A Hitchhiker's Guide to OpenNeuro:

Secondary Analysis on the Web's Largest Repository of Open Neuroimaging Data

by

Rae R. Buckser

Bachelor of Science, Purdue University, 2015

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This thesis was presented

by

Rae R. Buckser

It was defended on

December 1, 2021

and approved by

Julie A. Fiez, Professor, Psychology

Finnegan J. Calabro, Research Assistant Professor, Psychiatry and Bioengineering

Thesis Advisor: Marc N. Coutanche, Associate Professor, Psychology

Copyright © by Rachel Rose Buckser

2021

A Hitchhiker's Guide to OpenNeuro:

Secondary Analysis on the Web's Largest Repository of Open Neuroimaging Data

Rae R. Buckser, MS

University of Pittsburgh, 2021

Functional MRI (fMRI) is a foundational tool of cognitive neuroscience, but logistical constraints shut many researchers out of data collection. Secondary analysis of open data is a way for researchers to develop novel fMRI analysis methods, ask original questions, and contribute to the neuroimaging literature without collecting new data. To date, much of the documentation and guidance available for working with open neuroimaging data has focused on replication or on secondary analysis of large datasets. There is a lack of concrete, helpful material available for research groups who wish to enable secondary analysis by sharing smaller-scale datasets, or for researchers who wish to use smaller open datasets for their own projects. This document attempts to bridge this gap by offering examples, general recommendations, and concrete guidelines for users and sharers on OpenNeuro, the largest online data repository exclusive to neuroimaging. For users, important considerations include carefully planning analyses and checking preliminary results before committing to downloading and analyzing a dataset. For sharers, important considerations include looking at documentation and supplemental files from a user's point of view. The accompanying appendices offer checklists designed to walk researchers through the process of uploading data to share or using OpenNeuro data in four basic fMRI analysis tasks. Using and sharing open neuroimaging data represents an investment in a field that has become

more collaborative than ever before. The guidance and perspective offered should help users and sharers to make a productive start at navigating the open data landscape with OpenNeuro.

Table of Contents

List of Figures

Preface

This project was conceived, executed, and presented during the COVID-19 pandemic. As such, I would like to thank the faculty, fellow graduate students, and loved ones who navigated extraordinarily difficult and constantly changing circumstances to assist and guide me during this process. I would especially like to extend my gratitude to my mentoring and thesis committees, my advisor, Dr. Marc N. Coutanche, and all the current and past members of the LeNS Lab.

Acknowledgements: I would like to thank Griffin Koch for his generous guidance and feedback during the early stages of this project. I also wish to thank James Hengenius, who read and provided feedback on every draft of this manuscript. Finally, I would like to acknowledge Xueying Ren and Juan Carlos Angel Rojas for their feedback on the appendices.

1.0 Introduction: The Promise of Open fMRI Data

In the past twenty years, functional MRI (fMRI) has experienced an explosion of exciting innovations for data acquisition and analysis. Every study raises new questions about human cognition and new opportunities for computational advancement. fMRI is truly a foundational tool in cognitive neuroscience. However, it is also exclusive; logistical concerns often make fMRI research untenable.

Any researcher who has collected fMRI data is familiar with the massive amounts of money, time, and labor required to collect enough data to enable an analysis. Between securing funding, recruiting participants, and conducting experimental sessions, achieving a sufficient sample size may take months or even years, depending on the project. Even a pilot study can be costly and time-consuming. This is of particular concern to graduate students, who must balance data collection with coursework, teaching, and time to completion. Committing to funding and conducting a functional imaging study is an understandably daunting task, especially for students and early-career researchers. Further, access to fMRI is limited not only by lab funding and resources, but by environment and institutional factors. Those without ready access to a scanner will never have the means to collect data of their own – and with scanners largely concentrated at large research centers, R1 universities, and hospitals (Laird, 2021), fMRI research can easily become siloed in urban areas and large, well-funded institutions.

Secondary analysis of open data – experimental data collected and made freely available for use – presents one solution to these pervasive concerns. Researchers can now test planned analyses on pre-existing data before committing to funding and conducting an original imaging study. Pre-existing data can validate novel methods of analysis, determining whether new

computational methods are tenable and effective without the added labor of data collection. Open data is available to any researchers who can download it, ameliorating some of the bias introduced by the economic realities of scanning and the systematic stress of the pandemic. Even those without the means or access to scan can contribute to the literature by developing new analysis methods or asking new questions that recontextualize the results of previous experiments.

Shared data also allows multiple research groups to conduct analyses and ask questions within a shared reference space. The most prominent example is the Human Connectome Project (HCP). Since 2010, HCP has operated a multi-site effort to create a repository of openly available functional, structural, and diffusion brain imaging resources as well as open-source analysis tools and pipelines. "HCP-style" paradigms for data acquisition, processing, and analysis have been widely adapted for new data. As of 2021, over 1500 publications have cited HCP (Elam *et al.,* 2021). Between 2019 and 2021, an average of 20-30 publications per month have cited HCP data (Elam *et al.,* 2021). The widespread use of HCP resources has created an entire body of neuroscience literature with a common origin point (Elam *et al*., 2021) spanning over a decade of work. This has enabled a style of large-scale collaboration previously unimaginable in this field. Other large-scale multi-site data sharing projects such as the ABCD Study (Bjork *et al*., 2017) and the ABIDE I and ABIDE II data repositories (Di Martino *et al.,* 2014; Di Martino *et al*., 2017) provide more targeted knowledge spaces. Besides the institutional cooperation inherent in creating large data repositories across research sites, big data projects create opportunities for collaboration and comparison between researchers who would not otherwise have the means (Laird, 2021).

As the widespread use of HCP-style big data illustrates, there is great demand for shared data in the neuroimaging community. A more recent trend, however, is a departure from big data: individual researchers or groups conducting secondary analyses on smaller datasets that were

previously collected for separate experiments. In many ways, this new dynamic is a result of the larger trend towards open science. OpenNeuro (Stanford Center for Reproducible Neuroscience, 2021) is the most prominent tool for open data-sharing of smaller-scale neuroimaging datasets. As the first online data repository dedicated to neuroimaging, it originated as a direct response to issues of transparency, access, and publication bias in the field of neuroimaging.

1.1 A Brief History of Data-Sharing in fMRI

Rosenthal (1979) first identified what he termed the "File Drawer Problem". With limited resources and a glut of submissions, journals are far more likely to publish significant results than null findings, resulting in a culture of research where null results are not disseminated – or even submitted. As Scargle (2000) points out, this renders the results of meta-analyses unreliable, as effect sizes are inflated when only published results are considered. Concerns over reliability across neuroscience literature were complicated by the widely-publicized replication crisis. In 2009, Vul *et al.* published the controversial article, "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition" (originally "Voodoo Correlations"). The paper drew attention to what the authors considered unrealistically low p-values in a number of fMRI studies in which published methods were too opaque to allow for replication.

While Vul *et al.* were challenged on the tone and content of the article (Lieberman, Berkman, & Wager, 2009), the questions they raised had an undeniable impact on the field. In their 2009 response to the article, Poldrack & Mumford acknowledged that while their own conclusions showed low p-values were not outside the realm of plausibility, the statistics offered in the article were "undeniable". One takeaway offered by Poldrack & Mumford (2009) was the

disturbing fact that Vul *et al.* (2009) had to send repeated email surveys to the researchers in question to get an accurate picture of statistics used to calculate published p-values. It was clear there was a problem with transparency in the neuroimaging literature.

Open science, as a practice, represents an effort to address the replication crisis by increasing access, awareness, and transparency. The Center for Open Science maintains the Open Science Foundation (OSF), a data-sharing repository and toolbox where researchers can share their experimental results with the public free of charge (Center for Open Science, 2021). The stated goal of the Center for Open Science is "…a future scholarly community in which the process, content, and outcomes of research are openly accessible by default" (Center for Open Science, 2021). The central philosophy is twofold. First, if research is shared openly rather than siloed in academic journals with exorbitant subscription fees, both the general public and the research community will be enriched. In this sense, it is an effort to democratize science and eliminate the unequal financial barriers inherent when access to knowledge depends on access to journals. Second, methods and results shared on OSF are not limited by publishability. Because the volume of research produced is far greater than the capacity of academic journals to publish, novel or buzzworthy results are at times prioritized above the quality and rigor of the research. Moreover, institutions often judge researchers on where and how often they publish rather than the content of their research itself. Open science (theoretically) creates a level playing field where experiments and conclusions can be judged on merit by all.

To foster open science in the field of neuroimaging, Poldrack and colleagues founded the Stanford Center for Reproducible Neuroscience (SCRN) in 2015 (Stanford CRN, 2021). The goals of SCRN align closely with those of OSF: using computational tools and data-sharing resources to make research more reliable and accessible. One of SCRN's most tangible contributions to open

data in neuroimaging has been the adoption of the Brain Imaging Data Structure (BIDS) format (Gorgolewski *et al.,* 2016). BIDS is a standardized file organization structure for sharing neuroimaging data. Any lab with data to organize inevitably develops idiosyncrasies in file naming, directory structure, and general organization. Data-sharing between smaller labs and individual researchers becomes laborious as one group tries to decipher idiosyncrasies of the other. However, through BIDS Apps (Gorgolewski *et al.,* 2017), neuroimaging data can be standardized into BIDS format, making it easier to share across groups and institutions.

The introduction of BIDS format was motivated by the structure of SCRN's open data repository. Known as OpenfMRI when founded in 2017 (Poldrack & Gorgolewski), it was a repository of task-based fMRI data allowing anyone to access, download, and analyze the data at any time, for free. Data was distributed under a Creative Commons license that allowed the resulting analyses to be published provided the group originating the dataset was acknowledged.

Now known as OpenNeuro, the repository has expanded to include EEG, iEEG, MEG, ECoG, ASL, and PET data. All data is organized in BIDS format, which has since been adapted for use on formats other than fMRI (Holdgraf *et al.,* 2018; Pernet *et al.,* 2018; Niso *et al.,* 2018). Any research group can upload their data to OpenNeuro if it is BIDS formatted and conforms to a simple set of guidelines. Anyone with an internet connection and simple command line skills can download and analyze data from OpenNeuro at any time, free of charge. The SCRN also maintains NeuroVault (Gorgolewski *et al.,* 2015), a similar repository of uploaded and open statistical maps, and NeuroSynth, a keyword-searchable tool that aggregates maps from NeuroVault for use in meta-analyses and ROI identification.

OpenNeuro serves the goals of open and reproducible neuroscience. Uploading data whether or not it resulted in a publication, or even a significant result, furthers appreciation of the

breadth of research being conducted in the field – both published and unpublished. Data-sharing also increases visibility of null findings and allows researchers to check each other's work or run more rigorous analyses on data of interest. Such work can even lead to collaborations. All of this contributes to a more transparent, accessible, and reliable body of neuroimaging research.

However, a growing number of researchers now use OpenNeuro not as a source for replication, but as a source of pre-existing data to test and develop novel methods. In their 2017 outline of the project in Neuroimage, Poldrack & Gorgolewski explicitly positioned OpenfMRI as a complement to big-data projects like HCP:

> *"In recent years, there has been movement towards the sharing of large, well-curated, and highly general datasets such as the Human Connectome Project. The goal of OpenfMRI is to provide a complementary venue for every fMRI researcher to publish their task fMRI data, regardless of the size or generality of the dataset." (Poldrack & Gorgolewski, 2017)*

OpenNeuro is becoming more and more of a complement or even an alternative to large datasets like HCP or ABCD. Researchers validating proposed new methods of statistical analysis or data processing have found a rich source of material on OpenNeuro. The broad selection of tasks, experimental designs, and scanning protocols is a contrast to large but uniform datasets like HCP. While a given dataset on OpenNeuro may have 15 scans instead of 1500, the specifics of the experimental conditions may make the smaller dataset uniquely suited to a novel question or analysis tool (For example, see Coutanche & Thompson-Schill, 2013).

The diversity of data on OpenNeuro makes the platform a unique and invaluable tool for designing, validating, and implementing new methods of fMRI analysis. OpenNeuro is rapidly becoming a crucial part of neuroimaging research. According to a 2021 report on OpenNeuro from Markiewicz *et al.*, the authors were able to identify 165 publications that reused data acquired on OpenNeuro, including 112 journal or conference papers. The same report shows that data reuse experienced a massive increase in 2020, nearly tripling between 2019 and 2020 (Markiewicz *et al.*, 2021). Data were used for diverse applications including basic neuroscience, human cognition research, clinical research, modeling, and methods development (Markiewicz *et al.*, 2021), with many publications comparing two or more datasets (Markiewicz *et al.*, 2021).

One of OpenNeuro's greatest strengths is flexibility. The open-source platform has been designed to place minimum burden on sharers and users of open data by making it simple to format and upload a dataset or download one for your own use. However, the relative ease of sharing and acquiring data also open the platform up to specific weaknesses from the perspective of an end-user hoping to conduct secondary analysis on a shared dataset.

