Data Science and Sharing Team
Board of Scientific Counselors Review

December 2023

Contents

Executive Summary of the DSST BSC Report

- Our group was envisioned by Drs. Susan Amara and Peter Bandettini in 2016. They believed that both the NIMH IRP as well as the mental health research community as a whole would benefit from a group dedicated to advancing the creation, distribution, and re-use of shared, open datasets in order to accelerate discovery.

- Among our goals is to effect culture change with respect to sharing data and code and instilling other open science practices that enhance reproducibility and collaboration.

- These goals are shared and emphasized by NIH leadership, the White House, and funders throughout the world [1, 2, 3].

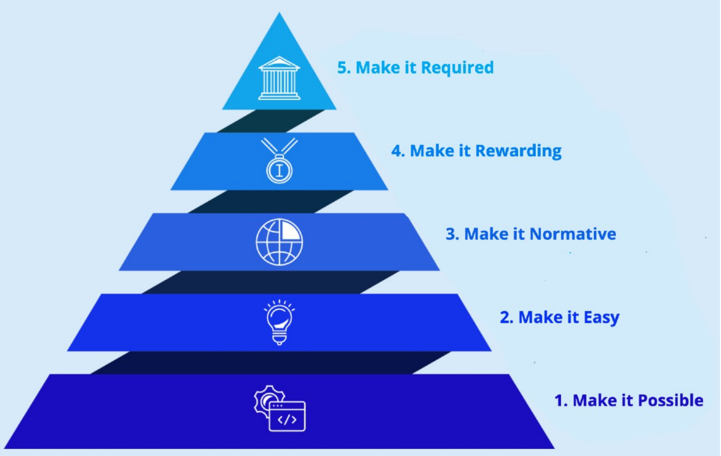

Figure 1: We use the Center for Open Science's "Pyramid of Social Change" [4] as a model to guide our strategy to effect culture change in the research community towards data sharing and transparency.

- Section 7.1 contains a list of hyperlinked citations to 75 publications, preprints, software packages, funded grants, and other deliverables produced by DSST staff in the past five years. They related to the five levels of the pyramid as follows:

  Possible: Nine of these references, including two funded grants [5, 6], relate to projects aimed at the bottom most level: making data sharing possible by expanding the space of data repositories and data standards. See Section 1 for details and highlights of these projects.

  Easy: 53 of these references, including one funded grant [7] and the 26 shared datasets listed in Section 7.2, relate to projects aimed at the next highest level: making data sharing easier by providing open science tools, training, and services to researchers in the IRP and the wider community. See Section 2 for details and highlights.

  Normative: 39 of these references relate to projects aimed at the middle level: making data sharing normative by providing examples of high-quality science that employs shared data and open science principles. See Section 3 for details and highlights of these projects.

  Rewarding: We believe the next highest level, making data sharing rewarding, has the greatest potential for growth. Our work aims to improve the available metrics for NIMH leadership to identify, acknowledge and reward investigators who are actively sharing and reporting shared datasets in their publications. See Section 4 for details and highlights of our projects in this area.

  Required: Finally, the DSST helped ensure that all 49 NIMH IRP Data Management and Sharing (DMS) plans complied with NIH's new DMS policy that makes data sharing required for all NIH funded research [8]. See Section 5 for details and highlights of our efforts to support the IRP in meeting these new requirements.

- Members of the DSST have received four separate **NIMH Directors Awards** since 2019 [9].

- In addition to the three grants mentioned above, Adam Thomas, Dylan Nielson, and Dustin Moraczewski were each awarded NIH Loan Repayment awards in 2016, 2019, and 2023, respectively.

- The six former members and trainees of the DSST have gone on to successful careers in government or industry and to graduate or medical school at prestigious academic institutions, respectively. See Section 7.1 for a complete list.

- All of the above have been accomplished with a budget that is among the lowest of any NIMH core, service, or team listed in NIH RePORTER [10]

Introduction

The DSST, envisioned by Drs. Susan Amara and Peter Bandettini in 2016, is dedicated to advancing the creation, distribution, and re-use of shared, open datasets in order to accelerate discovery. Subsequent years have demonstrated the prescience of this vision. Since 2016, there has been a growing recognition that lack of transparency represents a significant barrier to progress in biomedical science [11, 12]. Changing the culture and practice with respect to transparency and data sharing is now a major priority for NIH, the White House, as well as funders and policy makers throughout the world [1, 2, 3].

The mission of the DSST is to provide tools and training to help scientists within the IRP embrace open and reproducible science practices, including:

- Community-recognized data standards and formats
- Tools and procedures for making data processing and analysis code collaborative, version-controlled, and reproducible
- Timely deposition of all raw data, analysis code, and other methodological details required for reproducibility

The DSST's mission is quite different than other cores and support groups within the NIH IRP. In addition to providing cutting edge tools and techniques, we are working to change cultural and behavioral norms with respect to data sharing and transparency. Surveys have shown that the majority of scientists believe that data sharing is important and should be incentivized [8]. However, the current social norms within the research community provide relatively few incentives to expend limited and valuable resources on making raw data publicly available. Those same scientists who believe data sharing is important also exist within the current culture where those incentives don't exist. Thus they pursue the goals that are incentivized (e.g. high volume and/or high impact publications) and the current norm is perpetuated.

As social norms are self-reinforcing and slow to change, this culture shift requires a multi-tiered approach. A particularly effective strategy has been described and practiced by the Open Science Foundation [4]. In their "Pyramid of Social Change" model there are five progressive levels of intervention, with each level being dependent on sufficient success and maturity of the levels below (Figure 1). From bottom to top those levels are: 1. Making Data Sharing Possible, 2. Making it Easy, 3. Making it Normative, 4. Making it Rewarding, and 5. Making it Required. In this report, we summarize the initiatives and interventions that the DSST has led at each of these five levels. These efforts have simultaneously helped the NIMH IRP produce world class results with large, shared datasets (see Sections 2 and 3) while also creating new possibilities and standards for data transparency (see Sections 1, 4, and 5). In the section 6, we describe our plans for the future and where we believe the greatest opportunities exist to accelerate the wide adoption of open science practices.

1 Making Data Sharing Possible

For effective data sharing to be possible, it is critical that robust data standards and reliable infrastructure (e.g. data repositories) exist to make hosting, maintaining, and organizing that shared data possible. The DSST has worked to improve and expand these areas so that researchers have clear and well defined paths for sharing data across a wider range of modalities. Notably, we have expanded and complimented two existing data standards to include tabular phenotypic data and holographic stimulation in two-photon imaging. In addition, we have helped create two new repositories for researchers to more easily find and access PET and trial-level behavioral data.

1.1 Creating and Expanding Data Standards

Community-developed data standards are critical for shared data to be re-usable, yet many types of data collected by neuroscientists do not currently have established data standards. Data shared without sufficient information about how the data was collected and how it is organized (i.e., metadata) is effectively useless and, unfortunately, many data repositories are filled with such examples. The DSST has engaged in several projects to expand the space of available data standards in ways that are complimentary to existing standards and best practices.

In 2015 the data standard known as BIDS, or the Brain Imaging Data Structure [13] was created. The DSST came into existence almost simultaneously with BIDS and strongly advocated for it's adoption within the NIMH IRP. Further, DSST member Eric Earl has served as one of the ten BIDS Maintainers since 2021. As the original BIDS standard focused on neuroimaging, it gave relatively little guidance on how to represent tabular phenotypic data, which led to significant heterogeneity in the format and organization used for tabular data in BIDS-formatted shared datasets. To address this short-coming, Eric Earl led the creation of a new BIDS Extension Proposal (BEP036, [14]). The proposal provides guidance on how to format data according to best practices as well as software tools to modify tabular data formats [15]. The proposal was also featured in a 2022 presentation at the Organization of Human Brain Mapping conference [16].

In 2022, the DSST began a collaboration with Dr. Mark Histed and the Unit on Neural Computation and Behavior. We applied for and received a one-year seed grant from the Kavli Foundation to fund the creation of a data standard for holographic photostimulation patterns used in two-photon imaging experiments [6]. The new standard was implemented as an extension to the NeuroData Without Borders (NWB) framework and is now publicly available for download and use from the Python Package Index (PyPI) [17].

1.2 Creating and Expanding Data Repositories

In 2018, the DSST was approached by Bob Innis from the IRP's Molecular Imaging Branch about the absence of repositories for PET data sharing. In collaboration with Drs. Innis, Gitte Knudson, and Melanie Ganz, the DSST submitted a grant application to the BRAIN initiative to create a BIDS-based PET data repository that closely integrated with the existing OpenNeuro repository. Dylan Nielson, a data scientist on the DSST from 2017 to 2019, played a critical role in drafting the original grant proposal, which was funded in 2020 with DSST's Dr. Thomas as a co-PI [5]. This funding not only allowed for the creation of the OpenNeuroPET repository, but also allowed for a fundamental re-design of the pipelines that underlie OpenNeuro, such that it could support and validate a variety of additional modalities such as MEG, fNIRS, EEG, and iEEG. It also enabled the hiring of Anthony Galassi, a skilled programmer who also serves as one of the ten BIDS maintainers who have the critical role of managing the BIDS specification as it grows and expands.

