Poster #2025

Surveying data management practices and perceptions among neuroimaging researchers

Ana Van Gulick, Carnegie Mellon University, Pittsburgh, PA, USA

John Borghi, Stanford University, Stanford, CA, USA

Carnegie Mellon University
Center for the Neural Basis of Cognition
Carnegie Mellon University & University of Pittsburgh

Stanford University

QUESTIONS

How do neuroimaging researchers manage their data?

How do they plan for data management, store and document data during collection and analysis, and share their data and code?

What methods do they use for collection, analysis, and sharing of data and how consistent are those within a research group?

What are their perceptions of emerging open science and scholarly communications practice including open access publishing, data sharing, and preregistration?

Research Data Management

Research Data Management (RDM) includes the documentation, file naming, storage, security, backup, publication, sharing, and preservation of research materials including raw data, analyzed data, code, documentation of the data collection and analysis procedures, and any other research materials such as stimuli or other measures. RDM encompasses the entire research lifecycle from planning through collection, analysis, and publication and sharing of results.

Neuroimaging as Case Study

Neuroimaging research presents an ideal case study for assessing RDM practices and perceptions for several reasons:
(1) The datasets are large and complex, often containing information collected in a wide variety of forms and file formats.
(2) Analytical pipelines are highly iterative and rely on an array of software packages and custom code. Decisions have significant downstream effects.
(3) The field is currently grappling with concerns about its methodological and statistical practices, prompting a growing interest in open science, reproducibility, and data sharing.

Maturity Model

RDM maturity was defined as the extent to which data management practices are clearly defined, implemented, and (if applicable) optimized

Examples of maturity levels in Cognitive Neuroscience MRI research

	Ad hoc			Refined
Plan	There is a plan for transferring data from scanner to the lab. There is a plan for backing up the behavioral, imaging, and questionnaire data.	Data types (images, behavioral data, questionnaires, etc.) and outcomes are defined in advance. Task/design specific parameters are defined.	Roles and responsibilities around data are defined. "Data Sharing" is included in IRB documents, task plan, and forms (e.g. consent form).	Data Management Plan is revisited throughout project lifecycle. Experiment is pre-registered.
Collect	Behavioral, imaging, and task-related data are stored someplace. Data is backed up. Code is preserved.	Meta data is managed. [Incidental findings, motion, sleepiness, runs completed, etc.]	Naming conventions and file structure are standardized within lab/collaboration	File structure is kept to community standard (e.g. BIDS)
Analyze	Workflow parameters (e.g. pre-processing/analysis parameters, etc) are recorded [lab notebook, wiki, etc]	Analytical output (code, batch files, etc.) are stored.	Analyses and code are subject to version control/workflow management (e.g. OSF, Flywheel, Git, XNAT, LORIS, COINS).	Code is public and annotated.
Output	Data is described in a publication according to community best practices. [Table of peak coordinates that can be scraped by NeuroSynth, BrainMap, etc.]	Data is described in a publication that includes a data availability statement [Data is available to be sent to people who ask for it]	Data is deposited in subject-specific repository. [Unthresholded maps: Neurovault; Raw Data: OpenfMRI, Raw + Preprocessed Data: HCP]	Code and data are available in a public repository with persistent identifiers (e.g. DOI)

METHODS

SURVEY

- 74 multiple choice questions ordered roughly according the the phases of a typical research project including ratings of practices maturity from 1 (ad-hoc) to 5 (mature).
- Designed in consultation with MRI researchers and tailored to specific terminology, practices, and tools of the community.
- Modeled on the structure of the Data Curation Profiles and employing a capability maturity model framework.

Questions focused on a range of RDM topics including:

- Types of data collected
- Tools used to analyze and manage data
- Standardization of RDM practices within a lab group
- Factors that motivate and limit RDM
- Emerging scholarly communication practices including publishing pre-prints, datasets, and other research products

PARTICIPANTS

A total of 144 participants from 11 countries and 69 institutions participated in the survey.