1.2 The Problem: Documentation for Small-Scale Secondary Analysis

The relative ease of data-sharing on OpenNeuro results in a platform with a broad array of data on a plethora of topics. Available fMRI studies feature data from participants of all ages, with a wide selection of experimental tasks and designs. Researchers can find functional data investigating simple visual processing in healthy adults alongside longitudinal data investigating schizophrenia in specific age groups. By design, there is a relatively low bar to data-sharing on OpenNeuro. Data must be BIDS-formatted, de-identified, and in compliance with a simple set of

ethics and privacy guidelines. Data are uploaded via a simple command-line tool and can be easily transformed into BIDS format using BIDS Apps (Stanford Center for Reproducible Neuroscience, 2021). A simple approach means that more groups can share data and the repository grows every day. However, it entails obvious problems for data users and data sharers.

A dataset may meet OpenNeuro's standards, but those standards are designed to be a minimum. OpenNeuro has few concrete requirements for documentation or quality assurance, to a large extent leaving it up to the sharer to decide how much information about the experiment and the resulting data to include along with the dataset itself. OpenNeuro also provides little guidance on which files and images from a project should be included when uploading a dataset. It is left to the sharer to be familiar enough with the data and to decide which of the many pieces of information captured during any study are pertinent. Those who are uploading data in alignment with the goals of open science – ensuring that their data are preserved such that their results can be replicated in the future – may not go into such a decision with secondary analysis in mind. The amount of information necessary to replicate the findings of a specific experiment are not the same as that necessary to adapt the data for a new question, novel analysis method, model, or proof of concept. As such, there are many datasets provided on OpenNeuro that will meet the requirements of the platform and the requirements of replication, but not the secondary analysis requirements of a given user. To date, little practical guidance has been provided to help both users and sharers on OpenNeuro avoid pitfalls to work effectively with the platform and each other.

This document represents an attempt to bridge that gap by offering insight, recommendations, and concrete guidelines for both sharers and users hoping to use OpenNeuro to enable secondary analysis. Focus here has been placed on fMRI data shared on OpenNeuro. However, these issues are by no means unique to the platform. Other data-sharing platforms are

subject to the same concerns. Platforms in the vein of OpenNeuro include Dryad Data (Dryad, 2021), which includes neuroimaging data among many other types of open scientific datasets, and NITRC (NITRC, 2007) which hosts big-data resources like HCP and ABIDE but also allows individual groups to upload and share results with an account. Many labs elect to share their data along with code and documentation via GitHub. Every platform has unique issues, but central concerns remain the same: while it is beneficial for upload guidelines to be open-ended on open-source platforms, the resulting variation in the quality and documentation of available data becomes a source of confusion and frustration.

Many standards and practices are available for the use of open neuroimaging data. The Organization for Human Brain Mapping (OHBM) has published a comprehensive set of data-sharing standards and guidelines for both large and small datasets (Nichols *et al.,* 2017). Scholars at the SCRN have published numerous recommendations and guidelines (Poldrack & Gorgolewski, 2014; Pernet & Poline, 2015; Poldrack *et al.,* 2017), as have other open science advocates. These recommendations are primarily aimed at the goals of open science, intended to increase reliability and replicability through open data access. Little attention is devoted to the practice of using smaller datasets for novel questions, or to those researchers wishing to share smaller datasets with others for that purpose. More attention has been devoted recently to logistical questions when it comes to large datasets like HCP (Horien *et al.,* 2020; Laird, 2021). However, even now, smaller datasets are often left out of the conversation, especially when it comes to low-level, practical advice for beginners.

This is a new problem, but it will only become more relevant. As computing power increases and the scope of neuroimaging analysis widens, the need for shared data will continue to increase. Available data are always limited, but future computational exploration is not.

1.3 In Conclusion

The examples and guidelines offered in this document will form a base of knowledge for anyone who wishes to enable secondary analyses, whether when sharing their experimental data on OpenNeuro or when benefitting from others' results. The appendices contain checklists for both sharers and users. In combination with the insights and recommendations offered below, these checklists can be used by researchers setting out to consider secondary analysis on OpenNeuro for the first time. The goal of this document is not to give a comprehensive tutorial of fMRI analysis methods or the specifics of open data, but rather to prepare researchers on both sides of a secondary analysis to hit the ground running when it comes to uploading data to OpenNeuro or seeking data to download for a new project.

Having read this far, experimental researchers who are accustomed to sharing their data may be wondering whether their efforts are sufficient to enable secondary analysis and whether it is worth the extra time and labor to ensure that their data meet standards beyond replication. The answer is not the same for every group, but there are features that make secondary analysis a worthy goal even if extra effort is required.

2.0 Mission Statement: Why Should I Share My Data?

Open data has two related but distinct goals: replication and secondary analysis. As an experimental neuroimaging researcher, you may be motivated to share your data because you are dedicated to the goals of open science. You may be justifiably concerned about opacity, lack of access, and the persistence of the File Drawer Problem in a field where high-profile journals have become arbiters of academic success. Alternately, you may be abiding by requirements or recommendations to share data. As journals and funding bodies take steps to address a paucity of replicable results, more and more have begun incentivizing open data. Over 75 scientific journals, including *Neuropsychology*, *Cognitive Science*, and *Cortex*, have joined the Center for Open Science in providing badges to researchers who meet open data and preregistration standards (Center for Open Science, 2021) which has proven effective (Kidwell *et al*., 2016). Experimental studies funded by the National Institute of Mental Health (NIMH) are now required to share data on the NIMH Data Archive (NDA), which includes a number of sub-databases (NIMH Data Archive, 2021). Whether for ethical reasons, rewards, requirements, or all three, data-sharing may already be a part of your plan for publication.

However, when you set out to share data for these reasons, you likely have replication in mind. Your focus will not be on whether your uploaded dataset is clear and useful for secondary analysis, but whether it is sufficient to contextualize and reproduce the results of your experiment. These two standards can differ substantially. As you consider the end-user challenges and recommendations in this document, you may ask whether you have the time and labor available to devote to ensuring that your dataset is suitable for secondary analysis in addition to replication when you upload it to OpenNeuro or other platforms. There is a difference between a lab or

individual setting out to share experimental data and the consortiums like HCP, ABIDE, or ABCD that collect data to share for secondary analysis. With their large base of funding and resources across multiple institutions, consortiums and their associated labs can hire staff dedicated to data-sharing. Research staff can be paid to dedicate full-time or part-time labor to formatting data for sharing, ensuring that it is consistent with established standards, and offering support or updates over time. A smaller lab conducting a smaller-scale study with their own limited funding may not have the same resources. The work of adapting the large body of data acquired during an experiment into an uploaded dataset must be accomplished by researchers or staff with competing demands on their time. Given that extra time or labor will be required of them, above and beyond what is currently invested in data-sharing for replication, why is the additional goal of secondary analysis worthwhile?

In truth, there are several benefits to sharing data for secondary analysis. Much like open science overall, open data benefits the entire field in terms of ethics, equity, and access. In addition, the practices you will implement to prepare data for secondary analysis will benefit your research – investment in time and labor now will pay dividends in the future.

2.1 Open Data Democratizes Science

fMRI is highly exclusive. In 2021, typical 3T functional MRI scans cost $500 - $600 per hour. In addition to the cost of scanning, participants must be compensated for their time, and staff or students must be compensated for their work. Many labs and individual researchers simply do not have access to the scale of funding necessary to conduct a full-scale fMRI study. Even if an individual investigator has the resources to fund data collection, scanners are prohibitively

expensive for most institutions. The Siemens Magnetom Prisma, a widely used 3T functional scanner, costs on average 1.6 million – 2.2 million dollars in 2020. Even a used model costs between 900,000 and 1.4 million dollars (DirectMed Parts & Service, LLC, 2021). Running and maintaining a scanner costs additional tens to hundreds of thousands of dollars per year. It is unsurprising that most fMRI scanners are located at large research centers, R1 universities, or hospitals (Laird, 2021). Scanners are already in high demand from researchers within these institutions. Those located too far from the nearest scanner may not have resources or time to travel there. Researchers at smaller institutions may never be able to conduct scans at all. It is easy to see how financial and demographic features cut crucial diversity and depth out of the field, drastically limiting the amount of novel work. There are further implications for participants, contributing to unwanted biases in race, age, gender, and socioeconomic status, among other demographic factors (Laird, 2021).

Such struggles and biases have only been exacerbated by the COVID-19 pandemic. Throughout 2020, many researchers across the spectrum of research in STEM, social science, and the humanities were forced to postpone or abandon data collection in the interest of safety. Financial struggles at universities across the world left less funding available for students, staff, and the material needs of experimental research. A transition to remote work put enormous strain on the systems and people that enabled experimental research. As with most pandemic-related phenomena, this trend disproportionately affected students, staff, and faculty who were already marginalized by race, gender, immigration, and socioeconomic factors (Staniscuaski *et al.*, 2021).

In their 2021 report, Markiewicz *et al.* note that collecting the reused data they identified from scratch would have required more than 21,000 scans of individual participants (Markiewicz *et al.*, 2021). They estimate a total savings of almost 21 million US dollars. For those without

adequate funding or scanner access, the availability of free, open data represents a chance to conduct novel research that would otherwise be impossible. It also represents a savings to researchers who would otherwise need to devote months or years to collecting the data necessary to test a new analysis method before even knowing if the results would be reliable. Secondary analysis also enables techniques that would be near-impossible without shared data; meta-analyses, novel statistical or computational methods, and computational modeling all draw heavily on shared data. When smaller-scale datasets are shared, secondary analyses of this nature have access to a wider selection of specific, relevant results that may be uniquely suited to their needs. In a very real way, data reuse has opened the door to a broader, more complex, and more diverse body of neuroimaging literature.

Markiewicz *et al*. (2021) also note a "sharp increase over time" in the use of OpenNeuro data in published projects, including preprints, journal articles, conference presentations, theses, and software development. While OpenNeuro data has continued to increase in reuse since 2018, when the platform changed from OpenfMRI, data reuse nearly tripled between 2019 and 2020, and as of June 2021, almost 60 publications derived from OpenNeuro data had already been identified (Markiewicz *et al*., 2021). This is far from coincidental. As the pandemic forced scanning centers to close and researchers to work remotely, shared data remained as a resource for those forced to halt their own work. Even as the latter half of 2021 brings a return to in-person research for some, the aftereffects of the pandemic – material and psychological – are still palpable and will linger for decades to come. Sharing data for secondary analysis is a long-term investment in a field that is becoming more collaborative than ever before.

If you are invested in open science, you should be invested in open data. Secondary analysis is a clear step toward the future that the Center for Open Science envisions, in which barriers to access are minimized and scientific contributions are judged solely on merit.

2.2 Open Data Creates Connections

The COVID-19 pandemic has proven just how important connection and collaboration can be. Neuroscience and psychology have become more focused on collaboration across institutions, disciplines, and national borders. The open data landscape presents an opportunity to create connections with the broader neuroimaging community, as well as opportunities for collaboration.

Replication is still the most common goal of data-sharing. As such, smaller groups who make a point of sharing data that encourages secondary analysis will stand out. You will build goodwill on the platform and in the research community if you enable novel projects with your own work. Reuse can also foster collaboration; if you provide useful data and are responsive to users, you may find that you are presented with opportunities for authorship on publications and collaboration on future projects. An investment in secondary analysis is also an investment in your network and your place in the neuroimaging community.

Secondary analysis also yields visibility and citations for your work. Datasets that are reused will be cited in every publication built on the original project. Associated publications from the original researchers will also garner citations when they are drawn upon for reference. Taking the time to follow the guidelines provided in this document could improve the utility of the data you share, yielding more citations and more visibility for your work. Novel models or analyses built on your data may inspire others to seek it out for their own work. Beyond citations, the

secondary analyses your results inspire may recontextualize them for you. Observing others' use of your data may generate new questions, inspiring you to do a secondary analysis of your own.

2.3 Open Data Preserves Your Work

As you will see in the next sections of this document, there are multiple best practices to implement if you wish to share your data with the goal of enabling secondary analysis in addition to replication. Most are focused on making the dataset clear and comprehensible to a user who does not have access to your context and experience. They may seem like unnecessary extra labor when uploading your data, as you will often be updating or including information that was captured but not relevant to replication. However, imposing these standards on your data – and any data collected in the future – will have long-term benefits for your group. These standards will ask you to compose clear documentation about how the data were collected and which decisions were made. They will ask you to preserve all the information you captured during a study – not only that information used for a publication – in a readable, accessible, and well-organized format.

Consider the last time you joined a new group and tried to work with previously collected data. Most researchers will be familiar with the "telephone game" that occurs when the person most familiar with a body of data moves on to another project. Their knowledge of the structure and content of the data is lost, leaving others to wade through the files and impose their own order. The same can be true when the original investigator leaves off working with their own data and comes back to it after some time. Whether the data are being subjected to new analyses by a graduate student, re-evaluated by an institutional colleague, or revisited after a long period of peer

review to address reviewers' comments, it is likely that the experience of attempting to decipher the results will be very similar to that of a new user on OpenNeuro.

The standards of secondary analysis will minimize your "telephone" problem by forcing your group to develop and enforce a consistent system for organizing, naming, and formatting your data, a consideration that can fall by the wayside even in the most productive labs. Further, OpenNeuro's privacy standards ensure that your participants' identities will not be at risk if their data are re-analyzed.

