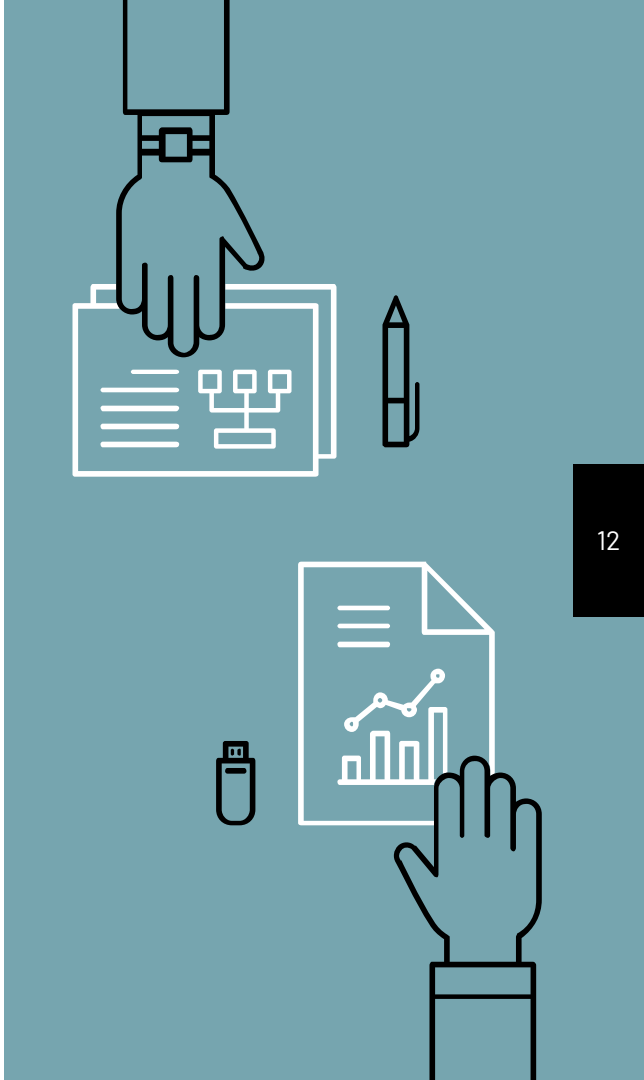
Module 3 Discussion

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?

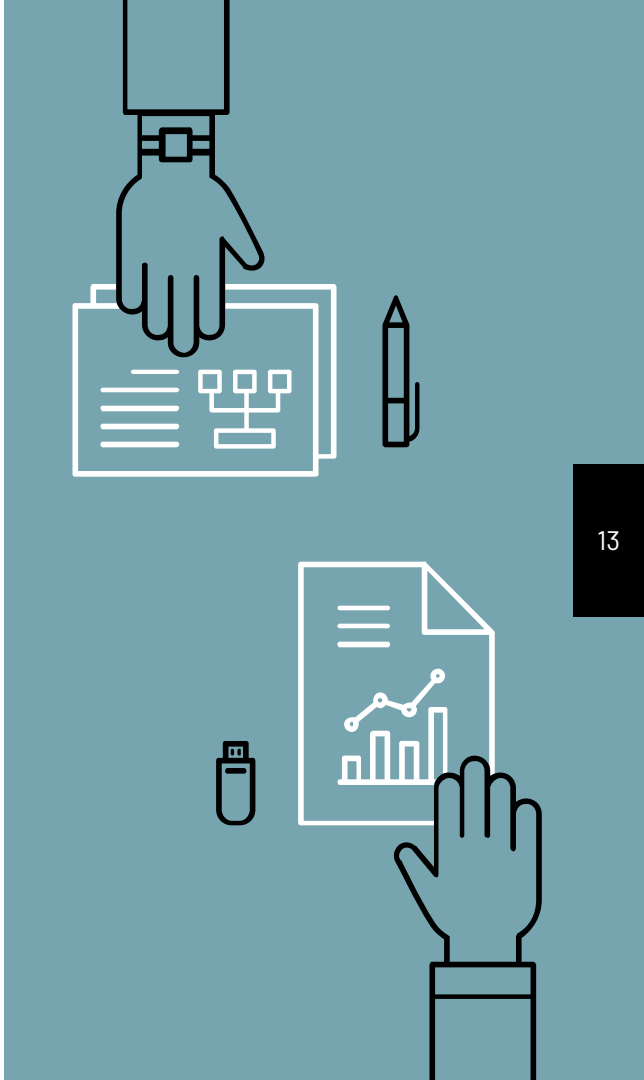
- Exploratory data analysis with plots, summary statistics, etc.
- Reading the documentation about how, when and by whom the data was collected/organized and how it has transformed at all (cleaning/wrangling)