In a collaboration with Sam Zorowitz, from Angela Langdon's lab in the NIMH, the DSST is working to make trial-level behavioral data more findable and accessible for computational modellers. At his previous position at Princeton, Sam hand-curated an impressive list of 661 open datasets containing trial-level behavioral data, which quickly became popular with researchers engaged in computational modeling of learning and decision-making. However, given the rapid pace of newly-published datasets, manual curation became untenable. The DSST is working to make the manual process of identifying and cataloging these datasets more automated. Once a month the system runs a query against the OpenAlex database [18] to retrieve the metadata on all new articles from journals that have previously published papers containing shared trial-level behavioral data. A similarity analysis is then conducted between these articles and the articles already in the database to determine which articles are topically similar. The full text of these papers is then analyzed to detect data sharing statements (See Section 4.1) that contain a DOI, URL, or other unique identifiers for the datasets. In our first pilot of this pipeline, we identified 193 publications that met our inclusion criteria and contained data sharing statements with URLs according to an existing regular expression-based algorithm [19]. Dr. Zorowitz then looked at each article manually and found that 167 of the articles referenced publicly available data. Seven of these were not sufficiently topically similar to be included in the database, 19 did not contain trial level data. The remaining 141 articles were added to the OpenCogData database, increasing the total number of datasets in the database to 793. The OpenCogData resource has been relocated from Princeton's web space to the DSST GitHub site where we intend to continuously update and maintain it with Dr. Zorowitz's guidance and assistance [20].

2 Making Data Sharing Easy

To improve the likelihood that researchers adopt open science practices, the DSST has engaged in multiple initiatives to train, advise, and assist researchers in incorporating these practices into their workflow. While the scope of these efforts primarily focuses on helping individuals and groups within the NIMH IRP, we have also provided similar support and guidance to other ICs as well as the regional and global scientific community. Our efforts to facilitate the adoption of open science practices can be grouped into the following categories: training, tool creation, using public datasets, and hands-on data sharing support.

2.1 Providing Training

The DSST provides organized individual and group training both on our own initiative and by request for researchers within the NIMH IRP. Typically our training is centered on a specific tool or technique relevant to our mission and expertise. For example, in December of 2019, Dustin Moraczewski and Arshitha Basavaraj organized a three day course on tools for making analysis code more reproducible for the Section on Developmental Neurogenomics, which also included a practical discussion on group-wide adoption of these tools. Similar topics were discussed in April of 2022 in which Eric Earl trained the NIMH's MEG Core on GitHub and best practices in collaborative coding. In another example of a focused training, in early 2023 Arshitha Basavaraj, Eric Earl, and Dustin Moraczewski provided workshops to multiple research groups on the implementation of the DSST defacing pipeline [21], a user-friendly pipeline developed to aid in the anonymization of structural MR images.

In addition to providing training on specific topics, the DSST also organizes more general events to build community and foster discussion surrounding open science practices. For example, in August of 2021 Adam Thomas led a team during the NIH Summer Internship Codeathon where trainees practiced data science and reproducible computing [22]. In July of 2022, Dustin Moraczewski and Arshitha Basavaraj held a discussion on best practices in open and reproducible science for Holly Lisanby's Noninvasive Neuromodulation Unit. To maintain community within the IRP during the COVID-19 pandemic, in 2021 the DSST began hosting a weekly event in which a speaker presents a primer or hosts a discussion on a data-science-related topic. Finally, the DSST has hosted a monthly series that brings researchers across institutes together for presentations and discussions on computer coding relevant to scientific fields. Topics include curating Python code into an easy to install package (i.e. pip), SQL database creation and management, technical documentation, and containerization (i.e. Docker).

The DSST seeks to build a larger community of researchers who embrace open science. To this end, the DSST has organized and/or participated in multiple hackathons to bring together the greater Washington D.C. and global scientific communities. In both December of 2021 and 2022 Dustin Moraczewski served as the president of the DC chapter of the Brainhack Global organizing committee, with other DSST members Arshitha Basavaraj and Jessica Dafflon also serving on the organizing committee. In April of 2022, Eric Earl organized an event for BIDS and OpenNeuro developers in which a prototype of a BIDS validator using the BIDS schema was produced [13]. In August of 2022 and February of 2023, Adam Thomas and Eric Earl organized similar coding events for the core BIDS and OpenNeuro developers [23]. Finally, DSST members have also participated in initiatives to educate the global neuroscience community. From October of 2020 to March of 2021, Dustin Moraczewski served as a teaching assistant for the ABCD-Repronim course in which over 100 students worldwide were educated on applying reproducible methods to a gold-standard, NIH-funded dataset (the ABCD Study), and larger principles of open science [24, 25]. In July of 2021, Eric Earl gave a presentation on an emerging community collection of data from the ABCD dataset for Neurohackademy, a summer school on neuroimaging and data science [26]. Finally, in July of 2023, Arshitha Basavaraj presented a poster at the annual Organization for Human Brain Mapping conference discussing best practices in preparing a dataset to be shared publicly [27].

2.2 Creating Tools for Managing Data

The DSST builds software tools to aid researchers in preparing and using shared datasets and practicing open, reproducible science. In this section we provide select examples of these initiatives. The NIH Office of Portfolio Analysis (OPA) developed a tool for obtaining bibliometric data from papers in PubMed. To facilitate access to the data by analysts, DSST member Travis Riddle wrote an R package that accesses and parses the data in a standardized format for further analysis [28]. Dustin Moraczewski has developed a series of scripts to assist researchers in interacting with NIH's High Performance Computing (HPC) cluster [29, 30]. Arshitha Basavaraj and Eric Earl have developed the aforementioned DSST Defacing Pipeline, which provides scripts and guidance to aid with anonymization of structural MR images [21]. In addition, DSST members have also regularly contributed to larger, community-driven open-source projects [31, 32, 33].

2.3 Using Large Shared Datasets

In the face of the concerns around reproducibility and the increasing prevalence of multivariate research approaches, many researchers require larger datasets than an individual lab can acquire [34]. Thus, large, publicly-available datasets and biobanks are becoming more important resources, however, utilizing these large datasets is rarely a simple endeavor. Each dataset is organized and distributed in idiosyncratic ways with complex data use agreements and considerable administrative maintenance. In addition, finding, downloading, storing, and accessing the data of interest is often time-consuming for groups with limited personnel, computational, and storage resources. The DSST is easing these burdens for investigators within NIMH IRP as well as other ICs with similar needs, which has proved to be among the most popular services DSST provides. A full list of datasets downloaded, curated, or otherwise managed by the DSST is available from our group's website.

Our efforts to simplify the use of these datasets can be divided into two broad categories: administrative and processing. On the administrative side, the DSST navigates and manages several data use agreements for multiple investigators. Some are relatively simple, while others have proved expensive and time consuming. In 2017, the DSST applied for access to the UK Biobank dataset [35], which was approved in 2018 [36] and then renewed in 2022 at a cost of $2.7k and $16.6k, respectively. For both the original application and the renewal, resolving issues with EU regulations (e.g. GDPR) and NIH regulation took approximate one year and many emails. However these efforts proved worthwhile as the UK Biobank dataset is unprecedented in it's sample size and depth of phenotypic data. Our application ultimately included 26 investigators across multiple ICs and is cited in several high-impact publications, with others currently under review and/or available as preprints [37, 38, 39, 40, 41].

With respect to technical administration, the DSST manages over two petabytes of disk space across multiple shared drives on the NIH HPC cluster, as well as a third petabyte on our dedicated server for archival and cold data storage. Researchers are granted access to these datasets after DSST staff confirm that the requester is in compliance with the data use agreement. Importantly, we have curated each dataset so that documentation and directory structure are user-friendly and, when applicable, we have organized the data according to established standards (e.g., BIDS). In addition, for many of the larger datasets, it is often unclear where to find variables of interest. The DSST has extensive knowledge of each dataset and we field many such inquiries, reducing researcher effort in locating variables of interest from potentially weeks to a single email or instant message. To date, DSST maintains access for 76 unique users across 31 different datasets with over 100,000 total MRI scan sessions.

On the processing side of our large, shared dataset services, the DSST works with collaborators to assist in strategizing, preparing, and implementing various processing pipelines. Creating and storing derivatives of large datasets is not trivial. For example, storing the outputs of fMRIPrep [42], a popular fMRI preprocessing pipeline, for the roughly 40k subjects in the UK Biobank requires 160 terabytes of disk space. We also facilitate interaction with phenotypic and genotypic data (e.g., psychometric data, clinical surveys, demographics, and Genotype-Tissue Expression). Thus researchers can look to the DSST to assist with custom analysis pipeline construction, implementation, and storage. Finally, MRIQC is a popular software program designed to calculate an extensive array of quality metrics on MRI data [43]. The DSST maintains a database of anonymized quality metrics from users across the world who have opted to have their data uploaded, which can then be used to train classifiers to automatically determine the quality of future images [42]. In 2019, we conducted an analysis of the the 282,000 unique entries in the database and found that flip angle and scanner manufacturer were most strongly associated with quality metrics for structural and functional scans, respectively [44]. As of December 2023, the database contains over 400,000 unique records for structural scans and over 490,000 records for functional scans.