You may even find that having all the information readily available and stored on OpenNeuro's servers in BIDS format provides a useful backup to lab-dedicated file storage. If you wish to share your data with colleagues or students in the future, it will be preserved and ready to download. If you choose to revisit your data for further analyses, you will find it organized, documented, and secure for years to come.

2.4 In Conclusion

Having taken all these factors into account, you may still weigh the amount of work required against the benefits and decide against the extra steps necessary to enable secondary analysis. However, if these concerns resonate with you, it is very likely that you will see the long-term benefits of investing that extra labor for you, your lab, and the field of neuroimaging.

The following sections will offer an end-user's perspective on shared data, along with recommendations for both sharers and users to optimize the benefits of secondary analysis. Following the general recommendations, you will find concrete guidelines for uploading and downloading data with a focus on four commonly used basic fMRI analysis techniques. If you are

setting out to share or use OpenNeuro for secondary analysis for the first time, you may use this information and the supplemental checklists provided as appendices to guide you through the process.

3.0 End-User Issues: Case Study

From the first-time user's perspective, the OpenNeuro user interface (UI) offers a minimal but straightforward experience. Every dataset is presented with a file tree organized in BIDS format, a readme provided when the data were shared, and any information about citation or associated publications that the sharers have provided. An update history is provided with a snapshot of each version of the dataset and a warning banner clearly indicates if data are not BIDS-compliant. Each dataset is accompanied by a comments section where users can post questions or requests. The first step to proficiency, for both the user and the sharer, is to navigate and become conversant with the UI. Conducting a few searches and examining available datasets will illustrate how OpenNeuro's commitment to remaining flexible, straightforward, and open-source is both a strength and a weakness. A plethora of data is available from diverse tasks, methods, and research topics. However, there is little consistency in the amount of information provided along with uploaded datasets, resulting in a few characteristic issues when considering data from the perspective of a secondary analysis.

Both users and sharers should be aware of these common pitfalls, which can be divided broadly into three categories: concrete analysis plans, documentation, and completeness of the dataset for the purposes of secondary analysis.

As an illustrative example, these boxes will present examples of datasets considered for a secondary analysis conducted in response to reviewers' comments on a proposed novel analysis technique submitted to NeuroImage in 2020. The exact details of the method and the original proof of concept are not germane to this discussion, but a few key details are important to keep in mind.

The proposed method utilized multivariate connectivity built on Representational Similarity Analysis (RSA). It assessed trial-by-trial change in the neural pattern similarity, passing a spherical searchlight across the brain and correlating a similarity metric between each searchlight and a seed region.

In response to reviewers' concern that the method could be measuring noise rather than true change over time, we designed a confirmatory analysis that was meant to assess results in a case where the ground truth should be unambiguous.

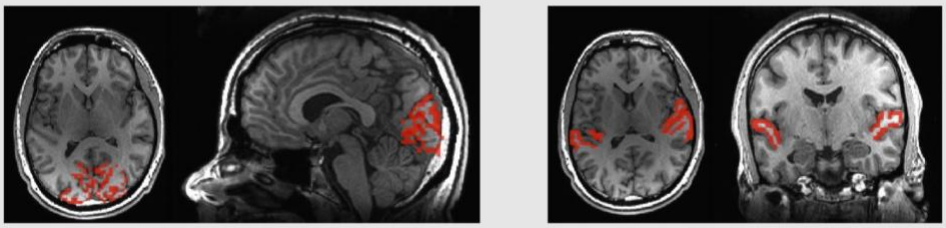
Figure 1. Case Study - RSA

3.1 User Considerations

3.1.1 User Analysis Plans

Not all preparation for secondary analysis is incumbent upon the sharer. Users will encounter frustration and wasted time if they embark on a novel analysis project without concrete plans for their data search, their processing steps, and their final analysis.

We proposed using the method with two low-level perceptual seed regions: V1 and primary auditory cortex.



Primary Visual ROI Primary Auditory ROI

The confirmatory analysis was intended to draw from a dataset in which we could compare brain activity elicited by visual stimuli with no sound (**visual-only**) with brain activity elicited by auditory stimuli with no visual input (**auditory-only**). We reasoned that if the method was sound, there would be a distinct difference in both cases when contrasting low-level visual and auditory ROIs. First, we sought to find a dataset in which there were visual-only and auditory-only conditions presented within the same experiment.

Figure 2. Case Study - Proposed Analysis

A secondary analysis project commences with a "shopping" phase, in which users conduct a systematic search for data on OpenNeuro, considering several datasets and deciding which data will ultimately form the material for a new analysis. Any search for data on OpenNeuro should begin with a "shopping list" of necessary features. Every analysis relies on a unique combination of "moving pieces" and will require a specific set of information to produce results. Users should consider not only the images and supplementary files necessary for the analysis but also the

exclusionary factors that would make a dataset unsuitable. The checklists provided here are designed to aid this process, which can begin with an assessment of the low-level processing and analysis tasks that form the basis of a novel analysis. Consider how data, both structural and functional, must be pre-processed. If your analysis relies on specific ROIs, ensure that there is sufficient coverage in the data to allow those regions to be defined in every brain.

Planning is the most important phase of the search for a dataset. Based on the needs of the proposed analysis, we developed a "shopping list" of factors that needed to be in place in the data we used. The initial list was simple:

- Functional and structural data
- Auditory-only and visual-only conditions to compare
- Stimulus timing files

The first dataset considered was the fMRI portion of an auditory/visual oddball task from Walz *et al*. (2018), which has been downloaded from OpenNeuro over 600 times. The task contained both auditory-only and visual-only blocks, each with a standard and an oddball stimulus. However, at least four categories of stimuli for each modality would be necessary for RCA. The oddball task offered only two conditions per modality.

We updated our list to include further necessary factors:

- At least four stimulus categories per modality
- Stimulus timing files that identified stimulus categories for each trial

Figure 3. Case Study - Shopping List

Users should also consider which preliminary statistical tests can be performed to ensure that the data selected will yield results in a novel analysis. As many datasets uploaded to OpenNeuro are derived from small-scale studies, they may not contain enough data to yield a

meaningful effect. Consider both the number of participants and the amount of data collected during each scan. A power analysis can inform the user's choice of datasets, and a sufficient quantity of data should be an element of the inclusion/exclusion criteria established during the "shopping" phase. Specific analyses may call for a certain number of experimental conditions or trials. Other statistical tests may be specific to the basic methods upon which the new analysis is built – an analysis built on MVPA, for example, depends on machine learning, so training and testing a basic classifier on key ROIs can indicate whether there is enough distinction between experimental conditions to allow for further, more complex analysis.

After eliminating the oddball dataset, we considered supplemental data from the Forrest Gump study (Hanke *et al*., 2018). This dataset is one of the most widely-cited on OpenNeuro. Although data was collected for the purposes of secondary analysis, the dataset contains only 20 subjects. The most frequently applied and cited portion of the data was collected while subjects watched a complete German-language version of the film *Forrest Gump*, followed by an audio-only version of the same film. These data have been provided on OpenNeuro as well as the study website and other platforms, including GitHub. In addition to the naturalistic audio and video portions, further tasks have been added as supplements to the main dataset.

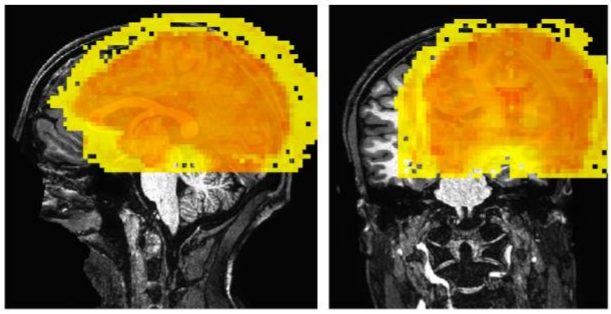
Two supplementary datasets seemed like an excellent fit for our purposes. In Sengupta *et al*. (2016), participants completed a visual area localizer in which they viewed grayscale images from six stimulus categories. In Hanke *et al*. (2015), participants passively listened to music clips from five musical genres. Because there was significant overlap between participants who completed the visual localizer and participants who completed the music listening task, auditory/visual cortex results from these data seemed ideal to compare. However, upon examining the data, it became clear that there would be challenges inherent in adapting them for our analysis.

Figure 4. Case Study - Forrest Gump

Finally, users should take the time to reflect upon their own technical proficiencies and limitations. Having decided on pre-processing steps, preliminary statistical tests, and the key parts of the novel analysis, consider each dataset through the lens of your technical skills. An

experimental paradigm may seem like a perfect match for your question, but the resulting data may or may not be in a state that will allow it to be adapted for your purposes. If the technical demands of preparing data for analysis are exceptionally time-consuming or far outside your comfort zone, it may be more productive to search for a new dataset. The utility of the data in terms of the original experimental paradigm must be weighed against the time and labor required to prepare and analyze it productively.

One issue arose while attempting to adapt our typical pre-processing pipeline for the visual localizer data (Sengupta *et al.,* 2016*)*. Only one structural MPRAGE was provided for each participant. The structural scan had been acquired during each participant's initial session, when they viewed the film. The localizer was acquired in a different session, in some cases a long time removed from the initial scan. As a result, the structural data and the functional localizer data were far out of alignment, presenting a more difficult task when it came to automating struct/func alignment during pre-processing.



Overlay: Visual localizer functional data for sub-01
Underlay: Structural scan for sub-01

While this was not an insurmountable issue, the time taken in revising, testing, and troubleshooting pre-processing set back the process of developing and troubleshooting the analysis itself.

Figure 5. Case Study - Pre-Processing Demands

3.2 Sharer Considerations

3.2.1 Documentation

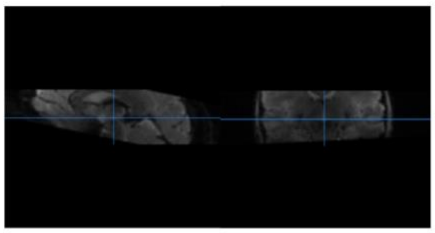
There are few hard and fast requirements for documentation on OpenNeuro. A .json file with basic information must be uploaded along with the data, but the content of most documentation – particularly the readme text included on the dataset homepage – is left up to the sharer. The readme provides the user's first impression of the dataset and may be the initial deciding factor in whether to continue. Including clear, sufficient detail is imperative. Of course, the explanatory text sufficient for users seeking to replicate an analysis will not be equivalent to that sufficient for users seeking data for secondary analysis. Extra considerations are necessary when considering what documentation to provide with secondary analysis in mind.

For replication purposes, one may assume that the user is familiar with the original results, how they were derived, and their implications. For the purposes of secondary analysis, the user is most likely looking at the content of this experiment for the first time. As such, documentation about the design and protocol for the experiment and for each task should be provided. Comprehending the nature and motivation of task design – the nature of stimuli, the protocol for each session, the structure of trials and runs – may not be strictly necessary for replication, but it is necessary for adapting the data to a new question. This is especially true if experimental design must be considered when creating new processing and analysis pipelines.

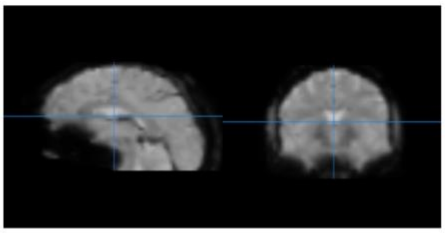
Forrest Gump's OpenNeuro documentation has been retained from the original upload to OpenfMRI (Hanke, *et al.,* 2018). The original data are well-documented, but little is provided on the supplements. The readme offers this information:

"The dataset… contains data from the same twenty participants while being repeatedly stimulated with a total of 25 music clips, with and without speech content, from five different genres using a slow event-related paradigm… there are additional acquisitions for fifteen of the the twenty participants: retinotopic mapping, a localizer paradigm for higher visual areas (FFA, EBA, PPA)…" (Hanke, *et al.*, 2018)

Auditory/visual cortex results from these 15 participants seemed ideal to compare. However, upon examining the functional data on OpenNeuro, it was clear that the two datasets were not comparable:



Listening data for sub-01



Visual localizer data for sub-01

We examined the study website and associated publications for both supplementary datasets. The visual localizer was acquired at 3T, while. The music listening task was collected at 7T. Fewer slices were collected at the higher resolution – enough to cover key auditory regions, but not enough to encompass our visual ROI.

Figure 6. Case Study - Documentation 1

A common technique is to provide minimal information in the readme but refer users to resulting publications for details about experimental protocol. This is understandable if the goal of data-sharing is replication, which entails prior familiarity with the results. However, information

shared in a publication is necessarily limited to the details that give context to the reported results. If details of protocol are irrelevant to the results at hand, it is only sensible to omit them from the methods provided. However, such details may be germane to any number of novel analyses or new questions. Further, the user must weigh the potential utility of the data against the labor of searching through one or more publications to find the parameters necessary to complete their analysis.

In a data set by Notter, Da Costa, & Murray (2019), participants were exposed to a visual-only condition and an audiovisual condition. We reasoned that a comparison would yield distinct patterns of activity across our two ROIs. The authors provide an abstract and a description of the experimental paradigm in the readme. This information initially led us to select this dataset.