Module 3 Discussion

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

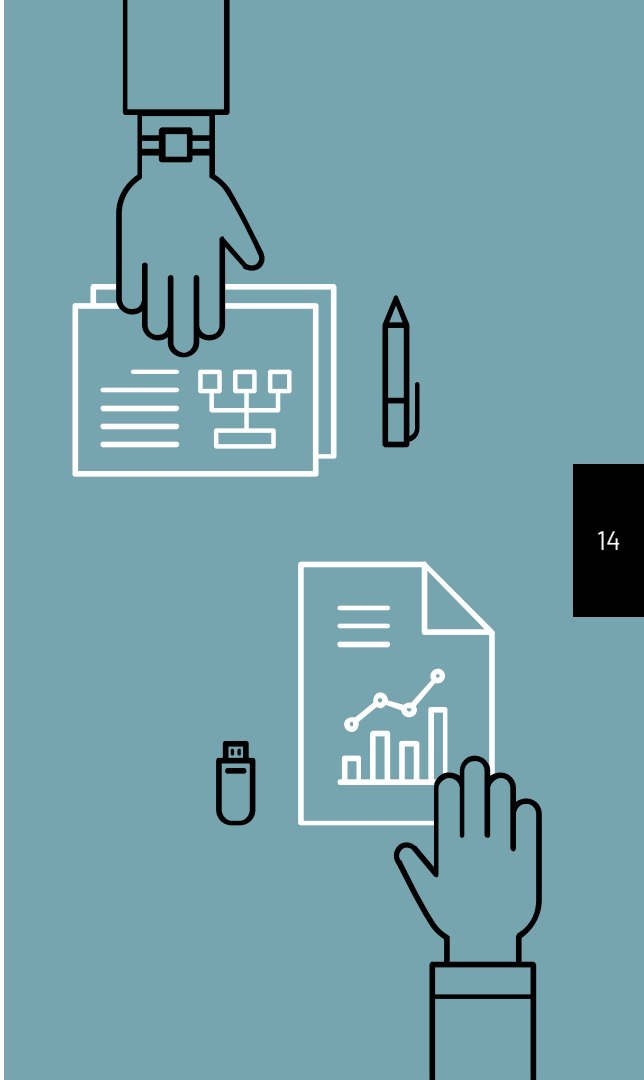
- The MIMIC Critical Care Database created and maintained by researchers at MIT is a great example
- They alter the data a bit to make it unidentifiable
  - Change an individual's date of birth to be a date in the future (e.g. 2100)
  - Change the dates of ICU admission
  - Add noise/perturbations to the measurements in a way that doesn't alter the conclusions reached from analyses
  - Group all individuals >90 years old into one group and not include their specific age
- 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?