Position	Percent	Research Area	Percent	Funding	Percent
Assistant Professor	21.5	Behavioral Neuroscience	2.1	National Institue of Health (USA)	64.1
Associate Professor	10.4	Bio/Neuroinformatics	1.4	National Science Foundation (USA)	11.7
Graduate Student	24.3	Clinical Neuroscience	16	Department of Defense (USA)	3.9
Post-Doc	21.5	Cognitive Neuroscience	55.6	Startup Funds	5.5
Professor	9	Computational Neuroscience	2.1	Institutional Funds	6.3
Research Assistant	2.1	Developmental Neuroscience	5.6	Foundation Grants	12.5
Research Associate/Scientist	6.9	Sensory Systems Neuroscience	1.4	International Grants	19.5
Research Technician	1.4	Social Neuroscience	5.6		
Staff Scientist	1.4	Affective Neuroscience	1.4		
Other	1.4	MRI Methods	2.1		
		Other [Please specify]	6.9		

RESULTS

RDM Limits & Motivations

		Data Collection	Data Analysis	Data Sharing
Limits	Time	69.6	71.3	79.5
	Cost	17.6	8.7	22.3
	Training	32.8	40.9	41.1
	Best Practice	43.2	48.7	49.1
	Incentives	36.8	32.2	37.5
	Other	7.2	6.1	5.4
Motivations	Access	76.8	73.3	70.5
	Prevent loss	100.0	85.8	78.6
	Publisher/Funder Mandates	35.2	28.3	42.0
	Institutional data policy	52.0	39.2	47.3
	Availability of tools	12.0	9.2	8.9
	Openness and reproducibility	63.2	64.2	67.0
	Other	3.2	3.3	0.0

Data Types

MRI Data	Percent	Non-MRI Data	Percent	Study Information	Percent
Anatomical	99.2	Demographics	97.0	Session info	90.2
Task	98.5	Clinical	60.6	Acquisition parameters	97.7
Resting	80.3	Behavioral	95.5	Task Information	97.7
Diffusion	59.8	Questionnaires	88.6	Stimuli	91.7
Field Map	64.4	Other Imaging	25.0	Code (presentation)	82.6
Other	11.4	Physiological Data	43.2	Code (data collection)	71.2
		Genetic	30.3	Other	7.6
		Eye Tracking	28.0		
		Other	3.8		

Software Types

MRI-specific Software (top 10)		Non-MRI-specific Software	
Software	**Percent**	**Software**	**Percent**
SPM	71.7	Matlab	83.3
FSL	70.8	r	70.0
Freesurfer	50.0	Excel	60.0
AFNI	45.0	SPSS	51.7
MRICro/MRICron	45.0	Python	48.3
Mango	13.3	SAS	5.8
CONN	12.5	JASP	5.0
OsiriX	10.8	Mathematica	0.8
BrainVoyager	6.7	Stata	0.0
Caret	6.7	Other	5.8
Other (see note)			

Many types of data and software are used.

Lack of workflow consistency even within a lab group.

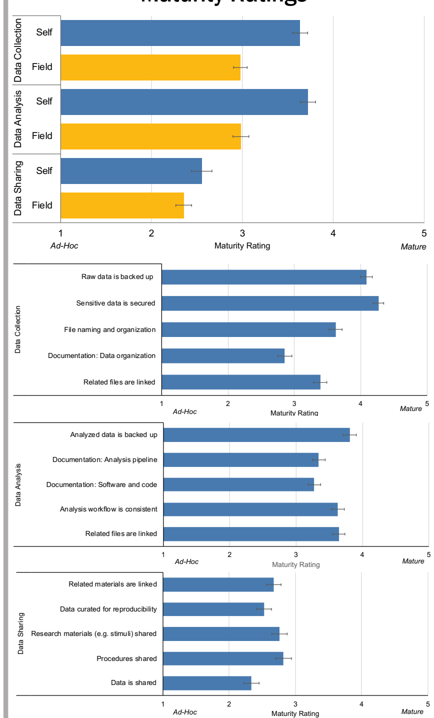
Survey respondents believe their RDM practices are more mature than the field (and they may be).