2.4 Sharing IRP Datasets

The central goal of the creation of the DSST was to promote and facilitate sharing raw data collected in the NIMH IRP. However, preparing and uploading raw data to a public repository can be deceptively labor-intensive, requiring anywhere from tens to over a thousand personnel hours depending on the size and state of the dataset. The DSST strives to minimize this resource burden on groups within the NIMH IRP through iterative, hands-on assistance with preparing datasets. Our strategy is to provide direct assistance while also training groups in best practices in data management and workflows so future datasets can be shared more easily. Below are a few examples of datasets the DSST has prepared and made available, with a complete list in Section 7.2.

Tracking the typical and atypical morphological development of the human brain is essential to understanding neurodevelopmental disorders. The NIMH has been on the cutting edge of this research question since 1989 when the Longitudinal Structural Magnetic Resonance Imaging Study of Human Brain Development was launched, which tracked the brain development of participants 5-25 years old using T1-weighted structural MRI. Containing >6,000 brain scans from >2,000 subjects, findings from this dataset have been published in many high-impact journals and have received over 10,000 citations [45]. Making this dataset publicly available had been a long-standing but elusive goal of NIMH leadership and was among the first projects the DSST undertook as our team took shape. The DSST was fortunate that Armin Raznahan's group had recently reviewed this dataset and identified a large subset of high quality scans from neurotypical subjects for use in a study on normative brain size variation [46]. The logistics and preparation of the dataset required more than 500 personnel hours including substantial efforts from both the DSST and Dr. Raznahan's group. Those efforts were ultimately successful in making 1,516 scans from 787 subjects publicly available on the NIMH Data Archive in 2019 [47].

In collaboration with Joyce Chung and several other NIMH investigators, the DSST has been instrumental in preparing and maintaining the NIMH's Intramural Healthy Volunteer Dataset [48, 49], which focuses on extensively characterizing a cohort of healthy volunteers for studies of the neurobiology of psychiatric illness. Consisting of data from 157 subjects, this dataset contains multiple modalities, including clinical assessments, psychometrics, structural and functional MRI, diffusion, a comprehensive MEG battery, and blood samples. Due to the large number of modalities per subject, collating and curating this dataset required over 1,000 DSST personnel hours plus significant efforts from staff in both the Office of the Clinical Director and the MEG Core Facility. At the time of this writing, the Healthy Volunteer Dataset has more than 22,000 views, 444 downloads, and was recently used in a secondary analysis published in Nature Communications [50].

In addition to our collaboration with Dr. Innis on the OpenNeuroPET repository (see Section 1.1), the DSST has also prepared and deposited several datasets containing PET images. In collaboration with Carolyn Beebe Smith's group we curated and uploaded four related L-[1-^{11}C]Leucine 4D PET datasets in PET-BIDS format that also contained structural MRIs, blood sampling data, and other derived PET data [51, 52, 53, 54]. These four datasets include data from healthy subjects, subjects with Fragile X Syndrome as well subjects in different sleep states and under propofol anesthesia. We estimate that curating these four datasets required a minimum of 500 personnel hours from the DSST staff plus a substantial time investment from Dr. Smith's group.

In an ongoing collaboration with Francis McMahon and Human Genetics Branch, the DSST is assisting with the curation of genotype-sequence and phenotypic data from two datasets, to be uploaded to the NIH database of Genotypes and Phenotypes (dbGaP). One study seeks to uncover the genetic determinants of antidepressant response in subjects with treatment resistant depression [55], while another explores the genetic basis of bipolar disorder [56]. Curating and uploading these datasets is estimated to require over 200 DSST personnel hours.

Finally, in collaboration with the NIMH IRP's Statistical and Scientific Computing Core, the DSST collaborated on a project to investigate best practices in the quality control (QC) of fMRI data. We assembled a subset of imaging data from seven publicly-available studies that included resting-state or tasked-based fMRI into a single dataset [57]. We then asked the MRI community to submit this dataset to their own QC pipeline and prepare a manuscript describing their QC results. Manuscripts from ten groups were compiled into a special issue of Frontiers in Neuroscience [58] and members of the DSST co-authored the special issue's editorial [59].

3 Making Data Sharing Normative

Many investigators perceive that sharing data and making their methods fully transparent involves a prohibitive level of additional effort, time, resources, and/or risk. The DSST strives to combat these perceptions by conducting and publishing projects that demonstrate best practices for data sharing and re-use, as well as other open science principles, to promote transparency and reproducibility.

3.1 fMRI Correlates of Positive and Negative Phenotypes

Understanding the link between brain function and measures of behavior and cognition is a key goal of cognitive neuroscience, however the multidimensional landscape of possible links is vast. Many studies in the past decade have pursued this research question, yet it is unclear which methods yield replicable results. To this end, in collaboration with Peter Bandettini and Emily Finn from the Section on Functional Imaging Methods, we sought to replicate and extend a previous high-impact study that pursued this question using canonical correlation analysis [60]. The first goal of the replication study was to computationally replicate the findings from the original paper using code provided by the authors and the same public dataset [61]. This analysis, led by DSST post-baccalaureate trainee Nikhil Goyal, was published as a preprint [62].

We then attempted to replicate the findings using a much larger, more heterogeneous sample: the baseline time point from the ABCD dataset [63]. These data were collected from approximately 11,000 9- and 10-year old children across 21 scan sites using a variety of MRI scanners with different manufacturers and head coils. Here we chose to submit our study as a registered-report. In this model the introduction and methods of the paper are submitted and peer reviewed before the analysis begins. Publication of the registered-report is guaranteed if the methods were followed, regardless of the outcome or significance of the results. We received excellent feedback from our reviewers and were ultimately successful in most of our replication criteria [64]. The DSST's newest member, post-baccalaureate trainee Mia Zwally, is conducting a follow-up analysis using cortical gradients and intends to leverage the longitudinal elements of the ABCD dataset [65, 66].

3.2 Factors Influencing Mental Well-being in Response to COVID Pandemic

In 2020 a group of IRP investigators sought to better understand the mental health implications of the emerging COVID-19 pandemic. Led by Dr. Joyce Chung, longitudinal survey data from 3,655 participants was collected to assess how an individual's mental well-being changed over the course of the pandemic. This large dataset was analyzed by DSST post-baccalaureate trainee Carl Harris and in collaboration with the Machine Learning Team. They found that it is possible to predict the variation in participant's mental health outcomes from time point to time point using their personal characteristics and circumstances. The model was first created and tuned using a subset of the sample and the hypotheses were pre-registered on OSF before final testing on the held out subset [67]. The dataset and code are both publicly available [68, 69].

3.3 Leveraging Large, Open Data for Machine Learning

The NIMH seeks to improve the lives of those affected with mental health disorders. Unfortunately, given the prevalence of less common disorders, assembling a sample size with enough statistical power is often difficult. To this end, our ongoing collaboration with the Machine Learning Team seeks to leverage as much publicly-available data as possible to train a deep neural network (DNN) capable of predicting phenotypic information from fMRI functional connectivity. The ultimate goal is to then adapt that network to predict clinical information in smaller datasets. For these projects we have preprocessed the imaging data from more than 50k subjects across several large datasets through the same preprocessing pipeline and standardized the format of the corresponding behavioral and cognitive data [15]. DNNs have shown great potential for uncovering complex, multidimensional relationships between brain and behavior, however it is often difficult to interpret what information a model is using to make a prediction. Utilizing the task-based fMRI data distributed from the HCP dataset [61], one project sought to improve the interpretability of DNNs using adversarial training and gradient-based saliency maps [70]. Another project aims to improve interpretability of the latent constructs from phenotypic data [71]. Finally, another ongoing project aims to explore the implications of the dimensionality reduction of phenotypic data on studies that use whole-brain resting-state functional connectivity to predict phenotypic measures [72].

3.4 Genetic Influences on Brain Morphology

Given that many mental health disorders affect males and females differently, our ongoing collaboration with Armin Raznahan's Section on Developmental Neurogenomics investigates the effects of sex chromosomes on neuroanatomical variation in humans. We are using the UK Biobank, since prior genome-wide association studies (GWAS) typically exclude these chromosomes. Using the T1- and T2-weighted structural images from approximately 40,000 subjects, the DSST generated multiple surface-based measures of cortical thickness, curvature, surface area, and local gyrification index, as well as extracted averages of these measures through multiple whole-brain parcellations. We found evidence that cortical regions that exhibit a larger male bias show stronger evidence for full X-chromosome dosage compensation and that failing to include the X-chromosome in GWAS analysis causes the loss of important information [38, 41]. Future work can use these findings for a more nuanced investigation of the relationship between the genome, brain structure, and mental health disorders.

4 Making Data Sharing Rewarding

We see creating more incentives for investigators to share data, thus making data sharing more rewarding, as the level of the social change pyramid (Figure 1) with the most potential for growth. Various data sharing prizes and awards exist, but only a handful of individuals are recognized, and the awards have a small impact on their careers compared to that of published papers. If data sharing incentives are to be effective, they must be created at scale.

4.1 Measuring and Rewarding Effective Data Sharing

To reward data sharing there must be a mechanism to measure it. Since 1999, NIH's Office of the Director as well as it's 27 constituent centers and institutes have issued at least 29 different data sharing policies and mandates for their staff and grantees [73]. Some early policies only applied to larger grants while some institute policies required that data associated with a given grant be uploaded to specific repositories. However, the NIH has invested relatively little effort in ensuring that shared data is reported in published papers so that other investigators might find and re-use it. Filling this gap in funders' ability to incentivize data sharing is a central goal of our team.

