



## Week 3: Open Data and Data Sharing in Rehabilitation

Sook-Lei Liew, PhD, OTR/L  
Associate Professor & Director, Neural Plasticity and Neurorehabilitation Lab  
Chair, ENIGMA Stroke Recovery Working Group  
University of Southern California  
[sliew@usc.edu](mailto:sliew@usc.edu) | <https://chan.usc.edu/npn/>

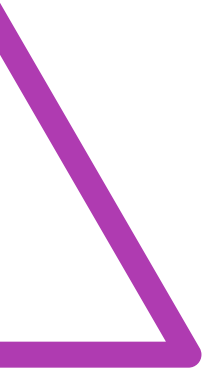


# This week's topics

- FAIR Data (Dr. David Kennedy, starring Dr. Maryann Martone)
- NIH's Data Sharing Plan: Part I (NIH) (Part II is optional!)
- Data Sharing and Open Data in Rehabilitation (Dr. Sook-Lei Liew)
- ReproRehabDB (Ms. Coralie Phanord)

Plus – **starting programming** in the environment of your choice (with your pod)!

(See our Youtube for R and Matlab streams, as well as a few very intro-to-python lectures (no full stream for this yet as it was not requested by this cohort))



# Refresher on reproducible science

## 1. What is the “reproducibility crisis”?

Do you think that scientific reproducibility and replicability is a problem in stroke research?

## 2. How can we use data science to address reproducibility? (last week)

## 3. How can we use open science to address replicability? (this week)



**USC** University of  
Southern California

# Potential solutions

## Methods (Reproducibility) → Data Science

- Underutilized reproducible methods:
  - Human error in manual processes (data entry, analysis)
  - Inconsistent keeping record across different team members

## Results (Replicability) → Big Data / Open Science

- Positive publication bias
- Logistical limitations:
  - Limited money, time, and participant availability can lead to biased and underpowered samples



# What can be done?

Results (Reliability) → **Big Data / Open Science**

- Overcoming positive publication bias and logistical limitations by testing samples from:
  - Retrospective datasets that have been archived
  - Pooled samples across retrospective/prospective datasets from **diverse** research sites (e.g., ENIGMA)
  - **Large** prospective datasets (e.g., UK Biobank)
- All of these would benefit from data science for accurate data management, analysis across sites



# Open Science: What is it?

- Open science movement: Sharing (published & **unpublished**) data, code, protocols, resources
- Why do it? To improve scientific reproducibility and replicability and build the capacity of the scientific community (especially trainees)
- What's involved? Usually free to download, with some agreement you won't abuse/sell the data.