Open Materials

Survey: https://doi.org/10.1184/R1/5845212.v1

Data: https://doi.org/10.1184/R1/5845656.v1

Maturity Ratings



Long Term Preservation

What Data should be preserved long term?	Percent
Raw MRI data	97.4
Behavioral data (e.g. response accuracy, reaction times)	94.8
Demographic data (e.g. age, gender)	90.5
Task-related stimuli (e.g. images, audio/visual files)	90.5
MRI acquisition parameters (e.g. TR, TE)	88.8
Code used for stimuli presentation	84.5
Questionnaire data	81.9
Code used for data collection	81.0
Data about the scan session (e.g. movement, crash)	74.1
Task-related information (e.g. timing parameters)	74.1
Analyzed MRI data (e.g. contrast maps)	73.3
Clinical or Medical data (Including mental health information)	62.9
Physiological data (e.g. heart rate, blood pressure)	38.8
Eye tracking/pupillometry data	30.2
Genetic/molecular data (e.g. blood samples, cheek swabs)	29.3
Other neuroimaging data (e.g. EEG, NIRS)	26.7
Other	76.7

RDM and data sharing are both limited by time and cost.

Motivated by preventing loss and providing access to data.

Data is backed-up but not curated.

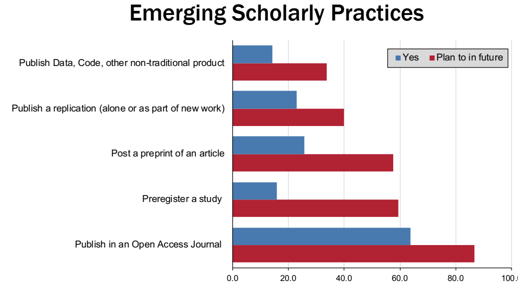
Barriers to Data Sharing

Reason	Percent
More to publish	50.4
Sensitive Info	30.4
Time	25.2
Supervisor	16.5
Difficulty	15.7
Other	14.8
Require Citation	9.6
Require Authorship	8.7
Intellectual Property	1.7
Don't know how	0.0

Read our Preprint

Borghi, J. A., & Van Gulick, A. E. (2018). Data management and sharing in neuroimaging: Practices and perceptions of MRI researchers. BioRxiv. https://doi.org/10.1101/266627

Emerging Scholarly Practices



CONCLUSIONS

- Neuroimaging is a highly complex space for data curation. There is a large variety of data types and formats, software, and tools being used and a lack of consistency across the field and even within research groups.
- Participants rated their RDM practices as more mature than the field as a whole, which is likely given our sample. RDM is most ad-hoc for data sharing due to lack of consensus in the field.
- There was a significant difference in how trainees vs. faculty saw the standardization of their practices.
- Both RDM and open science practices were limited by the amount of time these activities take and by the lack of training, community standards, and professional incentives.
- Perceptions of new scholarly communications practices including data sharing, data reuse, and Open Access publishing were largely positive but demonstrated little adoption of these activities into practice.

Next Steps

- Evolving space: Data sharing will be increasingly required, consensus is growing within the field about standards for analysis and documentation, technology is evolving to make documentation of workflows easier.
- Needs: More training in RDM, tools that link together across the research lifecycle, support for data storage, and more robust standards for data documentation, processing, and sharing.
- Collaboration with librarians on RDM training, standards for data documentation, and making datasets publicly available and reuseable.

References

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., … Yeo, B. T. T. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience*, *20*(3), 299–303. http://doi.org/10.1038/nn.4500

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. V. (1993). Capability maturity model, version 1.1. *IEEE Software*, *10*(4), 18–27. http://doi.org/10.1109/52.219617

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., … Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience, 18*(2), 115–126. http://doi.org/10.1038/nrn.2016.167

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *International Journal of Digital Curation*, *4*(3), 93–103. http://doi.org/10.2218/ijdc.v4i3.117