In 2020, DSST member Travis Riddle led a project designed to identify reported data sharing in the full text of publications. In collaboration with the Machine Learning Team we created a classifier and applied it to all NIMH-funded papers from the NIMH PubMed Central archive from 2008 to 2019. We were able to detect data sharing statements in 4.5% of the papers. We also created an online portal that provided a "leader-board" for investigator and institution as well as a mechanism allowing anyone to submit correction to the paper classifications. The project was presented at the Organization for Human Brain Mapping conference and the web portal code, and classification results are publicly available [74].

More recently, in order to better understand data sharing within the NIH IRP, we applied a novel regular expression classification technique from Serghiou et al. [19] to all publications listed in NIH IRP annual reports from 2019 to 2023. NIH IRP papers showed an increase in the presence of data sharing statements over time – from 21% in 2019 to 32% to 2023. We also observed significantly different rates of data sharing reporting between institutes, with those ICs that fund more genome-related studies having higher rates of data sharing (Figure 2). We note however that in Serghiou et al's analysis using all publications in PubMed Central, they reported that their regular expression matching technique showed a specificity of 98.6% and a sensitivity of 75.8% when applied to a validation set of 6,017 recent publications in PubMed Central. These rates are in line with the false positive rate we observed in evaluating papers with open trial-level behavior data (167 out of 193 or 87%, see Section 1.2).

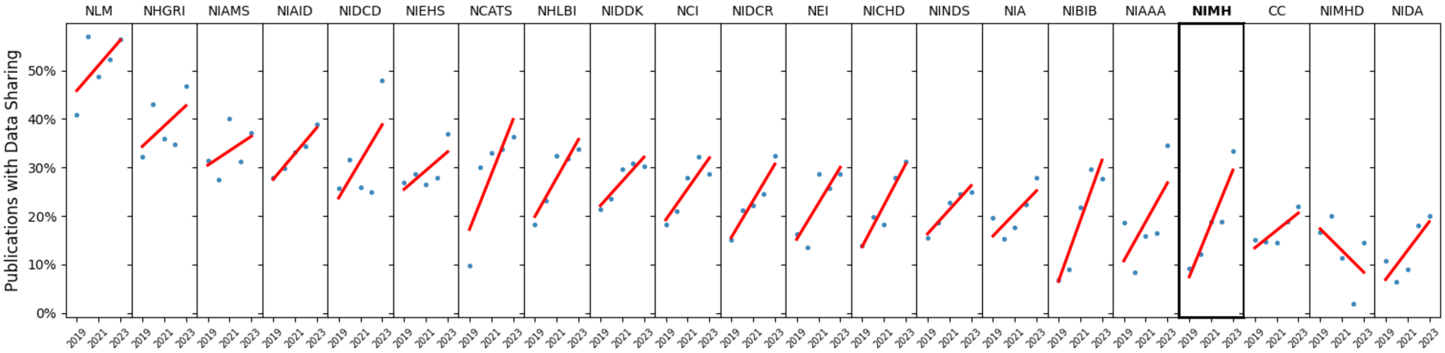

Figure 2: Blue dots indicate the percentage of papers in which a data sharing statement was detected for a given IC and year. Red lines show a simple linear fit across the five year period (2019 to 2023). IC intramural programs with very few publications (CIT, NCCIH, & NINR) have been excluded.

In collaboration with the Machine Learning Team, we next sought to improve upon measurement and characterization of data sharing reported in the biomedical literature by applying OpenAI's ChatGPT3.5 Large Language Model (LLM) [75] to the full text of the papers. For each paper we asked the LLM to report: 1. the species of the organism(s) studied in the paper, 2. the types of data collected (imaging, microscopy, genomic, behavioral, etc.), 3. whether or not a data sharing statement is included. If a data sharing statement was detected, we further asked the LLM to report: 4. the name of the data repository, 5. the DOI, URL, or other unique identifier of the dataset(s). We presented the model with a small, random subset of papers from the NIH IRP balanced to include an equal number of papers with and without data sharing statements according to the Serghiou et al. technique. Manual review of these papers revealed that the subset contained examples of papers that were incorrectly categorized by the regular expression based method, including both false positives and false negatives. The LLM-based approach correctly categorized all papers that did not contain data sharing statements, however it did produce false negatives. In these instances the data sharing statements were found in references and footnotes, which were not seen by the model. Subsequent testing will include these sections for a more comprehensive analysis.

5 Making Data Sharing Required

Since our team's beginning, we have witnessed an increasing number of investigators interested in sharing data. We believe that further adoption of data sharing in the biomedical science community will be best achieved by focusing on incentives rather than mandates or penalties. Nevertheless, the NIH's recently enacted Data Management and Sharing Policy demonstrates that data sharing is central to the US federal government's goal of making the results of taxpayer-supported research immediately and freely available to the American public [2]. To this end, our team plays a lead role in educating researchers and implementing these new requirement within the NIMH IRP.

5.1 Reviewing and Providing Guidance for Data Management and Sharing Plans

Compliance with the NIH Data Management and Sharing Policy that came into effect in January of 2023 requires every research group within the NIMH IRP to submit a Data Management and Sharing (DMS) Plan to the Scientific Director. The DSST held regular office hours in early 2023 to advise and assist groups in developing their DMS plans. In addition, Dr. Thomas led the creation of the NIMH DMS Plan Review Committee, composed of intramural investigators and staff from a variety of fields and backgrounds. Most research groups met with Dr. Thomas individually and/or received feedback on drafts of their DMS Plans before submitting for approval. Ultimately, the committee recommended DMS plans from every NIMH IRP investigator for the Scientific Director's approval. These efforts were recognized with a 2023 NIMH Director's Group Award to the committee.

As discussed in Section 4.1, the proportion of papers from the NIMH IRP containing data sharing statements was relatively low (15%) for the period of 2019 to 2022, owing in part to the relative paucity of genomics studies for which there is a well established culture of data sharing. Shortly before this report was drafted, the full listing of 2023 IRP publications was made available on the NIH Intramural Database [76]. Our analysis of the full text of these publications revealed that the proportion of NIMH publications containing data sharing statements increased more than two-fold to 33%. Relative to the other 21 IC's intramural programs, the NIMH IRP went from a rank of 20th for percentage of papers published in 2019 with data sharing statements, to 10th place for papers published in 2023 (Figure 2). We attribute this increase to the strong emphasis the NIMH IRP leadership has placed on compliance with the new DMS Policy and the commitment of IRP faculty to adopting its guidelines.

6 Summary and Future Directions

Our team is proud of our accomplishments over the past four years and we believe our mission is more important than ever. The NIMH IRP has made considerable progress in data sharing, but there is much more work to be done in order to meet the expectations of the NIH leadership. In addition to continuing to offer training, consulting, and hands-on support for data sharing and other practices to support transparency, the following section details new initiatives and future directions for our team as well as the additional resources needed to better serve the NIMH IRP.

6.1 Recognition and Awards for Data Sharing

Our analysis of data sharing in publications included in annual reports not only allows for comparison between institutes, as in Figure 2, but also between projects and investigators. The NIMH IRP projects and investigators with the highest proportion of papers that include data sharing statements can now be shared with the Scientific Director's office yearly. To spur the growth of data sharing, we recommend that these top ranking investigators be made known to the Board of Scientific Counselors and that they be acknowledged at the NIMH Director's Awards ceremony. If possible, we also recommend that the investigator and lab members receive monetary rewards in the form of cash bonuses or additional research funding. As our analysis spans the entire IRP, we can also share results with the intramural leadership of other institutes so they might recognize investigators who regularly report data sharing.

6.2 Improved Measurement of Data Sharing in Publications

Our project on cataloging open, trial-level behavioral datasets for computational modeling work (see Section 1.2) revealed that publications containing data sharing statements do not always provide the reader with access to the dataset. The regular expression matching approach of Serghiou et al. [19] cannot differentiate a number of edge cases where data is not in fact being shared. In collaboration with the Machine Learning Team, we are in the early phase of developing an alternative LLM-based technique to determine if a given publication includes the necessary information for the reader to find and download the dataset(s) analyzed in the paper (see section 4.1). Our technique will be able to identify when multiple data modalities are collected and shared, potentially in different data repositories. The technique will also be able to look for unique identifiers (e.g., DOIs, URLs, accession numbers) and run online queries to determine if those identifiers resolve to a repository which contains an accessible dataset. Once developed, these tools will be made available to extramural staff so they might also evaluate the inclusion of data sharing in the publications of grant recipients.

6.3 Data Sharing for More Modalities

In surveying the publications of the NIH IRP, we noted that the data reported can be grouped into five broad categories: tabular data (typically demographic or phenotype), medical imaging (e.g. MRI, PET, CAT, etc.), genomic/multi-omic (DNA, RNAseq, proteomics, etc.), electrophysiology (time series, mostly non-human), and microscopy (2D and 3D, light-based, EM, X-ray, etc). Each modality has idiosyncratic normative practices with respect to data sharing. The genomics/multi-omics field is perhaps the most advanced, with some types of medical imaging being in second place. By contrast, we noted a relative lack of agreed upon standards and infrastructure for microscopy data and, perhaps consequently, a decreased likelihood that raw microscopy data is publicly shared.