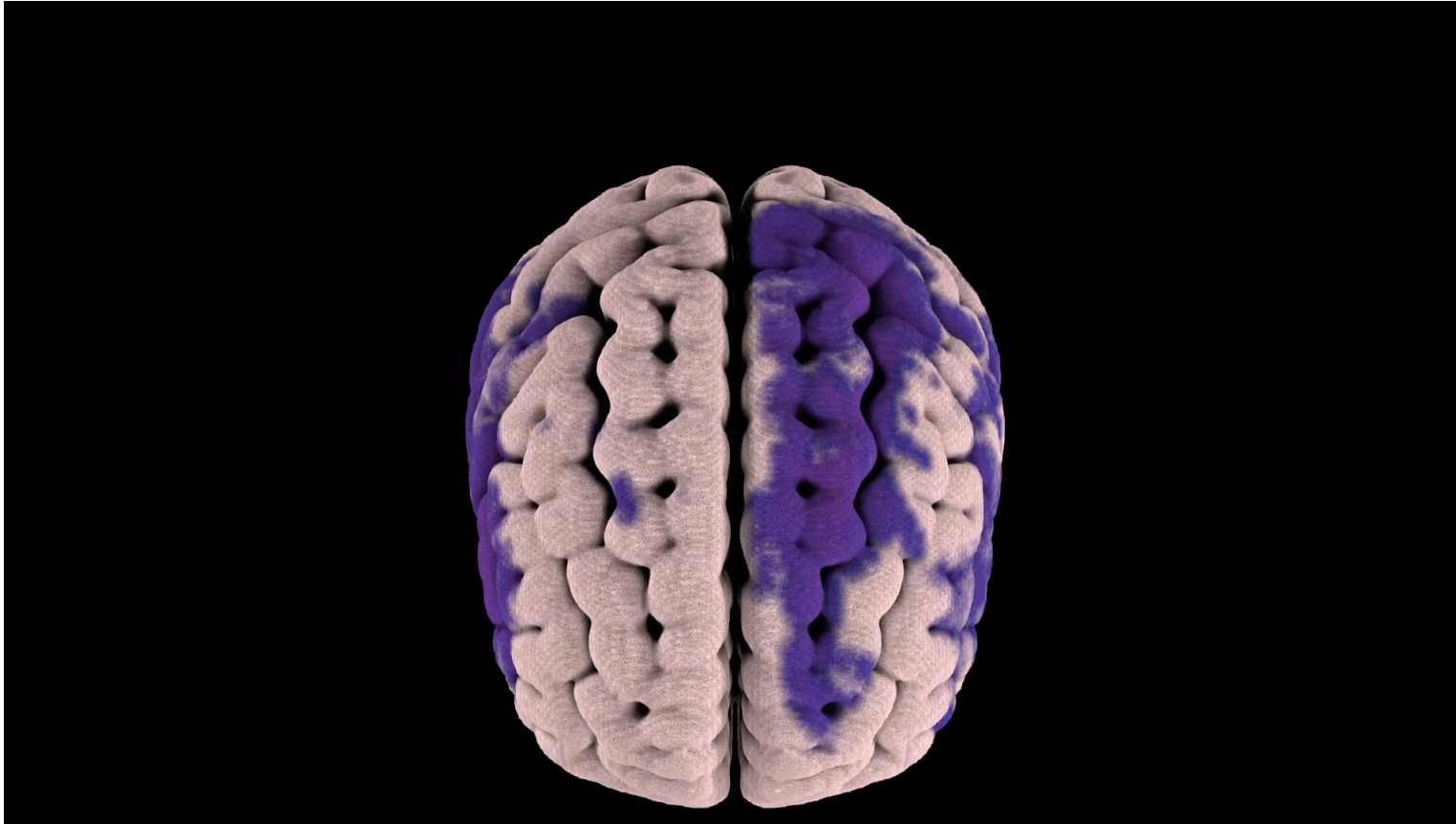


# Open-source data sharing to advance research

Anatomical Tracings of Lesions After Stroke (ATLAS) v2.0

N=1271 stroke T1-w high resolution MRIs and lesion masks

Liew et al., 2018, *Scientific Data*; Liew et al., 2022, *Scientific Data*



[http://fcon\\_1000.projects.nitrc.org/indi/retro/atlas.html](http://fcon_1000.projects.nitrc.org/indi/retro/atlas.html)  
<https://atlas.grand-challenge.org/>



**USC** University of  
Southern California

# Open Data: What types of [rehab] data are there?

- Data types: Surveys, behavioral measures, demographics, kinematic data, videos, physiological data (e.g., brain imaging)
- Prospective data collections (protocol is set prior to data collection)
- Retrospective data archives (usually study-specific data)
- Health services / medical records / observational data





# Open Data: What types of data are there?

- **Rehabilitation-Related Data Archives (NCMRR-funded)**

- CLDR: <https://www.utmb.edu/cldr/>
  - Center for Large Data Research and Data Sharing in Rehabilitation
  - Many types including health services research (e.g., medical records) and retrospective study-specific rehabilitation data
- ICPSR/ADDEP: <https://www.icpsr.umich.edu/web/pages/ADDEP/index.html>
  - Archive of Data on Disability to Enable Policy and research
  - Retrospective study-specific rehabilitation data
- OpenSim: <https://opensim.stanford.edu/>
  - Free motion simulation toolbox and trained models for different populations:  
<http://simtk.org/>
- NIH: <https://sharing.nih.gov/> - look for where to share your data



# Open Data: What types of data are there?

- **Prospective/Coordinated Brain Imaging, Clinical/Behavior**

- Human Connectome Project: <https://www.humanconnectome.org/>
  - Lifespan, young adult, clinical populations, with harmonized behavior
- UK Biobank: <https://www.ukbiobank.ac.uk/>
  - UK health records data including brain imaging, genetics, clinical variables
  - Working up to 100,000 individuals
- All of Us: <https://allofus.nih.gov/>
  - On beta release; will be US health records data including brain imaging, genetics, clinical variables and questionnaires
  - Working up to 1 million individuals



# Open Data: What types of data are there?

- **Community (Study-Specific) Brain Imaging**

- Open Neuro: <https://openneuro.org/>
  - 372 MRI, MEG, EEG, ECoG datasets
- INDI: [http://fcon\\_1000.projects.nitrc.org/](http://fcon_1000.projects.nitrc.org/)
  - International Neuroimaging Data-Sharing Initiative: Prospective and retrospective data
  - Resting state fMRI, structural MRI, diffusion MRI with behavioral measures
- NITRC: <https://www.nitrc.org/>
  - Neuroimaging Tools and Resources Collaboratory: Atlases, data, and tons of software/tools



# Open Data: What types of data are there?

**ReproRehabDB** (see short video by Coralie Phanord!)

<https://reprorehabdb.usc.edu/>

- Allows you to:
  - Search for data science courses OR rehabilitation-related datasets
  - Upload new courses or datasets (please upload!)
  - Rate and review courses and datasets (and see others' ratings)!
- Beta (in development) - if bugs, contact [reprorehab@gmail.com](mailto:reprorehab@gmail.com)
- Huge shoutout to Coralie Phanord, Sanying Yi, Swapnil Arya, and Paul Bailey for creating this!