Visual stimuli were either drawn from an existing stimulus set or adapted from an online library, but specific information on stimulus categories is not offered. This appears to be a case in which the authors planned to refer to the resulting publication to provide more detailed references and methods. This is referenced in the readme:

> *"The stimuli used in this experiment were the same as in Lehmann & Murray (2005). On each trial, subjects indicated whether the visual stimulus was appearing for the first time or had appeared previously… The experimental paradigm is schematized in the corresponding paper, Figure 1." (*Notter, Da Costa, & Murray, 2019*)*

As far as we could discern, the study was never published. The "corresponding paper" is never identified in the documentation, and in the "How to Acknowledge" section, the authors have listed the paper as "TBD". The data were never published on, so crucial context was lost. We had a choice of either attempting to identify the stimulus categories from the references provided and the vague timing files or abandoning this dataset. We elected to move on, changing our analysis plan.

Figure 7. Case Study - Documentation 2

Another element of documentation is quality control. Errors, artifacts, and persistent issues have most likely been tracked during data collection. However, that information may be superfluous or incidental to the original analyses, or a lab may find it preferable to provide that

information as needed by request. However, if data quality documentation is not provided along with data, a user may make the mistake of assuming the data is of uniform quality and free of systematic issues. This will have consequences later, when data quality issues affect subsequent processing or analysis steps, perhaps even to the point of biasing the results of the secondary analysis.

Overall, documentation is one of the main barriers to productive secondary analysis, but it is also one of the easiest to address.

Sharers should look at their data from the perspective of an outsider, a potential user who is considering the data for the first time. What information will they need to comprehend the experiment and the resulting data, even if they don't have the publication in hand? When the data is considered through this lens, it may become clear that more information must be offered than is necessary to replicate the analysis. However, the extra effort of crafting comprehensive documentation will pay dividends later.

Users should be ready to use everything at their disposal, recognizing that documentation may be lacking. Take a thorough look through the readme and all additional documentation provided with the BIDS-formatted data before downloading. If the necessary details are still unclear, be prepared to locate associated publications. If the data seem like a good fit but were not published on, consider contacting the authors for documentation. If authors have not provided documentation, do not offer associated publications, and are not responsive to comments or communiques, this may serve as a good indication that the dataset should not be subjected to secondary analysis.

3.2.2 Completeness for Secondary Analysis

A great deal of information is captured during any experimental study. Beyond data acquired during scans, a full dataset may contain such supplementary material as subject demographics, behavioral or psychophysical data, scanning parameters, or stimulus sets. Not all the information captured and stored will be relevant or useful for replicating results; some may be secondary, not required to replicate the experiment, and some may be collected but later deemed unhelpful or irrelevant to the question at hand. When uploading a dataset to OpenNeuro, it may seem intuitive to include only that supplementary information necessary to replicate reported results. However, information superfluous for replication may be vital to adapting the data for new purposes.

Overall, supplementary files – clear and detailed event timing, stimuli or stimulus information, participant demographics, and scan parameters – are a prerequisite for secondary analysis. Although a dataset may be complete for the purpose of replication, when secondary analysis in mind, it cannot be considered complete without these elements.

Perhaps the most important information to offer along with fMRI data are clear, detailed event timing files. While timing files may be required to replicate an analysis, the level of detail required for replication is different from that required for secondary analysis. For replication purposes, it may be sufficient to label events with numbers or arbitrary codes. If stimuli are identified, the sharer may choose to code only the experimental condition of the stimulus without identifying the item in a way that a new user will understand.

Stimulus category was important to our analysis. The number of categories and the number of presentations could determine whether the data would elicit a signal. Knowing the category presented at each TR was also vital for automating the analysis. We considered several datasets that, while complete for the purposes of replication, fell short for our purposes when it came to stimulus timing files.

Notter, Da Costa, and Murray (2019) present their timing files in this format:

ONSET	DURATION	CONDITION	TARGET	RESPONSE	ACC	RT	STIMULI
5.0	0.5	3	1	-1	0	0	17
16.0	0.5	1	1	1	1	822	2
23.0	0.5	3	1	1	1	558	22

Stimuli are identified only by number, and no information has been provided about the correspondence of numbers to images. Possible solutions could be adding more descriptive identifiers in the "stimuli" column; including the stimuli with the data set; or including a key that matched the number identifiers to stimulus categories.

Figure 8. Case study - Timing Files 1

Vague identifiers like file names or numbers will suffice when the stimuli are readily available and well-known to the group analyzing the data but will not provide enough information for new users to identify the features of stimuli presented at each trial. As discussed later, some of the most common fMRI analysis tasks depend on knowledge of stimulus features. Programming any novel analyses based on BOLD activation or multivoxel pattern analysis (MVPA) will almost certainly require knowledge of the stimuli presented during an experiment beyond their

experimental condition and onset time. Stimulus features may also form an element of connectivity-based analyses that analyze trial-by-trial changes. If a corpus of images is used, for example, novel analyses will need to identify which individual image was presented on a given trial. BIDS format allows for a stimulus set to be included with the data, and this can be enormously helpful to secondary analysis, especially if custom stimuli were created for the experiment. If a publicly available stimulus set is used, including a link to the original source in the documentation will prove useful for most secondary analysis.

In Pernet *et al*. (2019), participants listened to vocal and non-vocal sounds. Vocal sounds could be emotionally loaded, emotionally neutral, or distinct speech; non-vocal sounds came from animals, environments, man-made objects, or music. The array of categories seemed ideal for the auditory portion of our analysis. The timing files provided with the dataset come in this format:

onset	duration	trial_type	stimuli
12	8	nonvocal	sourcedata/stimuli/non_vocal_01
22	8	vocal	sourcedata/stimuli/vocal_01
32	8	nonvocal	sourcedata/stimuli/non_vocal_02

This would be enough to replicate the original analysis, but the fine-grained stimulus categories are not identified. These file names may have sufficed in the original lab, where the stimuli were accessible and familiar, but they made our analysis impossible to automate. Possible solutions could be adding more descriptive file names; adding columns to the timing file; or providing either a key to the stimulus identifiers used in the timing file or the stimuli themselves.

Figure 9. Case Study - Timing Files 2

Most studies capture some information about subject demographics, whether to inform analysis or to provide to a funding agency. In many cases, this information is of no direct relevance to the experiment. If a study does not deal with demographic factors, it is unlikely that demographic information will prove useful for replication, and formatting and uploading a file of subject demographics may seem like a waste of time. However, one of the more common forms of

secondary analysis involves comparing multiple datasets, most often collected during different experiments for different purposes. This may take the form of a meta-analysis. Alternately, testing a novel method may require comparing different modalities – a common test is to compare tasks targeting visual regions with tasks targeting auditory regions (see, *e.g.*, Sadoun *et al.*, 2020). When comparing brain data collected from multiple samples, it is crucial to perform at least some degree of matching between subjects. At the bare minimum, the distribution of participant age and gender should be preserved across datasets. If subject demographics are not provided, matching becomes difficult to impossible, and datasets that would otherwise form an informative contrast cannot be compared.

We decided to compare data from two studies – one that used only visual stimuli, and one that used only auditory stimuli.

Comparison requires matching. Conclusions drawn from comparing studies with a different number of participants, and/or participant populations that are not comparable, will not be reliable or generalizable. If two studies were conducted at different times with entirely different samples, matching is crucial. At a bare minimum, the data compared should include the same number of subjects, matched on age and biological sex.

Although Pernet *et al*. (2019) note the mean age and number of males in their sample in an associated publication (Pernet *et al*., 2015) indicating that demographics were collected, no participant demographics have been included with the OpenNeuro data. Providing the demographic data captured during the study – even if minimal – would have been another small change that increased this dataset's utility for future analyses.

Figure 10. Case Study - Participant Demographics

Finally, every fMRI study commences with a discussion of scanning parameters. How will resolution be balanced with efficiency? Will acquisition be interleaved? How will alignment be performed? Which areas of coverage are most important, and which outlying brain regions can be cut off or distorted without affecting results? How will motion be thresholded and corrected? The answers to these questions, and many more, will be unique to the study. These parameters are most likely recorded for use when processing the data, as when performing slice timing correction,

alignment, and segmentation. It may seem like too much extra labor to include information like slice timing/order or motion regressors along with the data. However, approaching the data with secondary analysis in mind, it becomes clear that scan parameters are an important element of a dataset. A new analysis may make use of derivative files produced during pre-processing, but it is more likely to start from scratch – processing the raw data with the user's own pipeline to prepare it for a novel analysis with a new purpose and specific needs. Data used for MVPA tasks, for example, is seldom blurred, while data used for BOLD activation analysis may be blurred to varying degrees. A user may wish to analyze the data in subject space rather than in template space, or vice versa. Processing data to align with new goals is the first element of a secondary analysis, and thus providing scan parameters along with raw data is a key element of ensuring data can be used for novel projects.

Just as when composing documentation, sharers should look at their data from the user's point of view. If you were setting out to process and analyze this raw data for the first time, what information would you need? Look at event timing files from a user's perspective and consider whether a new user will be able to understand them and incorporate them into an automated analysis without access to the experience and knowledge available to your lab. Include what participant demographics you collected, even if they were not relevant to your project, and take time to consider whether they are in a format that a new user will understand. Consider uploading your stimulus set or providing a link to the source of your stimuli. Finally, document your scanning parameters – knowledge of the choices you made will serve future users when they set out to construct new pipelines for your data.

Users should be aware of these important supplementary files and information. Keep in mind that datasets complete by the sharer's standards may not be complete for the purposes of

your analysis; do not trust that a dataset will be uploaded with all the necessary supplementary files, even if it has been published on in the past. In addition to the general requirements of supplementary files, consider the unique needs of your analysis. Be familiar with the basic building blocks: is this a connectivity-based analysis, which will require precise timing information? Is this an MVPA-based analysis, which will most likely required detailed information about the stimuli presented on each trial? Will subjects from different samples need to be matched for comparison?

Before searching for datasets, establish a "shopping list" for your analysis. Consider all the supplementary files that will be necessary beyond the raw data and basic event timing files. Establish inclusion and exclusion criteria for data. Look through the file tree, then make a rigorous check of the supplementary files provided before committing to downloading, processing, or analyzing any dataset. This is an idiosyncratic process, so flexibility on the user's part is important; when relying on 3rd party data, the ideal dataset for you may not exist. Be ready to change direction, if necessary, but remain cognizant of what will constitute a complete dataset for you.

3.3 Overall Recommendations

3.3.1 Users

Along with being familiar with the OpenNeuro interface, users must be well-acquainted with the demands of their proposed secondary analysis before beginning their search for data. Have a shopping list and a set of exclusion criteria established. Be aware that datasets complete and accurate from the perspective of replication may not be complete for your purposes. Ensure that all the moving pieces for your analysis are in place before committing to downloading and

processing data. Be flexible and remain open to changing your plans during the search. Once you have decided on a dataset, conduct preliminary statistical tests and pre-processing before committing to analysis. Early checks will minimize wasted time and effort, resulting in a more productive analysis that best addresses your unique research question or novel method.

3.3.2 Sharers

When secondary analysis is considered in addition to replication, sharers should look at their data through a user's eyes. Consider a dataset from the perspective of a new user setting out to analyze it for the first time without your documentation, experience, or previous publications in hand. From there, compose documentation that will give a user a positive first impression of your data. Consider the common types of supplementary files to include for secondary analysis and examine your work with an eye toward formatting timing files, stimulus information, participant demographics, and scan parameters so that they are informative to new users. Keep in mind that drafting helpful documentation and keeping supplementary files organized will benefit you in the future by creating a comprehensive, accessible record of the study and an organization system for your lab.

3.4 In Conclusion

Having read the notes and recommendations above, you have a framework of key considerations for users and sharers. You may now be wondering just how to get started when setting out to upload your data or search for a dataset for secondary analysis. In the following

section, we will expand upon these general recommendations to offer concrete guidelines to follow as you begin preparing your own project.

4.0 Concrete Guidelines for Secondary Analysis

The following sections will build upon the notes and examples offered above to offer concrete guidelines for working with shared data. These are accompanied by more specific considerations for four basic analysis methods: BOLD activation analysis, functional connectivity analysis (FC), multivoxel pattern analysis (MVPA), and representational similarity analysis (RSA). While computational exploration of shared data can take many innovative forms, these four techniques will most likely form the base on which any novel analysis is built. They can be seen as building blocks from which larger, more complex, secondary analyses and novel methods will be constructed.

Sharers need not have personal experience with a base method to upload their data in a form that will enable it. Users should be familiar with the base method(s) of their analysis before searching for data, using the concrete guidelines below to shape their expectations.

4.1 Guidelines for Users

4.1.1 Planning Your Analysis

Compose your shopping list before you begin searching for data. The checklists provided (Appendix B – F) will give you a basis, but you know your needs best. Map out the steps of pre-processing and analysis. Establish minimum criteria for a dataset and consider what would exclude data from consideration.

If your analysis will include specific ROIs – whether pre-defined or identified from the data – you will have an idea of where they fall on the average brain. This will determine the amount of coverage necessary in functional images and will also inform your exclusion criteria – major artifacts that affect your ROIs, such as susceptibility artifacts, will certainly be a concern. The same artifacts located in other parts of the brain may not present an issue.