We view this gap as evidence that more investment in microscopy data sharing is required. The NIMH IRP has world class microscopy facilities and we have found ready and willing partners including Ted Usdin's Systems Neuroscience Imaging Resource (SNIR) and Francis McMahon's Human Genetics Branch. Dr. McMahon's group has collected multiple datasets using SNIR's specialized confocal microscope as well as NIBIB's Instant Structured Illumination Microscope (iSIM). We are in the early stages of a project with Drs. Usdin and McMahon to make the data publicly available in a way that maximizes re-usability. We will also use this dataset to design procedures and pipelines such that similar microscopy datasets can be shared more easily in the future.

6.4 Resources Requested

Our team is excited about the challenges and projects before us, but the scale of data sharing need in the IRP far outstrips our current capacity. Our team currently consist of three permanent PhD data scientist slots (all filled), and four "non-backfillable" slots (three filled, including one post-baccalaureate trainee). We expect the staff and trainee currently occupying these temporary or loaned slots will seek higher level external positions within the next 18 months, at which point the slots must be returned.

The IRP survey responses reveal that our users are very happy with the services and skills we provide. However, some feel our responses to their requests are not as fast as they would ideally like. Our response time is limited by our small staff and relatively modest resources, which are reflected in the overall cost of our team to the NIMH. Of the sixteen cores, services, and teams listed for NIMH in NIH RePORTER [10], the DSST's operating costs were among the lowest over the past four years. We submit that the DSST offers tremendous value to the NIMH IRP at a modest expense. In order to quickly and efficiently respond to the growing demand for our services, we require additional staff.

To help us address the increase in data sharing requests precipitated by the 2023 NIH Data Sharing Policy, the Office of the Clinical Director loaned us one additional PhD-level position, which we greatly appreciate. We are eager to fill this slot once the contracting agency approves it. We believe granting our team an additional PhD-level slot in 2024 would allow us to respond to the growing demand more quickly and to more efficiently execute the above initiatives. We also request that our two data scientist and one post-baccalaureate slots be converted to permanent or "core" slots so they may be back-filled and avoid a net loss of staff in the near future.

To conclude, we would like to express our thanks to the BSC members for their service to the NIMH IRP, and we look forward to receiving their feedback on how our team can better serve the NIMH as well as the larger mental health research community.

7 References

7.1 References Co-authored by DSST Members

Current Team Members

Adam Thomas
Team Lead

Dustin Moraczewski
Data Scientist

Arshitha Basavaraj
Data Engineer

Jessica Dafflon
Data Scientist

Eric Earl
Data Scientist

Mia Zwally
Postbac IRTA

Former Team Members

Carl Harris
Postbac IRTA (2021 - 2023)
Now: PhD Candidate at JHU

Nikhil Goyal
Postbac IRTA (2019 - 2020)
Now: MD Student at UPenn

Dylan Nielson
Data Scientist (2017 - 2019)
Now: NIH Machine Learning Team

Travis Riddle
Data Scientist (2017 - 2019)
Now: Senior Research Fellow, Consumer Financial Protection Bureau

Nino Migineishvili
Summer Student (2017 & 2018)
Now: Research Data Analyst, California Policy Lab

John Lee
Data Scientist (2016 - 2018)
Now: Founder of Python AI Solutions

[5] Robert Innis, ..., **Adam G Thomas**, ... et al. *OpenNeuroPET: An Archive for PET data (NIH BRAIN Project MH002977)*. 2021. URL: https://intramural.nih.gov/search/searchview.taf?ipid=122105&nidbreload=true.

[6] Mark Histed and **Adam G Thomas**. *Kavli Seed Grant Awardees: Extending the NWB standard for holographic photo-stimulation*. 2023. URL: https://www.nwb.org/projects/.

[7] Robert W Cox and **Dylan M Nielson**. *BRAIN power: expanding reproducibility, quality control, and visualization in AFNI/SUMA (NIH Annual Report MH002974)*. 2019. URL: https://intramural.nih.gov/search/searchview.taf?ipid=112483&nidbreload=true.

[9] **Adam G Thomas**. *Curriculum vitae for Adam G Thomas*. 2023. URL: https://cmn.nimh.nih.gov/sites/default/files/inline-files/Adam_G_Thomas_CV_December_2023.pdf.

[13] Russell A Poldrack, ..., **Eric Earl**, ..., **Adam G Thomas**, ... et al. "The Past, Present, and Future of the Brain Imaging Data Structure (BIDS)". In: *arXiv [q-bio.OT]* (2023). arXiv: 2309.05768 [q-bio.OT]. URL: http://arxiv.org/abs/2309.05768.

[14] **Eric Earl** and Samuel Guay. *BIDS Extension Proposal 036 - Phenotypic Data Guidelines*. 2022. URL: https://bids.neuroimaging.io/bep036.

[15] **Eric Earl**, ..., **Jessica Dafflon**, ..., **Arshitha Basavaraj**, ..., **Dustin Moraczewski**, ..., **Adam G Thomas**. *BIDS tabular phenotype data software for big neuroimaging studies*. 2023. DOI: 10.17605/OSF.IO/VN4YQ.

[16] Samuel Guay, ..., **Eric Earl**, ..., **Adam G Thomas**. *New guidelines for phenotypic data in Brain Imaging Data Structure (BIDS).* Glasgow, Scotland, 2022. DOI: 10.17605/OSF.IO/35SXV.

[17] Paul laFosse, ..., **Carl Harris**, ..., **Adam G Thomas**, ... et al. *ndx-photostim - An Extension to the NeuroData Without Borders (NWB) standard for holographic stimulation.* 2023. URL: https://pypi.org/project/ndx-photostim/.

[20] Sam Zorowitz, ..., **Dylan Nielson**, ..., **Arshitha Basavaraj**, ..., **Jessica Dafflon**, ..., **Eric Earl**, ..., **Adam G Thomas**. *OpenCogData: A collection of publicly available cognitive task datasets maintained by the Data Science & Sharing Team at the National Institute of Mental Health.* 2023. DOI: 10.5281/zenodo.10257543.

[21] **Arshitha Basavaraj** and **Eric Earl**. *NIMH DSST Defacing Pipeline.* 2023. DOI: 10.5281/zenodo.10182998.

[22] Tarek Antar, ..., **Dustin Moraczewski**, ..., **Adam G Thomas**. *Visualizing the Nutritional Landscape of Food: An NIH Codeathon Project.* 2021. DOI: 10.5281/zenodo.5504204.

[27] **Arshitha Basavaraj**, ..., **Eric Earl**, ..., **Adam G Thomas**, ... et al. *Data curation of the NIMH Healthy Volunteer dataset.* Montreal, Canada, 2023. DOI: 10.17605/OSF.IO/YK7GW.

[28] **Travis Riddle**. *iCiteR: An R package that acts as a wrapper around the NIH's iCite API.* 2019. URL: https://github.com/riddlet/iCiteR.

[29] **Dustin Moraczewski**. *A Collection of Helper Scripts for the NIH HPC Cluster.* 2023. DOI: 10.5281/zenodo.10198362.

[30] **Dustin Moraczewski**. *A Collection of fMRIPrep Helper Scripts to Summarize Error Reports and Generate Motion Censor Files.* 2023. DOI: 10.5281/ZENODO.10198367.

[31] Tal Yarkoni, ..., **Dylan M Nielson**, ..., **John A Lee**, ... et al. "PyBIDS: Python tools for BIDS datasets". In: *J Open Source Softw* 4.40 (2019). DOI: 10.21105/joss.01294.

[32] Oscar Esteban, ..., **Dylan Nielson**, ..., **John A Lee**, ... et al. *Nipype: Neuroimaging in Python: Pipelines and Interfaces.* 2022. DOI: 10.5281/zenodo.6834519.

[33] Jakub Kaczmarzyk, ..., **Dylan M Nielson**, ... et al. *Neurodocker: a command-line program that generates custom Dockerfiles and Singularity recipes for neuroimaging and minifies existing containers.* 2023. DOI: 10.5281/zenodo.1477094.

[34] Scott Marek, ..., **Eric Earl**, ... et al. "Reproducible brain-wide association studies require thousands of individuals". In: *Nature* 603.7902 (2022), pp. 654–660. DOI: 10.1038/s41586-022-04492-9.

[36] **Adam G Thomas**. *Application 22875: Confirmation and expansion of NIH intramural results related to brain imaging, gene-dose effects and genetic scores.* 2018. URL: https://biobank.ndph.ox.ac.uk/ukb/app.cgi?id=22875.

[38] Travis T Mallard, ..., **Dustin Moraczewski**, ..., **Adam G Thomas**, ... et al. "X-chromosome influences on neuroanatomical variation in humans". In: *Nat. Neurosci.* 24.9 (2021), pp. 1216–1224. DOI: 10.1038/s41593-021-00890-w.

[39] Maxwell A Bertolero, ..., **Dustin Moraczewski**, ..., **Adam G Thomas**, ... et al. "Deep Neural Networks Carve the Brain at its Joints". In: *arXiv [q-bio.NC]* (2020). DOI: 10.48550/arXiv.2002.08891. arXiv: 2002.08891 [q-bio.NC].

[41] Rebecca Shafee, ..., **Dustin Moraczewski**, ..., **Adam G Thomas**, ... et al. "A sex-stratified analysis of the genetic architecture of human brain anatomy". In: *medRxiv* (2023). DOI: 10.1101/2023.08.09.23293881.