- 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

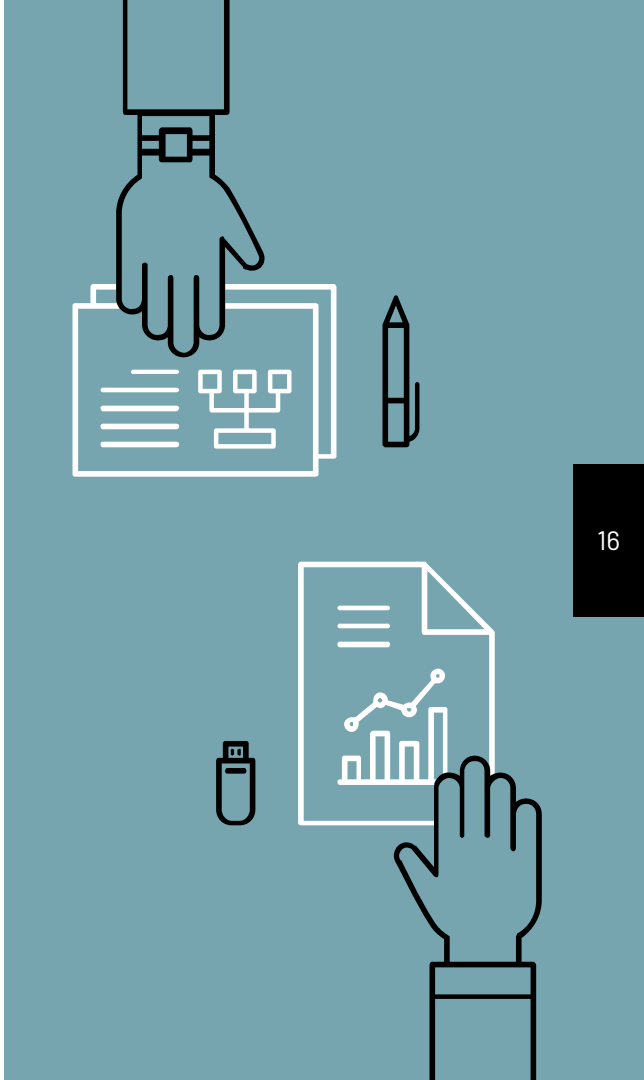
Yoshitaka Fujii, Japan	169
Joachim Boldt, Germany	96
Diederik Stapel, Netherlands	58
Chen-yuan Peter Chen, Taiwan	43
Yoshihiro Sato, Japan	43
Hua Zhong, China	41
Shigeaki Kato, Japan	39
James Hunton, United States	36
Hyung-in Moon, South Korea	35
Jan Hendrik Schön, United States	32

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

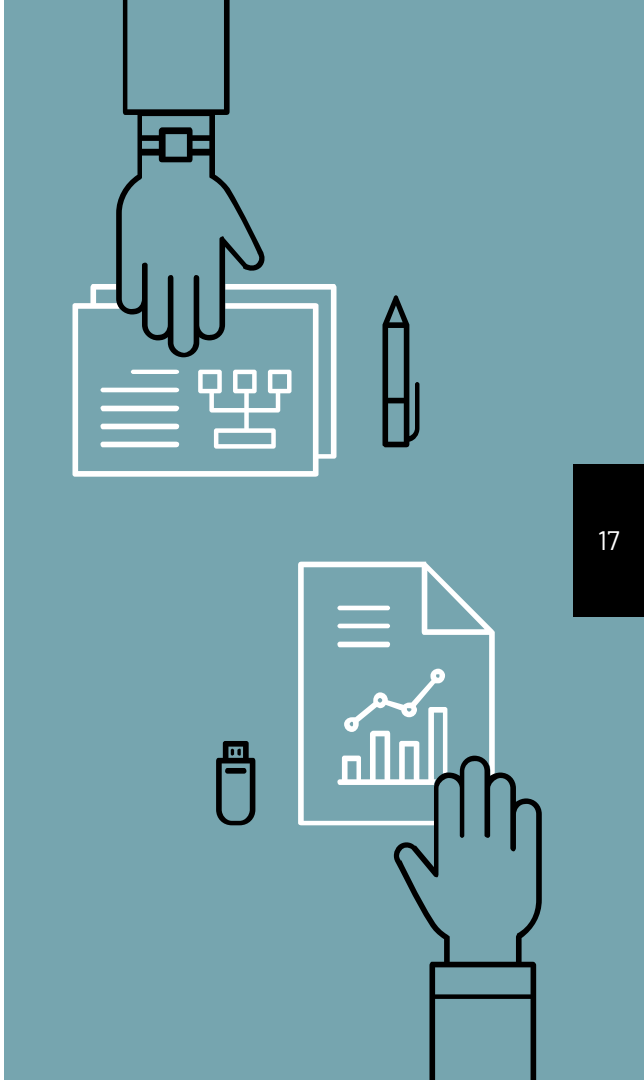
- I personally haven't used Sweave (it looked confusing to me 10 years ago) so I use Jupyter and knitr for documenting code I write