Consider your desired participant population. Develop inclusion/exclusion criteria, just as you would for a study you designed yourself. Is there a maximum/minimum participant age you will include? Are there diagnostic criteria to be considered? Could demographic factors like race or sex constitute confounding variables?

Finally, consider how much data you will need to achieve a significant effect size. How many individual scans or participants will be necessary? How many trials and runs must be present to allow for your statistical tests to be reliable? More data are always better but try to establish minimum standards before searching for datasets. If you will be comparing two or more datasets, ask how matching will be performed, keeping in mind that each dataset will need to satisfy your minimum standards. You will most likely want to match on age and biological sex, but other demographic criteria may be important to your question.

Your standards may need to be adjusted as you gain a better idea of the data available. Be aware that your ideal data may not exist. However, you will avoid wasted time and effort if you have concrete ideas in mind before beginning your search. If your plan is revised, take this time to pause and re-evaluate your needs before conducting a new search.

4.1.1.1 Analysis Planning: BOLD Activation

Start your planning by considering the contrasts you will need to perform. What conditions will need to be present in an experiment to allow for those contrasts? What would you need to see

in your data to confirm or reject your hypothesized results? Use these requirements to inform your search for data.

A univariate analysis may not need exact information about stimuli presented at each TR; knowledge of the condition, onset, and duration of each trial will likely be sufficient. As such, you will have a wider selection of data available than may exist for other base methods. Decide which template you will use for registration and ask how much you must smooth data in the context of your analysis.

4.1.1.2 Analysis Planning: Functional Connectivity

Connectivity analyses rely on change over the timepoints of an experiment. Noise, especially movement, that is time-locked to a few TRs will be of greater consequence than in an analysis where data are averaged over time. Decide on motion thresholds early in your search, determining the degree and frequency of head movement that will exclude a participant's data from your analysis.

Your planning and search will depend on whether you are analyzing FC in relation to a task or from resting-state (RS) data. If RS-based, specific events are not of concern, but you should still consider how much data will be required to generate reliable results. Resting state scans can range from five minutes to half an hour or more, although test-retest variability has been observed to decrease asymptotically after 9 – 13 minutes (Birn *et al.*, 2013). The amount of data required will depend on your question. Depending on the preferences of the experimenter, participants may spend the scan with their eyes closed, keep their eyes on a fixation cross, or watch a movie in the scanner to stay awake and alert; a naturalistic paradigm like movie-watching may be preferable to a paradigm where participants rested without input (Finn, 2021) and can enhance the predictive value of RS data (Finn & Bandettini, 2021). Sleep during RS is a clear and present concern

(Tagliazucchi & Laufs, 2014), but not all researchers take explicit measures to ensure or measure wakefulness. Ask if the exact paradigm is important to you. Would naturalistic visual and/or auditory input pose a problem? Would you prefer a scan where care was taken to keep participants awake and alert? You may be as flexible or stringent as you choose, but keep in mind that your expectations may change when you become acquainted with the data available to you.

4.1.1.3 Analysis Planning: MVPA

Consider which stimulus or trial categories would be best to test your hypothesis. Do you need access to stimuli belonging to specific semantic categories? Will you need access to a scan in which participants responded to a perception or memory task, or do you need data acquired while participants viewed or listened passively? You will also need sufficient samples to train your classifier. Compose the general framework of the experiment you need to analyze before searching for data. Publications in your area will help you work out which conditions and designs will work best. Realize that the experiment you envision may not exist as open data, so go into your search prepared to re-evaluate along the way.

Go into your processing and analysis aware of the original purpose of the data. You should know the context of the experiment and the analysis mode for which the data were originally collected. You may find yourself analyzing data whose original purpose was distinctly different from its purpose in your project. If the experiment and its original results were based on univariate analyses or FC, realize that you may not find an effect when you apply a very different statistical technique to the same data.

Finally, it is imperative that you examine all timing files and documentation before downloading data and beginning to work. This is important for every method, but MVPA will certainly require more information than is normally captured by default. Do not assume that data

will suit your purposes based solely on the file tree or the structure of the original experiment; as illustrated above in Figures 8-9, even if the data are well-suited to your analysis, timing files may still be incomplete for your purpose.

4.1.1.4 Analysis Planning: RSA

Ask how the items in an experimental task should differ to detect the effect you are looking for. Will variations in basic item-level features – shape, color, motion – make a difference? You may be as general or as specific as you like; your shopping list may read, "visual stimuli with at least for semantic categories" or it may read, "black-and-white static images of at least four types of inanimate object". Keep flexibility in mind and establish bare minimum criteria along with criteria that would exclude a dataset from consideration.

As with MVPA, look out for the original purpose of the data. What was the goal of the original analysis? What motivated the choice of items presented in the experiment? Even if RSA was never used on the data, any former multivariate analysis can indicate whether there was a strong distinction between conditions. Significant results from MVPA and related methods are a good indication that the dataset will be suited to RSA.

If you plan to conduct a searchlight analysis, keep in mind that you will need sufficient coverage and reliable segmentation that allows you to exclude white matter accurately from each searchlight as accurately as possible. Ensure that you are confident in the structural data and its match with each functional scan. Go into your search knowing your seed region(s) and the size of your searchlight – these will factor into your checks.

This is the newest and most infrequently applied of the four base methods discussed here, so you may find yourself encountering more difficulty as you search for data. You may need extra creativity in this case – if you cannot locate a dataset close to what you are looking for, you may

need to devise a different analysis to test your method or construct of interest with what is available.

4.1.2 Data Checks

It may be tempting to assume that shared data will pass the quality checks you would apply when collecting and processing your own original data, especially if they have resulted in a publication. However, there is no guarantee that your adapted pre-processing pipeline will yield the same results that the original authors observed. As such, you should apply the same checks to shared data as you would to your own.

Check motion and noise thresholds, keeping in mind that your standards may result in excluding scans. Confirm that pre-processing steps worked well and look for the sufficient coverage you identified in the planning phase. The checklists provided offer a baseline, but keep in mind any additional standards and quirks specific to your processing stream.

If you have mapped out the preliminary steps of your analysis, you should be able to make a prediction about their results. Overall, any basic analysis should show a consistent pattern within and across scans; for example, canonical resting state networks should be discernible in any functional connectivity analyses. Though present, these results may not rise to significance at the individual level. Check for patterns in your basic results before moving on to the more complex steps of your analysis. If the results differ substantially from your expectations – especially if they appear random, without discernable patterns – take time to consider both the features of the data and the nature of your analysis scripts. You may need to do further statistical checks to identify the issue before moving on.

The same holds true for the results of the final analysis. Whether or not you see significant results, there should be an overall pattern within and across the scans you consider. Even if you do not obtain the hypothesized result, you should see some consistent outcome. If the results appear random, especially at the group level, go back to the data and your scripts. The issue may come down to correcting a bug or altering the steps of processing and/or your base analysis.

As with quality checks, it is tempting to assume that because the data have already been evaluated and shared, results derived from them will be robust. Be aware that you should be applying the same rigorous standards of reliability and replication to shared data that you would to data you collected yourself. Hold to the standards of significance you established in the planning phase. Be wary of doing too many tests or cherry-picking your results. Be open to the fact that, just as in an original study, your analysis may produce null results; even data that resulted in publishable results for the original authors may fail to yield results in new analyses. If you have been diligent up to this point, null results are not a failing on your part or on the part of the sharer.

4.1.2.1 Data Checks: BOLD Activation

Visualize hypothesized activation maps for each experimental condition you will consider. Where can you reasonably expect to see high activation? Which regions should be affected least? For example, both within and across participants, visual stimuli should result in a high degree of signal in V1; faces should evoke activity in the fusiform face area (FFA). Do the same for each of your contrasts, then make a visual inspection of your results at different thresholds. Evaluate your data with and without blur, considering that it is possible to smooth too much. If you do not observe simple confirmatory trends on a gross level, *e.g.,* V1 being highly sensitive to visual trials, the error may lie in your scripts.

4.1.2.2 Data Checks: Functional Connectivity

Keep an eye out for large spikes in movement or noise. Compensating for these time-locked spikes is more important for connectivity than for an analysis that will collapse across time. If you must exclude scans due to excessive movement, ensure that enough data has been retained to detect a reliable effect. If you find yourself excluding many participants from the final analysis, this may not be a good dataset for your needs.

Visualize where significant connectivity should be found - *e.g.*, the default mode network for RS data, or the visual system if visual stimuli are presented. If connectivity patterns appear random across participants, you may be encountering an issue with too much noise. If results appear random or highly unexpected at the group level, you may need to look for outliers and/or revise your thresholds for motion. When correlation is mapped on the brain using a seed-based approach, you should see very high values within your seed region. If a seed ROI does not exhibit near-perfect connectivity with itself, there may be an issue with your processing and/or analysis scripts.

If you are assessing task-based FC, maps should differ visibly in response to the task. Try your basic analysis with and without task condition as a variable. If maps look the same, you may only be detecting intrinsic network activity. Assess whether it is the task or the analysis that is failing to yield results – at an early stage, you can still set your current dataset aside and look for a new task that shows a stronger signal.

4.1.2.3 Data Checks: MVPA

Before beginning a more complex analysis, especially if testing a novel technique based on MVPA, look at your data and think of classification tasks that have obvious results. If there is a region where you know stimuli should be highly discriminable, such as low-level sensory cortex

or a well-established functional region like FFA, train and test your classifier on that region. If you know when rest trials occur, you can do a basic check by trying to classify rest vs. non-rest. Results – both within and across subjects – should match up to your expectations.

If classification is poor on your confirmatory task, the issue may be with the dataset. You may not have enough runs, trials, or categories available to detect an effect. Alternately, if the data were originally acquired for an analysis based on BOLD activation contrasts or FC, your experimental conditions may not be distinct enough to elicit classification. For example, an analysis by Daly *et al.* (2021) used simultaneous EEG-fMRI to identify several sub-cortical regions associated with differing emotional content in music clips. We considered this dataset, but our primary concern was that the four conditions of classical music would be too similar in their auditory features to elicit accurate classification in lower-level sensory regions like primary auditory cortex. You may also wish to revise how you train and test your classifier; if this is not sufficient, you may want to re-evaluate and look for a new dataset.

4.1.2.4 Data Checks: RSA

Take careful note of alignment and segmentation during pre-processing – ensure you are happy with how gray and white matter are being extracted. Take special care to ensure that automated pre-processing is defining all ROIs accurately in both structural and functional data. Accurate results may depend on excluding white matter accurately and ensuring that all the correct voxels, and only the correct voxels, fall within your ROIs.

Consult RSA literature relevant to your question or method and think about regions in which the results should be obvious – this could be lower-level sensory cortex, somatosensory cortex, or ventral temporal cortex, among others – and try your analysis in these regions.

If you are conducting a second-order RSA, consider how the analysis looks within your seed region(s). Minor variations may mean that similarity is not perfect when an ROI is compared with itself, but it should be near-perfect, certainly much higher than any other brain regions. After running a searchlight analysis, check for predictable networks – for example, you may expect to see similarity along the dorsal or ventral stream.

If you see none of the expected patterns in your preliminary, confirmatory results, you may wish to revise the size of your ROIs and your searchlights. Look at the values in your RDMs – if you see very little difference between them, even in ROIs where there should be a clear distinction, try training and testing a basic classifier on the same data. You may find that the data – especially if originally collected for a univariate or connectivity analysis – simply do not offer enough power to distinguish between items on a fine-grained level. In this case, it may be best to revise your plans and search for a new dataset.

4.1.3 Publication and Citation

The dataset you use should be cited in subsequent publications in the manner specified on its OpenNeuro page. If elements of the original study's method are particularly relevant to your analysis, include them in your Methods section. If the original study resulted in a publication, you may wish to draw from the figures provided. Be sure to abide by any journal requirements and seek permission before drawing from previously published figures. If you refer to publications resulting from the data in addition to the dataset itself, be sure to include all relevant citations.

Users and sharers benefit from creating goodwill, connection, and collaboration. If your work is to be published or offered as a pre-print, consider reaching out to the sharing group to share the manuscript and express your appreciation. If you have specific questions, you may

consider reaching out to the sharing group during the earlier stages of your project. If you receive substantial help and feedback from the sharing group, it may warrant joint authorship.

4.2 Guidelines for Sharers

In Appendix A, you will find a checklist for data uploading. It can be used as a general-purpose guide as you begin taking inventory of your study and preparing to organize your data for upload to OpenNeuro.

4.2.1 Documentation

Before composing documentation, look through your data and records from the perspective of a new user. Note any missing or problematic data, persistent artifacts, or issues identified in your quality checks. Even if an issue did not affect your analysis or results, it may be relevant to a user. Many analyses may only find significant results at the group level; when presenting data quality issues, it is particularly important to note artifacts or protocol issues that persist across scans. A problematic or outlying participant can most likely be excluded, but issues that affect group-level analysis will bias results if not accounted for.

Summarize your methods, including the design of the experiment. Offer a basic summary of the behavioral protocol. Specify whether the design was block or event-related. Offer a description of the structure of a run and the order that items were presented – were trials and/or blocks random or pseudo-random? Was there a fixed order, or did it differ across participants?