[44] **Dylan Nielson**, ..., **Adam G Thomas**. "Distributions of image quality metrics in the MRIQC Web-API, OHBM 2019". In: OSF, 2019. DOI: 10.17605/OSF.IO/NE6TR.

[48] Allison C Nugent, ..., **Adam G Thomas**, ..., **Arshitha Basavaraj**, ..., **Eric Earl**, ..., **Travis Riddle**, ... et al. "The NIMH intramural healthy volunteer dataset: A comprehensive MEG, MRI, and behavioral resource". In: *Sci Data* 9.1 (2022), p. 518. DOI: 10.1038/s41597-022-01623-9.

[59] Paul A Taylor, ..., **Arshitha Basavaraj**, ..., **Dustin Moraczewski**, ... et al. "Editorial: Demonstrating quality control (QC) procedures in fMRI". In: *Front. Neurosci.* 17 (2023). DOI: 10.3389/fnins.2023.1205928.

[62] **Nikhil Goyal**, …, **Dustin Moraczewski**, …, **Adam G Thomas**. "Computationally replicating the Smith et al. (2015) positive-negative mode linking functional connectivity and subject measures". In: *bioRxiv* (2020), p. 2020.04.23.058313. DOI: 10.1101/2020.04.23.058313.

[64] **Nikhil Goyal**, …, **Dustin Moraczewski**, …, **Adam G Thomas**. "The positive-negative mode link between brain connectivity, demographics and behaviour: a pre-registered replication of Smith et al. (2015)". In: *R Soc Open Sci* 9.2 (2022), p. 201090. DOI: 10.1098/rsos.201090.

[65] **Mia Zwally**, …, **Dustin Moraczewski**, …, **Adam G Thomas**. *Characterizing the Relationship Between Cortical Gradients and Cognitive Traits in Children*. 2023. DOI: 10.5281/zenodo.10277223.

[66] **Mia Zwally**, …, **Dustin Moraczewski**, …, **Adam G Thomas**. *Preregistration of Characterizing the Relationship Between Cortical Gradients and Cognitive Traits in Children*. 2023. DOI: 10.17605/OSF.IO/T9DHK.

[67] **Carl Harris**, …, **Adam G Thomas**, … et al. *Preregistration of Predictions of mental well-being from individual characteristics and circumstances during the COVID-19 pandemic*. 2022. DOI: 10.17605/OSF.IO/ATXCG.

[69] **Carl Harris**, …, **Adam G Thomas**, … et al. "Prediction of mental well-being from individual characteristics and circumstances during the COVID-19 pandemic". In: (2022). DOI: 10.31234/osf.io/7enqw.

[70] Patrick McClure, …, **Dustin Moraczewski**, …, **Adam G Thomas**, … et al. "Improving the interpretability of fMRI decoding using deep neural networks and adversarial robustness". In: *Aperture Neuro* (2023). DOI: 10.52294/001c.85074.

[72] **Jessica Dafflon**, …, **Eric Earl**, …, **Dustin Moraczewski**, …, **Adam G Thomas**, … et al. *Predictability of phenotype information from functional connectivity in large imaging datasets*. 2023. DOI: 10.17605/OSF.IO/EWR8M.

[74] **Travis Riddle**, …, **Adam G Thomas**. *Identifying Data Sharing and Data Reuse in Full-text NIMH-funded papers*. 2020. DOI: 10.5281/zenodo.3905326.

[77] Martin Norgaard, …, **Adam G Thomas**, … et al. "PET-BIDS, an extension to the brain imaging data structure for positron emission tomography". In: *Sci Data* 9.1 (2022), p. 65. DOI: 10.1038/s41597-022-01164-1.

[78] Martin Schweinsberg, …, **Travis Riddle**, … et al. "Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis". In: *Organ. Behav. Hum. Decis. Process.* 165 (2021), pp. 228–249. DOI: 10.1016/j.obhdp.2021.02.003.

[80] **Eric Earl**, …, Adam Thomas. *Portland BIDS Sprint 1 Report*. 2022. DOI: 10.5281/zenodo.10371983.

[81] **Eric Earl**, …, **Jessica Dafflon**, …, **Arshitha Basavaraj**, …, **Dustin Moraczewski**, …, **Adam G Thomas**. *Big neuroimaging dataset BIDS tabular phenotype tools*. 2023. DOI: 10.5281/zenodo.8084023.

[82] **Jessica Dafflon**, … et al. "A guided multiverse study of neuroimaging analyses". In: *Nat. Commun.* 13.1 (2022), p. 3758. DOI: 10.1038/s41467-022-31347-8.

[83] Bruno Aristimunha, …, **Adam G Thomas**, …, **Jessica Dafflon**. "Synthetic Sleep EEG Signal Generation using Latent Diffusion Models". In: *Deep Generative Models for Health Workshop NeurIPS 2023*. 2023. URL: https://openreview.net/pdf?id=mDwURmlapW.

[84] Walter H L Pinaya, …, **Jessica Dafflon**, … et al. "Generative AI for Medical Imaging: extending the MONAI Framework". In: (2023). arXiv: 2307.15208 [eess.IV]. URL: http://arxiv.org/abs/2307.15208.

[85] Walter H L Pinaya, …, **Jessica Dafflon**, … et al. "Brain Imaging Generation with Latent Diffusion Models". In: *Deep Generative Models*. Springer Nature Switzerland, 2022, pp. 117–126. DOI: 10.1007/978-3-031-18576-2_12.

[86] Pedro F Da Costa, …, **Jessica Dafflon**, … et al. "Transformer-based normative modelling for anomaly detection of early schizophrenia". In: (2022). arXiv: 2212.04984 [cs.LG]. URL: http://arxiv.org/abs/2212.04984.

[88] Gang Chen, ..., **Dustin Moraczewski**, ... et al. “Improving accuracy and precision of heritability estimation in twin studies: Reassessing the measurement error assumption”. In: *bioRxiv* (2023), p. 2023.06.24.546389. DOI: 10.1101/2023.06.24.546389.

[97] Melanie Pincus, ..., **Eric Earl**, ... et al. “Chronic psychosocial stress and experimental pubertal delay affect socioemotional behavior and amygdala functional connectivity in adolescent female rhesus macaques”. In: *Psychoneuroendocrinology* 127 (2021), p. 105154. DOI: 10.1016/j.psyneuen.2021.105154.

[98] Merel C Postema, ..., **Eric Earl**, ... et al. “Analysis of structural brain asymmetries in attention-deficit/hyperactivity disorder in 39 datasets”. In: *J. Child Psychol. Psychiatry* 62.10 (2021), pp. 1202–1219. DOI: 10.1111/jcpp.13396.

[99] David F Montez, ..., **Eric Earl**, ... et al. “Using synthetic MR images for distortion correction”. In: *Dev. Cogn. Neurosci.* 60 (2023), p. 101234. DOI: 10.1016/j.dcn.2023.101234.

[100] Matthew Cieslak, ..., **Eric Earl**, ... et al. “QSIPrep: an integrative platform for preprocessing and reconstructing diffusion MRI data”. In: *Nat. Methods* 18.7 (2021), pp. 775–778. DOI: 10.1038/s41592-021-01185-5.

[101] Carolina Badke D’Andrea, ..., **Eric Earl**, ... et al. “Real-time motion monitoring improves functional MRI data quality in infants”. In: *Dev. Cogn. Neurosci.* 55 (2022), p. 101116. DOI: 10.1016/j.dcn.2022.101116.

[102] Sydney Kaplan, ..., **Eric Earl**, ... et al. “Filtering respiratory motion artifact from resting state fMRI data in infant and toddler populations”. In: *Neuroimage* 247 (2022), p. 118838. DOI: 10.1016/j.neuroimage.2021.118838.

[103] Kristina M Rapuano, ..., **Eric Earl**, ... et al. “An open-access accelerated adult equivalent of the ABCD Study neuroimaging dataset (a-ABCD)”. In: *Neuroimage* 255 (2022), p. 119215. DOI: 10.1016/j.neuroimage.2022.119215.

[104] Omid Kardan, ..., **Eric Earl**, ... et al. “Resting-state functional connectivity identifies individuals and predicts age in 8-to-26-month-olds”. In: *Dev. Cogn. Neurosci.* 56 (2022), p. 101123. DOI: 10.1016/j.dcn.2022.101123.

[105] Oscar Miranda-Dominguez, ..., **Eric Earl**, ... et al. “Carotenoids improve the development of cerebral cortical networks in formula-fed infant macaques”. In: *Sci. Rep.* 12.1 (2022), p. 15220. DOI: 10.1038/s41598-022-19279-1.

[106] Aiden Ford, ..., **Eric Earl**, ... et al. “Functional maturation in visual pathways predicts attention to the eyes in infant rhesus macaques: Effects of social status”. In: *Dev. Cogn. Neurosci.* 60 (2023), p. 101213. DOI: 10.1016/j.dcn.2023.101213.

[107] Nora Byington, ..., **Eric Earl**, ... et al. “Polyneuro risk scores capture widely distributed connectivity patterns of cognition”. In: *Dev. Cogn. Neurosci.* 60 (2023), p. 101231. DOI: 10.1016/j.dcn.2023.101231.