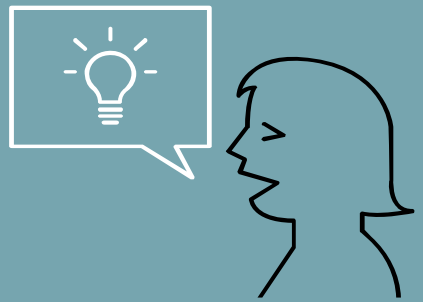


Module 3 Discussion

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





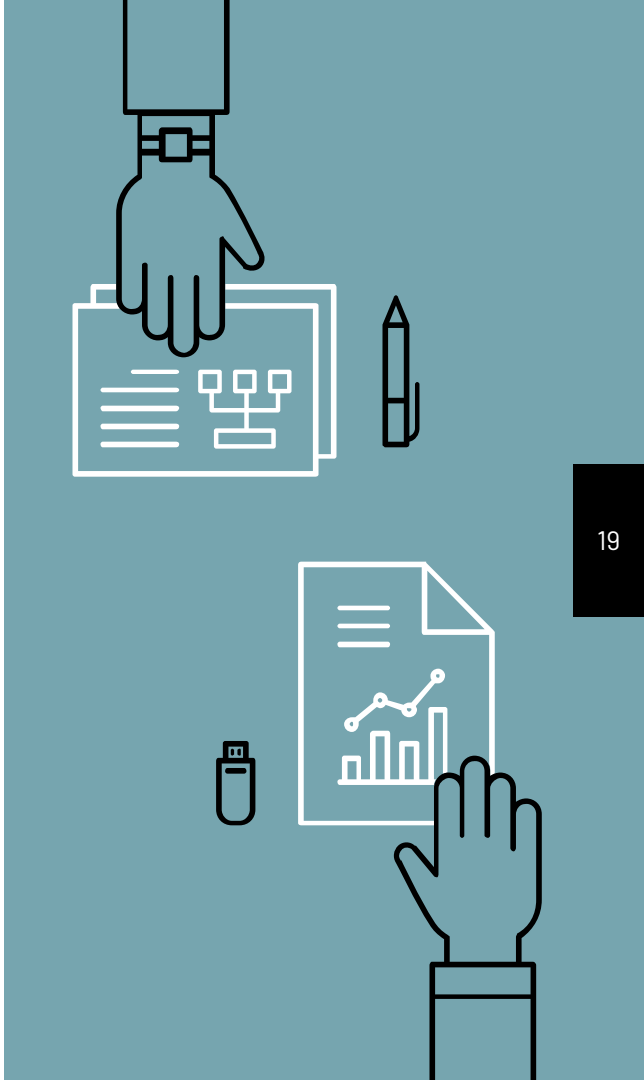
In-Class Project

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

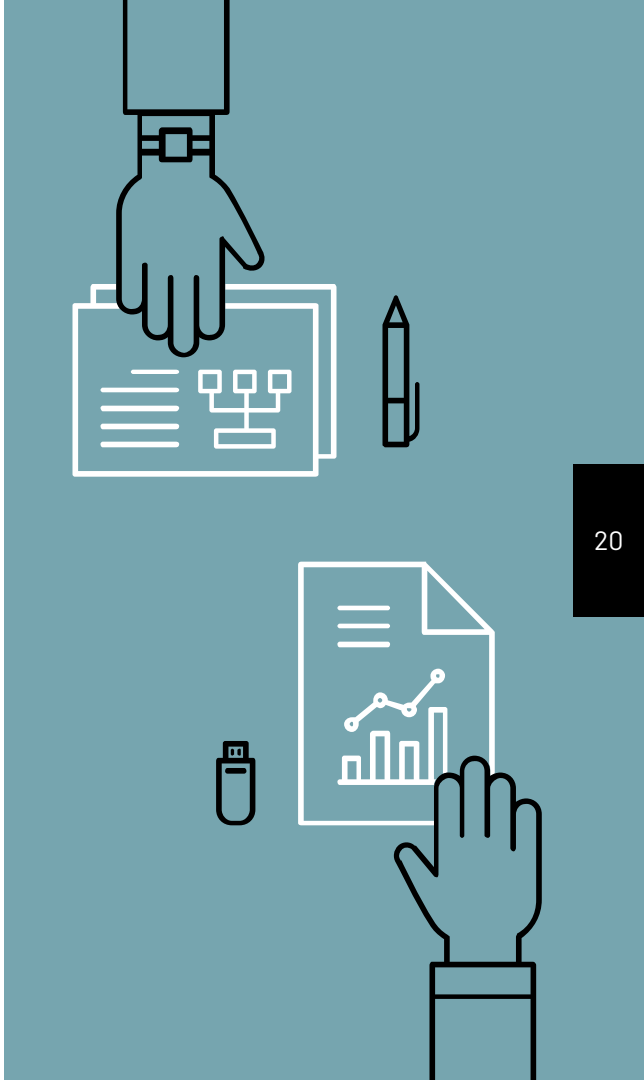


We will attempt to reproduce the results from and critique the reproducibility of:

[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.

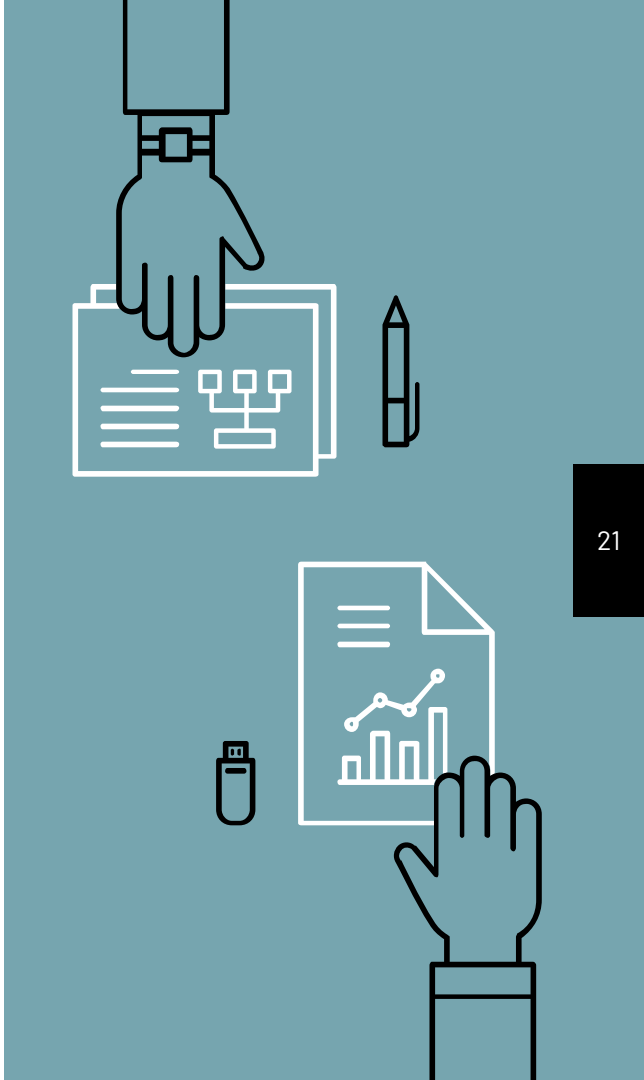
Specific Tasks:

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



Sample Size

- What did we get for a sample size? Does it match the paper?
- Why?



Homework

- Watch Module 4 videos
- Submit Module 4 discussion points