Note additional data collected during scans, (e.g., simultaneous EEG, physiological data, or behavioral data), even if it is not included with the data as uploaded – it may affect the types of artifacts visible in the raw data. It is important for users to know when and how structural data were collected; the assumption is that a structural scan will be available for each session, so exceptions should be noted. Another common assumption is whole-brain coverage. If your experiment entailed collecting fewer slices or allowing for less coverage, this should be noted as well.

Having this information present in the readme will help users decide quickly whether the design of your experiment meets their needs and whether to invest further time in examining the data. Consider offering a brief description of the participant population, especially if the study dealt with populations other than typical adults.

If your data are derived from an experimental task that presented stimuli at various timepoints, documentation should describe the basic features of the stimuli, as well as the categories from which they were drawn and the experimental conditions into which they fit. You most likely took these categories into account when choosing or creating your stimulus set. For example, if your stimulus set features audio clips, you may have included multiple types of sound for to control for confounding factors; one of the datasets featured in the case study (Pernet *et al*., 2019) balanced animal sounds, environmental sounds, non-speech vocalizations, and speech across trials. Only two experimental conditions were compared in the original analysis – human vocalizations and non-vocal sounds – but secondary analyses could make use of more fine-grained category information. Stimulus features will be among the most helpful pieces of information to offer with task-based data.

If these data have been or will be published on, you may consider repurposing your Methods section for the readme; however, keep in mind that more information will be necessary to allow for secondary analysis than for a Methods section in the context of a larger paper. This is not necessarily a negative; composing the documentation for your dataset may serve as an opportunity to take inventory of the study. Crafting a concise description of the design, motivation, and methods of your experiment can be a valuable exercise, and it will be valuable if you or your colleagues revisit the data in the future.

4.2.1.1 Documentation: BOLD Activation

The information necessary to recreate your results will be near sufficient for a secondary univariate analysis, but you may wish to expand on the methods you would offer in a publication. Be sure to provide a description of the experimental paradigm, including both behavioral and scanning protocols. Knowing task, run, and trial order will aid the construction of a predictive general linear model (GLM).

Users will need to comprehend the experimental conditions of the task, though exact knowledge of stimuli is less important here than in other methods. Provide a description of your stimulus set; if it was drawn from a publicly available source, provide a link if possible.

If your original results incorporate a BOLD activation analysis, consider sharing the contrasts you performed. Similar contrasts may be employed in a novel analysis as a check.

4.2.1.2 Documentation: Functional Connectivity

FC is more likely to be performed on RS data. If an RS scan was included in your protocol, pay special attention to the checklists for FC. If you are documenting the protocol for an RS scan, note the conditions of data acquisition. What were participants doing during the scan? Did it entail

auditory or visual input? Were they offered something to place their focus on, and were they given instructions about focus or wakefulness? For task-based data, summarizing protocol can cue users into any issues that might result in excessive movement or noise; for example, participants' use of response devices can result in time-locked head motion.

If your study dealt with a clinical population, note their diagnosis along with any other relevant factors. Secondary factors used in recruitment may seem trivial, but they could be crucial for a novel question. For example, consider two datasets of a similar size drawn from clinical studies. One study recruited adults currently undergoing treatment for paranoid schizophrenia, while the other recruited adults who self-reported a diagnosis of schizophrenia or schizoaffective disorder. These samples may be similar, but a user searching for data from schizophrenic patients for a secondary FC analysis will benefit from the ability to make a distinction between them.

4.2.1.3 Documentation: MVPA

Be sure to document the original parameters of your analysis. This is most important for those who collected functional data with a univariate or connectivity analysis in mind. Identify not only your experimental conditions, but the types of trials that fell into those conditions. Be specific about the stimuli and categories. You most likely took these categories into account when choosing or creating a stimulus set or assembling prompts for your behavioral protocol.

Keep in mind that MVPA is seldom performed on smoothed data, although studies have found that a small degree of smoothing can aid decoding (Hendriks *et al.*, 2017) and some labs do choose to blur data for MVPA. In general, note factors that may have resulted in excessive motion along with consistent sources of noise that you accounted for during your own quality checks.

4.2.1.4 Documentation: RSA

You will not be able to predict which semantic or structural features of your stimuli will be relevant to a secondary analysis, so try to be as detailed as possible. If your goal is to enable secondary analyses based on RSA, keep in mind that most will require knowledge of the semantic and sensory features of the items presented during a scan. Consider your motivation for the protocol, how you selected your stimuli, what features the items share, and what you controlled for – all these factors need not be included in documentation, but it may help to give thought to them when summarizing the protocol and offering a brief description of the items presented to participants in the scanner.

If possible, offer some information on your initial question and analysis. If your original aims were far removed from knowledge representation and/or your original analysis was not concerned with information at the multivoxel pattern level, your data may not be a good fit for users hoping to conduct RSA-based analyses. You will save them time and frustration if they are able to make this determination immediately.

The relevance of a given brain area will depend on the user's analysis, but a searchlight analysis will take the whole brain into account. Be sure to note anything that you found to cause blur or distortion in your data, or anything that you found that affected segmentation. If you notice susceptibility artifacts – for example, minimal distortion from simultaneous EEG – make sure to note it as a possible issue even if it did not affect your results.

4.2.2 Supplemental Files

Before uploading a dataset, take time to catalogue the information captured during the associated study. You will likely find many records that did not, ultimately, inform your analysis.

However, many of these will be helpful for secondary analysis. Keep in mind the most important files discussed in the previous section: event timing, participant demographics, and scan parameters. These will be, in some form or other, a part of the base of information you have accumulated over the course of a study. If your goal is to enable secondary analysis, now is the time to format them for public consumption.

Ensure that event timing files are formatted such that a new user without your data or publications to hand can discern what happened at each timepoint; if events like stimulus presentation, fixation, response, and rest are represented by numbers or codes specific to your scripts, consider reformatting files to make them unambiguous. Include the scan parameters necessary to develop a new pipeline suited to your raw data – you may choose to put this information in the .json file(s) that accompany your data, or you may include text files for each session and/or for the dataset as a whole. Participant demographics should include, at minimum, age and biological sex for matching; if other information was captured during recruitment, take inventory of it and include elements that might inform a secondary analysis. Think about possible confounding variables captured in participant information – whether or not they affected your analysis – and what new users might need to know about them.

You may have captured additional results outside the scanner. For example, behavioral performance, diagnostic measures, or developmental testing may form part of the dataset available to you. Including additional results may not be necessary for enabling secondary analysis, but it will expand the number and types of secondary analyses that can be conducted. If your goal is to invest in open data, supplying optional supplementary files will maximize secondary use and benefit the community.

4.2.2.1 Supplemental Files: BOLD Activation

When formatting timing files, it is important for users to know the experimental condition of every trial. Exact information on stimuli is less important, as data will be averaged across trials. Onsets and durations are necessary to calculate beta weights for trials of interest. If timing is essentially the same across scans, it may be easier to include a single timing file with the dataset. However, more exact timing will aid the construction of a GLM. If exact trial timing was recorded during a scan, it should be included.

While exact knowledge of stimuli presented on each trial is not crucial, exposure to the stimulus set will aid the user. If you created custom stimuli, consider providing them along with the dataset.

A univariate analysis is a common way to re-evaluate or recontextualize older clinical or developmental results. If you have captured participant information for your study beyond basic demographics – e.g., scores on diagnostic or developmental measures – consider providing it to maximize future use.

4.2.2.2 Supplemental Files: Functional Connectivity

For task-based data, users will need to know the state of the experiment at each TR. For RS data, it is common to remove the first few TRs of the scan, so users should have clear onset and end times for each scan.

If you aim to empower novel analysis as much as possible, consider additional data collected at the scanner. If you recorded eye-tracking data, even if these were not included in your analyses, they can be used to assess wakefulness during an RS scan. Physiological data can also be used to assess wakefulness or to do fine-grained motion correction based on respiration.

If your study dealt with atypical populations, aim to include as much detail as possible in demographic files. If you captured a diagnosis or assessment score for each participant, consider adding a table to the dataset as uploaded. While some studies on OpenNeuro offer only age ranges (e.g., "25-30") with participant demographics, if your study concerned developing or geriatric participants, exact age may be a concern. Include this information if it was captured and note in documentation if it was not.

4.2.2.3 Supplemental Files: MVPA

Timing files are the most important consideration for secondary MVPA. Users commencing an MVPA-based analysis will need an accessible reference that specifies the category and/or identity of the item presented on each trial. If your stimuli for each experimental condition were drawn from a balanced set of semantic categories, the category of each stimulus should be noted in your timing files. If you used a single category of stimuli but balanced features of them – for example, presenting faces either head-on or in profile – these features should be included, as they could form factors in a multivariate analysis.

4.2.2.4 Supplemental Files: RSA

As with MVPA, users will need to know not only the experimental condition of each trial, but the more specific category of the item presented. In this case, even knowing the identity of the stimulus or item presented on each trial may be relevant. If your timing files identify the individual item presented at a given time, consider making the identifiers accessible to an outside user – rather than a number or cryptic filename, try to use readable text. For example, a study presenting multiple animal stimuli might offer identifiers like "fish1" and "fish2" for two separate images within a category.

The specific features of your stimuli – whether physical or semantic – may be relevant to RSA-based analysis. In addition to offering a brief description of your stimuli in the readme, consider providing the stimulus set itself. Having the option of examining the stimuli will be a great help to users deciding whether your data will be a good fit.

4.2.3 Future Planning

Formatting and uploading your data may seem like the final step to sharing, but the process continues, especially if you wish to enable secondary analysis. In the future, you may receive comments on the dataset's OpenNeuro page along with direct email communications about the data. Potential users will ask questions, point out bugs, and request additional supplemental materials. As the platform grows and changes – especially in the wake of the recent redesign of the website UI – you may also need to make updates to keep in line with the standards of OpenNeuro and BIDS format. If your goal is to help new users apply your data, plan ahead by assigning a member of your group to check in on the data regularly. Look for formatting warnings from OpenNeuro, new comments, and use statistics. You may also choose to delegate the task of responding to questions about the dataset – designate a "corresponding author" for the data who will communicate with researchers seeking more information or offering potential opportunities for collaboration. Supporting and updating your dataset will increase its use and build goodwill on the platform.

4.3 In Conclusion

Once you have become familiar with the concrete guidelines outlined above, sit down with your own data or analysis project and begin to map out the process of applying them. In Appendix 2, you will find checklists to guide you through this process. The material is intended to help you hit the ground running. However, remember that you know your needs best – do not be afraid to make modifications or go beyond the guidelines offered in this article.

The materials offered here also assume that users are familiar with the basics of pre-processing and analysis in their chosen mode. You should know how to construct a pre-processing pipeline or have access to one that has already been constructed. You should also have a solid grasp on how to set up your base analysis, e.g., how to set up a GLM for BOLD analysis or train a classifier for MVPA.

Ensure that you understand the core steps of your pre-processing and analysis before you begin. Do not be afraid to let your technical shortcomings influence your search for data – it is far better to acknowledge your strengths and weaknesses ahead of time than to set your project back by struggling to apply your existing skills to a difficult dataset.

Finally, remember that enabling and engaging in secondary analysis are skills that can be learned. Going through this process once or twice will strengthen your existing skill set, familiarize you with the demands of the platform, and make future studies more streamlined. Remember that you are making an investment in your own research and in the field as a whole. As you practice and progress, that investment will only become more meaningful.

5.0 Conclusion

5.1 Platform Recommendations

A discussion of secondary analysis on OpenNeuro cannot conclude without addressing what the platform can do to enable secondary analysis, taking some of the burden off sharers and diminishing the effort barrier that keeps some from sharing smaller-scale datasets. The OpenNeuro UI recently underwent a redesign (Markiewicz *et al*., 2021) that made some helpful changes to the platform. As a database committed to established standards of data-sharing, OpenNeuro seeks to abide by the FAIR principles, endeavoring to ensure that data are findable, accessible, interoperable, and reusable (Wilkinson *et al*., 2016). The BIDS validator identifies and extracts metadata, and the redesign streamlines searching by features like date, author, modality, and participant population. Sharers may choose to add additional metadata when uploading, including associated publications. However, only a limited subset of such metadata is automatically indexed by the BIDS validator. This diminishes the effectiveness of the new search tools; for example, as of November 2021, a general keyword search for "oddball" returns 10 datasets, including the visual/auditory oddball task discussed in Box X, while a search for "oddball" within the "Task" sub-field returns only one dataset. The "Diagnosis" sub-field contains only three specific diagnoses apart from Healthy Control, despite the broad array of clinical populations available.

Much can be automated to provide a more streamlined experience. As it stands, providing well-composed documentation and supplementary files is the best way to ensure that a dataset is useful for secondary analysis. However, providing and automatically indexing more metadata

fields for sharers would diminish the time and effort users spend searching while minimizing post-study labor for sharers.