[108] Carolina Badke D’Andrea, ..., **Eric Earl**, ... et al. “Thalamo-cortical and cerebello-cortical functional connectivity in development”. In: *Cereb. Cortex* 33.15 (2023), pp. 9250–9262. DOI: 10.1093/cercor/bhad198.

[109] Viola Neudecker, ..., **Eric Earl**, ... et al. “Early-in-life isoflurane exposure alters resting-state functional connectivity in juvenile non-human primates”. In: *Br. J. Anaesth.* 131.6 (2023), pp. 1030–1042. DOI: 10.1016/j.bja.2023.07.031.

[110] Z A Kovacs-Balint, ..., **Eric Earl**, ... et al. “The role of puberty on physical and brain development: A longitudinal study in male Rhesus Macaques”. In: *Dev. Cogn. Neurosci.* 60 (2023), p. 101237. DOI: 10.1016/j.dcn.2023.101237.

[111] Aki Nikolaidis, ..., **Eric Earl**, ... et al. “Proceedings of the OHBM Brainhack 2021”. In: *Aperture Neuro* 3 (2023), pp. 1–20. DOI: 10.52294/258801b4-a9a9-4d30-a468-c43646391211.

[112] Robert J M Hermosillo, ..., **Eric Earl**, ... et al. “A Precision Functional Atlas of Network Probabilities and Individual-Specific Network Topography”. In: *bioRxiv* (2022), p. 2022.01.12.475422. DOI: 10.1101/2022.01.12.475422.

[113] Zsofia Kovacs Balint, ..., **Eric Earl**, ... et al. “Brain Development During Adolescence in Male Rhesus Macaques: The Role of Puberty”. In: *Biol. Psychiatry* 89.9, Supplement (2021), S291. DOI: 10.1016/j.biopsych.2021.02.725.

[124] Emily S Finn, ..., **Dylan Nielson**, ... et al. “Idiosynchrony: From shared responses to individual differences during naturalistic neuroimaging”. In: *Neuroimage* 215 (2020), p. 116828. DOI: 10.1016/j.neuroimage.2020.116828.

[125] Vinai Roopchansingh, ..., **Dylan M Nielson**, ... et al. “EPI Distortion Correction is Easy and Useful, and You Should Use It: A case study with toddler data”. In: *bioRxiv* (2020), p. 2020.09.28.306787. DOI: 10.1101/2020.09.28.306787.

[126] **Dylan M Nielson** and Per B Sederberg. “MELD: Mixed effects for large datasets”. In: *PLoS One* 12.8 (2017), e0182797. DOI: 10.1371/journal.pone.0182797.

[127] Patrick McClure, ..., **John A Lee**, ..., **Dylan Nielson**, ... et al. “Distributed Weight Consolidation: A Brain Segmentation Case Study”. In: *Adv. Neural Inf. Process. Syst.* 31 (2018), pp. 4093–4103. URL: https://www.ncbi.nlm.nih.gov/pubmed/34376963.

[128] Patrick McClure, ..., **John A Lee**, ..., **Dylan M Nielson**, ..., **Adam G Thomas**, ... et al. “Knowing What You Know in Brain Segmentation Using Bayesian Deep Neural Networks”. In: *Front. Neuroinform.* 13 (2019), p. 67. DOI: 10.3389/fninf.2019.00067.

[129] Benjamin P Kay, ..., **Eric Earl**, ... et al. “Motion Impact Score for Detecting Spurious Brain-Behavior Associations”. In: *bioRxiv* (2022), p. 2022.12.16.520797. DOI: 10.1101/2022.12.16.520797.

[130] Peter A Bandettini, ..., **Adam G Thomas**. “The challenge of BWAs: Unknown unknowns in feature space and variance”. In: *Med* 3.8 (2022), pp. 526–531. DOI: 10.1016/j.medj.2022.07.002.

[131] Alex DeCasien, ..., **Adam G Thomas**, ... et al. “Linking X-Y Gametologue Co-Expression Patterns to Sex Differences in Disease”. In: vol. 93. San Diego, California: Elsevier, 2023, S93. DOI: 10.1016/j.biopsych.2023.02.240.

[132] Satrajit S Ghosh, ..., **Adam G Thomas**, ... et al. “A very simple, re-executable neuroimaging publication”. In: *F1000Res.* 6 (2017), p. 124. DOI: 10.12688/f1000research.10783.2.

[133] Laurentius Huber, ..., **Adam G Thomas**, ... et al. “Fast dynamic measurement of functional T1 and grey matter thickness changes during brain activation at 7T”. In: *F1000Res.* 6 (2017). DOI: 10.7490/f1000research.1114359.1.

[134] Rotem Botvinik-Nezer, ..., **Dylan M Nielson**, ... et al. “Variability in the analysis of a single neuroimaging dataset by many teams”. In: *Nature* 582.7810 (2020), pp. 84–88. DOI: 10.1038/s41586-020-2314-9.

[135] **Dylan M Nielson**, ..., **John A Lee**, ..., **Adam G Thomas**, ... et al. “Detecting and harmonizing scanner differences in the ABCD study - annual release 1.0”. In: *bioRxiv* (2018), p. 309260. DOI: 10.1101/309260.

[136] Nino Migineishvili, ..., **Dylan Nielson**, ..., **Adam G Thomas**, ... et al. *Parsimony and Machine Learning in Neuroimaging*. 2022. DOI: 10.17605/OSF.IO/R5BPC.

[137] David C Jangraw, ..., **Adam G Thomas**, ..., **Dylan M Nielson**, ... et al. “A highly replicable decline in mood during rest and simple tasks”. In: *Nat Hum Behav* (2023). DOI: 10.1038/s41562-023-01519-7.

[138] Gitte M Knudsen, ..., **Adam G Thomas**, ... et al. “Guidelines for the content and format of PET brain data in publications and archives: A consensus paper”. In: *J. Cereb. Blood Flow Metab.* 40.8 (2020), pp. 1576–1585. DOI: 10.1177/0271678X20905433.

7.2 References for Datasets Prepared and Uploaded by DSST

[47] Armin Raznahan, ... et al. *Normative brain size variation and brain shape diversity in humans - the NIMH IRP Longitudinal Development t1w MRI dataset 1990-2010 #783*. 2019. DOI: 10.15154/1504177.

[49] Allison C Nugent, ..., **Adam G Thomas**, ..., **Arshitha Basavaraj**, ..., **Eric Earl**, ..., **Travis Riddle**, ... et al. *The National Institute of Mental Health (NIMH) Intramural Healthy Volunteer Dataset*. 2023. DOI: 10.18112/openneuro.ds004215.v1.0.2.

[51] Shrinivas Bishu, … et al. *Effects of propofol anesthesia on rates of cerebral protein synthesis (rCPS)*. 2023. DOI: https://doi.org/10.18112/openneuro/ds004730.

[52] Dante Picchioni, … et al. *Rates of cerebral protein synthesis (rCPS) and memory formation during sleep*. 2023. DOI: 10.18112/openneuro.ds004731.v1.0.0.

[53] Dante Picchioni, … et al. *Rates of cerebral protein synthesis (rCPS) in stages of sleep*. 2023. DOI: 10.18112/openneuro.ds004733.v1.0.0.

[54] Kathleen C Schmidt, … et al. *Rates of cerebral protein synthesis (rCPS) in subjects with fragile X syndrome*. 2023. DOI: 10.18112/openneuro.ds004654.v1.0.1.

[55] Francis J McMahon, … et al. *Genetic Determinants of Antidepressant Response*. 2023. URL: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003329.v1.p1.

[56] Francis J McMahon, … et al. *National Institute of Mental Health (NIMH) Amish Mennonite Bipolar Genetics Study (AmBiGen)*. 2019. URL: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000899.v1.p1.

[57] Paul Taylor, …, **Dustin Moraczewski**, …, **Arshitha Basavaraj**. *FMRI Open QC Project*. 2022. DOI: 10.17605/OSF.IO/QAESM.

[68] **Carl Harris**, …, **Adam G Thomas**, … et al. *Dataset used in Prediction of mental well-being from individual characteristics and circumstances during the COVID-19 pandemic*. 2023. DOI: 10.7910/DVN/L4LRM2.

[87] Philip Shaw, …, **John Lee**, …, **Nino Migineishvilli**, …, **Dylan Nielson**, … et al. *NIMH NHGRI data-sharing project*. 2017. DOI: 10.15154/1463004.

[91] Elise M Cardinale, … et al. *The NIMH Intramural Multivariate Assessment of Inhibitory Control in Youth: Links with Psychopathology and Brain Function Dataset*. 2023. DOI: 10.18112/openneuro.ds004724.v1.0.0.

[92] Stefano Marenco, … et al. *Cellular Diversity in Human Subgenual Anterior Cingulate and Dorsolateral Prefrontal Cortex by Single-Nucleus RNA-Sequencing*. 2021. DOI: 10.15154/5a45-ek87.

[93] Javier Gonzalez-Castillo, … et al. *Large Single-Subject Functional MRI Datasets at 7T*. 2018. DOI: 10.18112/openneuro.ds001555.v1.0.1.

[94] Peter A Bandettini, … et al. *Task Dependence, Tissue Specificity, and Spatial Distribution of Widespread Activations in Large Single-Subject Functional MRI Datasets at 7T*. 2018. DOI: 10.15154/1464520.