# Open Data: But I want something more specific?

- If you have a specific need, you may consider reaching out to someone who has published a dataset that you'd like to utilize
- General guidelines:
  - Collaborate on the data (including authorship)
  - Receive useful insight on the data wrt how you use it
  - No one's data is perfect!
  - Maybe help organize their data into a data archive that you both can also publish (see journals like *Scientific Data*, *GigaScience*) or cite



# Data Sharing: I want [need] to share data

- Everyone should think now about data sharing (before you even start a study)
- NIH's new Data Management and Sharing Policy goes into effect January 2023 – all grants will need a DMS plan, which is great practice for any study
- <https://sharing.nih.gov/>
- Many great principles shared in Dr. Martone's FAIR data and the NIH data sharing videos

# Sharing rehabilitation data – human subjects research

Ensure your informed consent form and IRB protocol has language for data sharing now!

- Ask your IRB about any data archiving policies
- Opt-in/opt-out of data sharing, or a blanket policy
- Sample wording for consent and/or protocol:

*DATA STORAGE AND RETENTION: Research data will be maintained in paper format in a secure location at USC or electronically on secure, password-protected computers and servers. Only authorized individuals will have access to it, and all electronic data will be de-identified. **The researchers intend to keep the de-identified research data indefinitely. Other researchers may have access to the de-identified data for future research, and the de-identified data may be included in future repositories or archives for use by other researchers. Any data shared with other researchers will not include your name or other personal identifying information.***

# Sharing rehabilitation data – data transfer agreements

- Ideal to share data under open licenses (anyone can access) such as a CC0 license
- However, check with your university/institute to see what they allow.
- Some universities/institutes require sharing data via data transfer agreements, specific licenses, etc. – contact your tech transfer office
- If data can be identified, you may need a more stringent data use agreement
- See OpenNeuro FAQ for examples: <https://openneuro.org/faq>



# Sharing rehabilitation data – good data management

Ensure you have good data management practices in place now!

- Systematic, machine-readable file naming/conventions
  - No spaces or special symbols, consistent naming
- Keeping track of meta-data as you go
  - Noting units, abbreviations, etc. Have an outside colleague try to use your data
- Have a system for data analysis and version control

# Sharing rehabilitation data – what to share

- Raw, preprocessed and/or processed data
- Individual subject data
- NULL DATA! Unpublished data!
- Group-level data
- In addition to open data, consider sharing your code/software
  - Github
    - Allows not only for data sharing but for collaboration – people can pull and improve or modify your code, all version-controlled
    - People can also report issues, which you can respond to in real-time
  - Open Science Framework

# Sharing rehabilitation data – publishing a paper about your data

Many journals are emerging with data descriptor formats:

- *Scientific Data*, a Nature journal
- *GigaScience*
- *Data, Data in Brief*
- See also Walters et al., 2020:  
<https://insights.uksg.org/articles/10.1629/uksg.510/>
- Publishing your data allows it to be cited, but also ensures you provide the necessary meta-data for accessing your dataset

# Sharing rehabilitation data – associated costs

- Someone to scrub/collate the data, generate meta-data
  - Ideally, you are doing this from the beginning (see Data Management!) so costs should be quite low
  - Requires someone who is familiar with the data (hard to hire an outside person to do this)
  - It always takes longer than you think! (Again have someone naïve check it)
- Cost for archives (\$0-\$10,000?) – reach out to archive!
  - Someone to go through your data and meta-data and reformat as needed
  - Web storage and maintenance
  - Facilitating requests and troubleshooting for downloads
- Article publishing fee for related data paper (\$500-3500)

# Sharing rehabilitation data – timeline

- Start the process early (ideally before you start collecting data!)
- Identify all of the necessary requirements from your institute
  - IRB, consent, data transfer, license, decide what you will likely share (raw, preprocessed, processed, group-level, individual-level)
- Upon completion of data collection, organize the raw data
- Upon completion of data analysis, organize the data analysis pipeline (for code sharing) + data outputs
- By the time you write the paper, your data should be ready to go!

# In summary

- Open data is a powerful tool for analyses and overcoming positive publication biases and improving replicability in the field
- Data sharing is mandated by the NIH, AND it's just general good practice 😊
- It doesn't have to be painful and hard, especially if you start early!
- And, data often takes on a life of its own that contributes to science above and beyond your immediate intentions 😊



# Thank you!

## Visit us at <http://npnl.usc.edu>



### Contact Us:

Me: [sliew@usc.edu](mailto:sliew@usc.edu)

Lab: [npnl@usc.edu](mailto:npnl@usc.edu)

Twitter: [@NPNLatUSC](https://twitter.com/NPNLatUSC)

### Special thanks to:

- ReproRehab Team
  - Coralie Phanord
  - Grace Song
  - ENIGMA Stroke Recovery Team
- 
- NIH R25
  - NIH R01 NS115845
  - NIH K01 091283



[enigma.ini.usc.edu](http://enigma.ini.usc.edu)