BIDS format also provides guidelines for event timing files (Gorgolewski *et al.*, 2016), but only onsets and durations must be present for the timing file to be validated. Other fields, including trial type, are optional but encouraged. Further tags can be added using Hierarchical Event Descriptors (Robbins *et al.*, 2020), an annotation framework, but including such elements as trial and stimulus identifiers remains optional. Participant demographic files remain entirely optional, with some demographic metadata indexed but no set format for the file itself (Gorgolewski *et al.*, 2016). While keeping standards flexible is crucial to the OpenNeuro model, further requirements and more automation for event and demographic files would make a big difference to efficacy while adding comparatively little burden. As it stands, the image files themselves are far more comprehensively regulated, validated, and indexed than the accompanying behavioral data or any supplemental files. Enabling secondary analysis could entail taking a closer look at how supplemental files can be more FAIR.

5.2 Recommendations for the Field

In keeping with the growing commitment to open science within academic publishing and funding, one may also ask where the field as a whole can go in the next years and decades to increase collaboration through data reuse. Prioritizing secondary analysis should form part of this effort. Incentivizing reuse with as much zeal as replication will be an important step towards the goals of equity, innovation, and collaboration discussed in this document. A further step could be working to publicize the reuse side of open data as much as the replicability side. The more

common data-sharing in this mode becomes, the more the ecosystem supporting it will grow – ultimately diminishing the barriers of effort and frustration that currently discourage some groups from data-sharing or secondary analysis on a smaller scale.

5.3 Future Directions

As this document stands, it is our hope that it serves as both encouragement and call to action. In the future, these standards for uploading and reuse can be modified for other neuroimaging modalities, other platforms, and even open data in other fields. The psychology open data landscape is vast and includes large and small repositories for behavioral, psychophysical, clinical, and demographic data. No matter the mode, the standards of clear documentation and supplemental information will hold when it comes to empowering secondary analysis.

5.4 A Final Message

Whether you are using or sharing on OpenNeuro, it is our hope that this document and the accompanying materials will help you make a strong start. We hope that we can help you gain skills that carry forward into future studies, making your lab more efficient and more open. Feel free to adapt the checklists and recommendations to better suit your own work. The more you get to know your own needs and strengths, the better your results will become. Sharing your data with the world and engaging with research through secondary analysis are both meaningful tasks, especially now. The effort you put in now will pay off. Navigating the open neuroimaging

landscape can be exceptionally difficult, but the connections it creates can also be exceptionally rewarding. Good luck, and don’t panic.

Appendix A OpenNeuro Upload Checklist

If relevant data were not collected during your study, mark N/A.

DATA TO PROVIDE

Must-Have

For each session:

Present	N/A	
•	•	Functional data from all tasks, by run
•	•	Header files for every scan, by run
•	•	T1w MPRAGE
•	•	Resting State data

Should-Have

For each session:

Present	N/A	
•	•	Field maps
•	•	T2w structural images

Optional

For each session:

Present	N/A	
•	•	Simultaneous EEG
•	•	Physiological data (Heart rate, respiration, skin conductance)
•	•	Eye-tracking data

Derivatives:

Present	N/A	
•	•	Motion-corrected functional images
•	•	Smoothed functional images
•	•	Warped functional images
•	•	Affine matrix used for warping

For each participant:

Present	N/A	
•	•	Gray matter mask
•	•	White matter mask

•	•	CSF mask
•	•	Whole-brain mask
•	•	ROI masks from original analysis

SUPPLEMENTAL FILES

Timing Files

Must-Have		
Present	N/A	
•	•	Onsets and durations
•	•	Experimental condition of each stimulus presented
•	•	Rest trials
•	•	Onset and duration for resting state scan
•	•	Fixation or stimulus onset for resting state scan

Should-Have		
Present	N/A	
•	•	Category of each stimulus presented
•	•	Stimulus presented
•	•	Identify stimuli with clear names
•	•	Response required?

Participant Demographics

Must-Have		
Present	N/A	
•	•	Age or age range for each participant
•	•	Biological sex of each participant

Should-Have		
Present	N/A	
•	•	Exact age for each participant
•	•	Race
•	•	Nationality
•	•	Diagnostic status
•	•	Experimental group

•	•	Gender identity

Must-Have		
Present	N/A	
•	•	Link to associated publications
•	•	.json file with basic study information and citation

Should-have		
Present	N/A	
•	•	Stimulus set or link to publicly available stimuli
•	•	Slice order

Optional		
Present	N/A	
•	•	Behavioral data
•	•	Scores on diagnostic measures

DOCUMENTATION

Must-Have		
Summarize Methods:		
Present	N/A	
•	•	Describe each task
•	•	Identify files provided for each session
•	•	Identify experimental conditions
Data Quality:		
Present	N/A	
•	•	Missing data
•	•	Persistent artifacts
•	•	Irregular scans
•	•	Data quality issues that affected pre-processing
Study Information:		
Present	N/A	
•	•	Authors
•	•	Associated publications
•	•	Corresponding author information
•	•	Citation information

Should-have		
Task design:		
Present	N/A	
•	•	Experimental conditions
•	•	Describe trials
•	•	Block or event-related
•	•	Trial order random, pseudo-random, or the same across participants
•	•	Sources of extra movement or noise
•	•	Describe stimuli presented
•	•	Describe categories and features of stimuli
•	•	When/if participants responded
Resting State:		
Present	N/A	
•	•	Length of scan
•	•	Participant activities during scan
Functional:		
Present	N/A	
•	•	Number of slices
•	•	Coverage
Structural:		
Present	N/A	
•	•	Number of structural scans per participant
•	•	Length of scan
Study design:		
Present	N/A	
•	•	Participant population
•	•	Task order
•	•	Sessions per participant
•	•	Other data collected – simultaneous EEG, physio, eye-tracking, etc.
•	•	Describe field-mapping

Optional		
Present	N/A	
•	•	Motivation for original study
•	•	Where data were collected
•	•	Analyses originally performed

- Derivatives available
- How responses were recorded

FUTURE PLANNING

Tasks to delegate:

- Monitor BIDS compliance warnings
- Update dataset and associated publications as necessary
- Respond to OpenNeuro comments
- Respond to communications about dataset

Appendix B OpenNeuro Download Checklist: BOLD Activation-Based Analysis

PLANNING YOUR ANALYSIS

Identify:
• Your base analysis • Your pre-processing steps • Core steps of your planned analysis, from beginning to end • Regions of interest that will be considered • Your relevant technical skills • Your relevant technical weaknesses • Whether matching must be performed, and how

Construct Your Shopping List
Inclusion/exclusion criteria:
• Data quality • Coverage • Number of participants • Participant population *Examples: Diagnostic status, age, developmental stage, typical adult*
Amount of data:
• Number of participants/scans • Number of trials • Number of runs • Length of scans • Structural scans • Number of conditions
Stimuli:
• Experimental conditions *Examples: Reward, loss; visual, audiovisual; words, numbers, faces* • Stimulus modality *Examples: Visual input; auditory input; spoken words; motion; text*

Plan your base analysis and checks
• Identify your contrasts and hypothesized results. Visualize hypothesized results on the brain and in your ROIs. • For checks, identify regions and contrasts that should have obvious results. *Examples: Visual stimuli should activate V1, auditory stimuli should activate primary auditory cortex; comparing two experimental conditions should yield different results than either condition alone; comparing reward to loss should affect striatum.* • Consider stimulus modalities. Consult relevant publications too see what basic contrasts can be performed.

For each dataset:
• Read entire readme • Examine file tree and compare to shopping list • Check header files for scan parameters • Examine timing files and participant demographics • Visually examine the structural and functional data • Skim related publications and external documentation
Ask:
• Has this dataset been updated as necessary? • Do authors respond to comments on this dataset? • Will the labor of adapting and troubleshooting your pre-processing and analysis for this dataset outweigh its utility?

DATA CHECKS

Pre-Processing
• Download and process several participants' data with your pipeline
Do your customary quality checks:
• Motion correction/de-spike • Skull-stripping • Adequate coverage of your ROIs • Accurate warping • Accurate structural/functional alignment

- Accurate segmentation

Masks:

- Gray matter
- White matter
- Whole-brain
- Your ROIs

Base Analysis

- Construct your GLM and run the contrasts you identified on your pre-processed participants.
- Look for activation in the confirmation regions you identified
- Try the confirmation contrasts you identified and look for expected results
- Try your contrasts of interest and observe results at different thresholds
- Check for patterns within participants, even if not significant at the individual level. Patterns of activity should be relatively similar across scans.
- Look for patterns across your test scans. There should be recognizable patterns, even if slightly inconsistent and not significant at individual level.
- If participant-level checks give expected results, download and pre-process remaining data
- Run checks at group level
- Results should be consistent with participant-level patterns. Inconsistencies should even out at group level.

If your results look unexpected at any stage:

N/A

- Revisit your scripts. Go through step-by-step and check for bugs, then correct and try checks again.
- Look for outliers or highly irregular participants, then try checks again with outliers excluded.
- Revisit your smoothing kernel – try smoothing data more or less

After troubleshooting, look for a new dataset:

True False N/A

- If your results appear random, with no patterns within or across participants
- If your new contrasts show no difference between conditions
- If it becomes too time-consuming or difficult to automate pre-processing
- If too many participants, timepoints, or scans must be excluded to meet your desired effect size
- If the labor or time you will put into adapting and troubleshooting your analysis for this dataset will set back your project significantly

PUBLICATION AND CITATION

If you plan to publish your results

- Cite dataset using format specified on OpenNeuro
- Cite all related publications referenced for your results
- If drawing from published figures, contact original journal and authors for permission
- When publishing or uploading as a pre-print, contact original authors

Appendix C OpenNeuro Upload Checklist: Functional Connectivity-Based Analysis

PLANNING YOUR ANALYSIS

Identify:
• Your base analysis • Your pre-processing steps • Core steps of your planned analysis, from beginning to end • Regions of interest that will be considered • Your relevant technical skills • Your relevant technical weaknesses • Whether matching must be performed, and how • Thresholds for motion

Construct Your Shopping List
Inclusion/exclusion criteria:
• Data quality • Coverage • Number of participants • Participant population *Examples: Diagnostic status, age, developmental stage, typical adult*
Amount of data:
• Number of participants/scans • Structural scans
Format of RS scan:
• Length N/A • • Participant activity *Does it matter what the participant saw or heard during the scan? Do you need to know they were fixating and/or awake during the scan?*

Format of task-based scans:
N/A • Experimental conditions • Stimulus modality *Visual, auditory, audiovisual, etc.*

Plan your base analysis and checks
• Identify a seed region that should be part of a canonical network *Examples: Default mode network should be active during RS scans; V1 should be strongly associated with the visual system when visual stimuli are presented* • Visualize hypothesized results for your ROIs

For each dataset:
• Read entire readme • Examine file tree and compare to shopping list • Check header files for scan parameters • Examine timing files and participant demographics • Visually examine the structural and functional data • Skim related publications and external documentation • Look for results from any FC that was applied to the data for the original project
Ask:
• Has this dataset been updated as necessary? • Do authors respond to comments on this dataset? • Will the labor of adapting and troubleshooting your pre-processing and analysis for this dataset outweigh its utility?

DATA CHECKS

Pre-Processing
• Download and process several participants' data with your pipeline
Do your customary quality checks:
• Motion correction/de-spike • Skull-stripping

- Adequate coverage of your ROIs
- Accurate registration
- Accurate structural/functional alignment
- Accurate segmentation

Masks:

- Gray matter
- White matter
- Whole-brain
- Your ROIs

Base Analysis

- Run your FC analysis in your pre-processed scans using the confirmation seed region(s) you identified. Inspect the results at different thresholds and look for the canonical networks you expect to see.
- Run your FC analysis between your ROIs and themselves. Correlation between a seed region and itself should be near-perfect.

N/A

- For task-based FC, run your analysis with your chosen seed region(s), with and without task as a variable. Inspect the results at different thresholds and look for differences.

- Check for patterns within participants, even if not significant at the individual level. Patterns of activity should be relatively similar across scans.
- Look for patterns across your test scans. There should be recognizable patterns, even if slightly inconsistent and not significant at individual level.
- If participant-level checks give the expected results, download and pre-process remaining data.
- Run checks again at the group level.
- Results should be consistent with individual-level patterns. Inconsistencies should even out at group level.

If your results look unexpected at any stage:

N/A

- Revisit your scripts. Go through step-by-step and check for bugs, then correct and try checks again.

- Look for outliers or highly irregular participants, then try checks again with outliers excluded.
- Check de-spike results. Make a visual check of data and look for motion spikes that might have an outsized influence on results. Try checks again with high-motion TRs excluded.
- Explore methods of compensating for intrinsic connectivity.
- For resting state, ask if scan is sufficient, too short, or too long.