[95] Jonathan Power, … et al. *Multi-echo Cambridge*. 2018. DOI: 10.18112/openneuro.ds000258.v1.0.1.

[114] Daniel Pine, … et al. *Ecological Momentary Assessment of Youth Anxiety: Evaluation of Psychometrics for use in Clinical Trials dataset*. 2023. URL: https://osf.io/av5r4.

[115] Amicia Elliott and Benjamin White. *Pupal behavior emerges from unstructured muscle activity in response to neuromodulation in Drosophila*. 2021. DOI: 10.6084/m9.figshare.c.5489637.v1.

[117] Berron Brown, … et al. *Associations between neighborhood resources and youth's response to reward omission in a task modeling negatively biased environments*. 2023. DOI: 10.18112/openneuro.ds004847.v1.0.0.

[118] Peter A Bandettini, … et al. *Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis*. 2018. DOI: 10.15154/1464517.

[119] Audrey Thurm. *Markers of Autism Spectrum Disorders in At-Risk Toddlers: A Pilot Study*. 2018. DOI: 10.15154/1464602.

[120] Susan Swedo and Audrey Thurm. *Clinical and Immunological Investigations of Subtypes of Autism*. 2023. URL: https://nda.nih.gov/edit_collection.html?id=2368.

[121] Cameron C McKay, … et al. *Emotion and Development Branch Phenotyping and DTI (2012-2017)*. 2023. DOI: 10.18112/openneuro.ds004605.v1.0.0.

[122] Javier Gonzalez-Castillo, … et al. *100 runs at 3T*. 2018. DOI: 10.18112/openneuro.ds001553.v1.0.1.

[123] Emily S Finn, … et al. *Layer-dependent activity in human prefrontal cortex during working memory*. 2019. DOI: 10.18112/openneuro.ds002076.v1.0.1.

[140] Audrey Thurm, …, **Arshitha Basavaraj**, …, **Eric Earl**, …, **Adam G Thomas**. *Structural MRI scans in autism during early childhood in BIDS format.* 2022. DOI: 10.15154/1528371.

7.3 References from Outside DSST

[1] Alan E Guttmacher, … et al. "Why data-sharing policies matter". In: *Proc. Natl. Acad. Sci. U. S. A.* 106.40 (2009), p. 16894. DOI: 10.1073/pnas.0910378106.

[2] White House Office of Science and Technology Policy. *OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay.* Tech. rep. 2022. URL: https://www.whitehouse.gov/ostp/news-updates/2022/08/25/ostp-issues-guidance-to-make-federally-funded-research-freely-available-without-delay/.

[3] Jan G Bjaalie, … et al. "Perspectives on Data Sharing and the New NIH policy from the European Union". In: *Harvard Data Science Review* (2022). DOI: 10.1162/99608f92.bcd0b999.

[4] Brian Nosek. *Strategy for Culture Change.* 2019. URL: https://www.cos.io/blog/strategy-for-culture-change.

[8] Jocelyn Kaiser and Jeffrey Brainard. "READY, SET, SHARE!" In: *Science* (2023). DOI: 10.1126/science.adg8470.

[10] Office of Behavioral and Social Sciences Research. *NIH Reporter.* 2020. URL: https://reporter.nih.gov.

[11] Francis S Collins and Lawrence A Tabak. "Policy: NIH plans to enhance reproducibility". In: *Nature* 505.7485 (2014), pp. 612–613. DOI: 10.1038/505612a.

[12] Lyric A Jorgenson, … et al. "Incentivizing a New Culture of Data Stewardship: The NIH Policy for Data Management and Sharing". In: *JAMA* (2021). DOI: 10.1001/jama.2021.20489.

[18] Jason Priem, … et al. "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts". In: (2022). arXiv: 2205.01833 [cs.DL]. URL: http://arxiv.org/abs/2205.01833.

[19] Stylianos Serghiou, … et al. "Assessment of transparency indicators across the biomedical literature: How open is open?" In: *PLoS Biol.* 19.3 (2021), e3001107. DOI: 10.1371/journal.pbio.3001107.

[23] Christopher Markiewicz. *BIDS Sprint 3 Report.* 2023. DOI: 10.5281/zenodo.10207821.

[24] David N Kennedy, … et al. "Everything Matters: The ReproNim Perspective on Reproducible Neuroimaging". In: *Front. Neuroinform.* 13 (2019), p. 1. DOI: 10.3389/fninf.2019.00001.

[25] Angela Laird, … et al. *ABCD-ReproNim Course.* 2022. URL: https://www.abcd-repronim.org/.

[26] Ariel Rokem. *Neurohackademy 2021 Schedule.* 2021. URL: https://neurohackademy.org/neurohack_year/2021/.

[35] Clare Bycroft, … et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726 (2018), pp. 203–209. DOI: 10.1038/s41586-018-0579-z.

[37] Zhiwei Ma, … et al. "Outlier detection in multimodal MRI identifies rare individual phenotypes among more than 15,000 brains". In: *Hum. Brain Mapp.* (2021). DOI: 10.1002/hbm.25756.

[40] Siyuan Liu, … et al. "Integrative structural, functional, and transcriptomic analyses of sex-biased brain organization in humans". In: *Proc. Natl. Acad. Sci. U. S. A.* 117.31 (2020), pp. 18788–18798. DOI: 10.1073/pnas.1919091117.

[42] Oscar Esteban, … et al. "Analysis of task-based functional MRI data preprocessed with fMRIPrep". In: *Nat. Protoc.* 15.7 (2020), pp. 2186–2202. DOI: 10.1038/s41596-020-0327-3.

[43] Oscar Esteban, … et al. "MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites". In: *PLoS One* 12.9 (2017), e0184661. DOI: 10.1371/journal.pone.0184661.

[45] Jay N Giedd, … et al. "Child psychiatry branch of the National Institute of Mental Health longitudinal structural magnetic resonance imaging study of human brain development". In: *Neuropsychopharmacology* 40.1 (2015), pp. 43–49. DOI: 10.1038/npp.2014.236.

[46] P K Reardon, … et al. "Normative brain size variation and brain shape diversity in humans". In: *Science* 360.6394 (2018), pp. 1222–1227. DOI: 10.1126/science.aar2578.

[50] Anna Zapaishchykova, … et al. "Automated temporalis muscle quantification and growth charts for children through adulthood". In: *Nat. Commun.* 14.1 (2023), p. 6863. DOI: 10.1038/s41467-023-42501-1.

[58] Paul Taylor, … et al. "Demonstrating Quality Control (QC) Procedures in fMRI [Special Issue]". In: *Front. Neurosci.* 17 (2023). URL: https://www.frontiersin.org/research-topics/33922/demonstrating-quality-control-qc-procedures-in-fmri.

[60] Stephen M Smith, … et al. "A positive-negative mode of population covariation links brain connectivity, demographics and behavior". In: *Nat. Neurosci.* 18.11 (2015), pp. 1565–1567. DOI: 10.1038/nn.4125.

[61] David C Van Essen, … et al. "The WU-Minn Human Connectome Project: an overview". In: *Neuroimage* 80 (2013), pp. 62–79. DOI: 10.1016/j.neuroimage.2013.05.041.

[63] Nora D Volkow, … et al. "The conception of the ABCD study: From substance use to a broad NIH collaboration". In: *Dev. Cogn. Neurosci.* 32 (2018), pp. 4–7. DOI: 10.1016/j.dcn.2017.10.002.

[71] Ka Chun Lam, … et al. "Interpretable (meta)factorization of clinical questionnaires to identify general dimensions of psychopathology". In: 2022. URL: https://openreview.net/pdf?id=c5-qKzTbP2O.

[73] NIH Office of the Director. *NiH Institute and Center Data Sharing Policies*. 2023. URL: https://sharing.nih.gov/other-sharing-policies/nih-institute-and-center-data-sharing-policies.

[75] OpenAI. "GPT-4 Technical Report". In: (2023). DOI: 10.48550/arXiv.2303.08774.

[76] Office of Intramural Research. *NIH Intramural DataBase (NIDB)*. 2001. URL: https://intramural.nih.gov/.

[79] Robert B Innis, … et al. "Consensus nomenclature for in vivo imaging of reversibly binding radioligands". In: *J. Cereb. Blood Flow Metab.* 27.9 (2007), pp. 1533–1539. DOI: 10.1038/sj.jcbfm.9600493.

[89] John Darrell Van Horn and Michael S Gazzaniga. "Why share data? Lessons learned from the fMRIDC". In: *Neuroimage* 82 (2013), pp. 677–682. DOI: 10.1016/j.neuroimage.2012.11.010.

[90] Mark Hahnel, … et al. *The State of Open Data 2023*. 2023. DOI: 10.6084/m9.figshare.24428194.v1.

[96] Mark D Wilkinson, … et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Sci Data* 3 (2016), p. 160018. DOI: 10.1038/sdata.2016.18.

[116] NIH ACD Artificial Intelligence Working Group. *NIH ACD Artificial Intelligence Working Group Final Report*. 2019. URL: https://acd.od.nih.gov/working-groups/ai.html.

[139] Bruno J Strasser. "Genetics. GenBank–Natural history in the 21st Century?" In: *Science* 322.5901 (2008), pp. 537–538. DOI: 10.1126/science.1163399.