After troubleshooting, look for a new dataset:			
True	False	N/A	
•	•	•	If your results appear random, with no patterns within or across participants
•	•	•	If you do not see any sign of canonical networks, especially at the group level
•	•	•	If it becomes too time-consuming or difficult to automate pre-processing
•	•	•	If too many participants, trials, or TRs must be excluded to meet the minimum number you established when planning
•	•	•	If you see intrinsic network activity, but no effect of manipulations
•	•	•	If the labor or time you will put into adapting and troubleshooting your analysis for this dataset will set back your project significantly

PUBLICATION AND CITATION

If you plan to publish your results:
• Cite dataset using format specified on OpenNeuro
• Cite all related publications consulted for the body of your paper
• If drawing from published figures, contact original journal and authors for permission
• When publishing or uploading as a pre-print, contact original authors

Appendix D OpenNeuro Download Checklist: MVPA-Based Analysis

PLANNING YOUR ANALYSIS

Identify:
• Your base analysis • Your pre-processing steps • Core steps of your planned analysis, from beginning to end • Regions of interest that will be considered • Your relevant technical skills • Your relevant technical weaknesses • Whether matching must be performed, and how • Your approach to classification

Construct Your Shopping List
Inclusion/exclusion criteria:
• Data quality • Coverage • Number of participants • Participant population *Examples: Diagnostic status, age, developmental stage, typical adult*
Amount of data:
• Number of participants/scans • Length of scans • Structural scans • Number of trials per scan • Number of runs per scan
Stimuli:
• Experimental conditions • Stimulus modality *Visual, auditory, audiovisual, etc.*

- Desired stimulus features
  Examples: Black-and-white images, naturalistic audio clips, video clips with sound, black text on white background
- Desired stimulus categories
  Examples: Human faces, manipulable objects, nouns, scrambled images
- Minimum number of categories
- Desired task
  Examples: Remembering stimuli, identifying stimuli, answering questions about stimuli

Plan your base analysis and checks
• Identify ROIs where classification should be highly accurate *Examples: Faces will be distinct in FFA; ventral temporal cortex represents category knowledge; V1 will be sensitive to low-level visual features* • Hypothesize the results of classification in your target ROIs

For each dataset:
• Read entire readme • Examine file tree and compare to shopping list • Check header files for scan parameters • Skim related publications and external documentation • Identify original analyses applied to the data, and their results • Identify stimulus categories within each condition • Examine participant demographics
Examine timing files. Look for:
• Trial condition • Stimulus category for each trial • Stimulus identified for each trial • Timing files clear and decipherable • Visually examine the structural and functional data
Ask:
• Has this dataset been updated as necessary? • Do authors respond to comments on this dataset? • Will the labor of adapting and troubleshooting your pre-processing and analysis for this dataset outweigh its utility?

DATA CHECKS

Pre-Processing
• Download and pre-process several participants' data with your pipeline
Do your customary quality checks:
• Motion correction/de-spike • Skull-stripping • Adequate coverage of your ROIs • Accurate warping, if data will be warped • Accurate structural/functional alignment • Accurate segmentation
Masks:
• Gray matter • White matter • Whole-brain • Your ROIs

Base Analysis
• Run MVPA in your pre-processed scans using the confirmation ROI(s) you identified. Look for high classifier performance where it should be expected. • Try classifying rest vs. task trials • Classifier performance should be similar across participants, though exact values will be inconsistent. If one participant stands out, consider whether to exclude or try to adapt for the data. • Run MVPA in your ROIs and check for similarity to your hypothesized results. • If participant-level checks give the expected results, download and pre-process remaining data. • Run checks again at the group level. • Results should be consistent with individual-level patterns. Inconsistencies should even out at group level.

If your results look unexpected at any stage:
N/A • • Revisit your scripts. Go through step-by-step and check for bugs, then correct and try checks again.

•	•	Confirm that all TRs are labeled correctly and shifted to compensate for HRF
•	•	Look for outliers or highly irregular participants, then try checks again with outliers excluded.
•	•	Ask if your classifier algorithm is right for the task
•	•	Ask if you have enough runs and enough TRs to train your classifier

After troubleshooting, look for a new dataset:			
True	False	N/A	
•	•	•	If your classifier performance is consistently at or below chance in confirmation analyses
•	•	•	If you must exclude too many participants to achieve your planned effect size
•	•	•	If it becomes too time-consuming or difficult to automate pre-processing

PUBLICATION AND CITATION

If you plan to publish your results
• Cite dataset using format specified on OpenNeuro
• Cite all related publications referenced for your results
• If drawing from published figures, contact original journal and authors for permission
• When publishing or uploading as a pre-print, contact original authors

Appendix E OpenNeuro Download Checklist: RSA-Based Analysis

PLANNING YOUR ANALYSIS

Identify:
• Your base analysis • Your pre-processing steps • Core steps of your planned analysis, from beginning to end • Regions of interest that will be considered • Your relevant technical skills • Your relevant technical weaknesses • Whether matching must be performed, and how • Seed regions for searchlight analysis

Construct Your Shopping List
Inclusion/exclusion criteria:
• Data quality • Coverage • Number of participants • Participant population
Amount of data:
• Number of participants/scans • Length of scans • Structural scans • Number of trials per scan • Number of runs per scan
Stimuli:
• Experimental conditions • Minimum number of unique stimuli • Minimum number of discrete stimulus categories • Stimulus modality *Visual, auditory, audiovisual, etc.*

- Desired stimulus features
  Examples: Black-and-white images, naturalistic audio clips, video clips with sound, black text on white background
- Desired stimulus categories
  Examples: Human faces, manipulable objects, animals, nouns, scrambled images
- Desired task
  Examples: Remembering stimuli, identifying stimuli, answering questions about stimuli

Plan your base analysis and checks

- Identify ROIs where stimulus categories should be highly dissimilar
  Examples: Sounds should be highly distinct in primary auditory cortex; ventral temporal cortex should represent object categories; words should be highly distinct in VWFA
- Identify ROIs where stimulus categories should not be highly dissimilar
  Examples: auditory stimuli should not be well-distinguished in V1; visual object categories should not be well-represented in primary auditory cortex
- If conducting a searchlight analysis, identify predictable patterns of high correlation with your seed regions
  Examples: *ventral stream, dorsal stream, association cortex*

For each dataset:

- Read entire readme
- Examine file tree and compare to shopping list
- Check header files for scan parameters
- Skim related publications and external documentation
- Identify original analyses applied to the data, and their results
- Identify stimulus categories within each condition
- If provided, examine stimulus set
- If no stimuli provided, look for examples and information in readme or associated publications
- Examine participant demographics

Examine timing files. Look for:

- Trial condition
- Stimulus category for each trial

- Stimulus identified for each trial
- Timing files clear and decipherable
- Visually examine the structural and functional data

Ask:

- Will you be able to identify the stimulus category or event presented at each TR?
- Is there reason to think that stimulus categories will be distinct enough for RSA?
- Has this dataset been updated as necessary?
- Do authors respond to comments on this dataset?
- Will the labor of adapting and troubleshooting your pre-processing and analysis for this dataset outweigh its utility?

DATA CHECKS

Pre-Processing

- Download and process several participants' data with your pipeline

Do your customary quality checks:

- Motion correction/de-spike
- Skull-stripping
- Adequate coverage of your ROIs
- Accurate warping
- Accurate structural/functional alignment
- Accurate segmentation

Masks:

- Gray matter
- White matter
- Whole-brain
- Your ROIs

Base Analysis

- Run RSA in your pre-processed scans using the confirmation ROI(s) you identified. Look for high or low dissimilarity where they should be expected.

- Run second-order RSA within seed regions. Dissimilarity should be near zero when a region is compared with itself.
- Run searchlight analysis in your pre-processed scans and look for the expected, canonical patterns of similarity for seed region(s).
- Results should show a consistent pattern across participants, even if values are different and/or not significant at the individual level.
- If participant-level checks give the expected results, download and pre-process remaining data.
- Run checks again at the group level.
- Results should be consistent with individual-level patterns. Inconsistencies should even out at group level.

If your results look unexpected at any stage:

N/A

- Revisit your scripts. Go through step-by-step and check for bugs, then correct and try checks again.
- Confirm that all TRs are labeled correctly and shifted to compensate for HRF.
- Look for outliers or highly irregular participants, then try checks again with outliers excluded.
- Try running an MVPA classifier within your seed regions to see whether stimulus categories are well-represented.
- Ask if you have selected the right type of stimuli and categories for your analysis.
- Ask if you have enough stimuli, categories, and trials to detect an effect in your analysis.
- Confirm that WM has been excluded from searchlights.
- Consider the size of your searchlights and ROIs.

After troubleshooting, look for a new dataset:

- If you do not see expected patterns of dissimilarity in your checks, across participants and at the group level
- If your data appears random, with no patterns within or across participants
- If it is clear that stimuli are not distinct enough to detect an effect in this dataset
- If you must exclude too many outlying participants to achieve your planned effect size

- If it becomes too time-consuming or difficult to automate pre-processing
- If the labor or time you will put into adapting and troubleshooting your analysis for this dataset will significantly set back your project

PUBLICATION AND CITATION

If you plan to publish your results

- Cite dataset using format specified on OpenNeuro
- Cite all related publications consulted for the body of your paper
- If drawing from published figures, contact original journal and authors for permission

Bibliography

Birn, R. M., Molloy, E. K., Patriat, R., Parker, T., Meier, T. B., Kirk, G. R., Nair, V. A., Meyerand, M. E., & Prabhakaran, V. (2013). The effect of scan length on the reliability of resting-state fMRI connectivity estimates. *NeuroImage*, 83, 550–558. https://doi.org/10.1016/j.neuroimage.2013.05.099

Bjork, J. M., Straub, L. K., Provost, R. G., & Neale, M. C. (2017). The ABCD Study of Neurodevelopment: Identifying Neurocircuit Targets for Prevention and Treatment of Adolescent Substance Abuse. *Current Treatment Options in Psychiatry*, 4(2), 196–209. https://doi.org/10.1007/s40501-017-0108-y

Center for Open Science. (2021). *Open Science Badges*. Center for Open Science. Retrieved November 20, 2021, from https://www.cos.io/initiatives/badges

Coutanche, M. N., & Thompson-Schill, S. L. (2013). Informational connectivity: Identifying synchronized discriminability of multi-voxel patterns across the brain. *Frontiers in Human Neuroscience*, 7. https://doi.org/10.3389/fnhum.2013.00015

Daly, I., Nicolaou, N., Williams, D., Hwang, F., Kirke, A., Miranda, E., & Slawomir J. Nasuto. (2021). *An EEG dataset recorded during affective music listening* [Dataset]. Openneuro. https://doi.org/10.18112/OPENNEURO.DS002721.V1.0.2

Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., Balsters, J. H., Baxter, L., Beggiato, A., Bernaerts, S., Blanken, L. M. E., Bookheimer, S. Y., Braden, B. B., Byrge, L., Castellanos, F. X., Dapretto, M., Delorme, R., Fair, D. A., Fishman, I., … Milham, M. P. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific Data*, 4(1), 170010. https://doi.org/10.1038/sdata.2017.10

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., … Milham, M. P. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6), 659–667. https://doi.org/10.1038/mp.2013.78

DirectMed Parts & Service, LLC. (2020, October 26). *Siemens MRI Machine Cost & Models*. DirectMed Parts & Service.

Dryad. (2021). *Dryad*. Data Dryad. https://datadryad.org/stash/

Elam, J. S., Glasser, M. F., Harms, M. P., Sotiropoulos, S. N., Andersson, J. L. R., Burgess, G. C., Curtiss, S. W., Oostenveld, R., Larson-Prior, L. J., Schoffelen, J.-M., Hodge, M. R., Cler,

E. A., Marcus, D. M., Barch, D. M., Yacoub, E., Smith, S. M., Ugurbil, K., & Van Essen, D. C. (2021). The Human Connectome Project: A retrospective. *NeuroImage*, 244, 118543. https://doi.org/10.1016/j.neuroimage.2021.118543

Finn, E. S. (2021). Is it time to put rest to rest? *Trends in Cognitive Sciences*, 25(12), 1021–1032. https://doi.org/10.1016/j.tics.2021.09.005

Finn, E. S., & Bandettini, P. A. (2021). Movie-watching outperforms rest for functional connectivity-based prediction of behavior. *NeuroImage*, 235, 117963. https://doi.org/10.1016/j.neuroimage.2021.117963

Gorgolewski, K. J., Alfaro-Almagro, F., Auer, T., Bellec, P., Capotă, M., Chakravarty, M. M., Churchill, N. W., Cohen, A. L., Craddock, R. C., Devenyi, G. A., Eklund, A., Esteban, O., Flandin, G., Ghosh, S. S., Guntupalli, J. S., Jenkinson, M., Keshavan, A., Kiar, G., Liem, F., … Poldrack, R. A. (2017). BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLOS Computational Biology*, 13(3), e1005209. https://doi.org/10.1371/journal.pcbi.1005209

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3(1), 160044. https://doi.org/10.1038/sdata.2016.44

Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S. S., Maumet, C., Sochat, V. V., Nichols, T. E., Poldrack, R. A., Poline, J.-B., Yarkoni, T., & Margulies, D. S. (2015). NeuroVault.org: A web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics*, 9. https://doi.org/10.3389/fninf.2015.00008

Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W., & Stadler, J. (2018). *Forrest Gump* [Dataset]. Openneuro. https://doi.org/10.18112/OPENNEURO.DS000113.V1.3.0

Hanke, M., Dinga, R., Häusler, C., Guntupalli, J. S., Casey, M., Kaule, F. R., & Stadler, J. (2015). High-resolution 7-Tesla fMRI data on the perception of musical genres – an extension to the studyforrest dataset. *F1000Research*, 4, 174. https://doi.org/10.12688/f1000research.6679.1

Hendriks, M. H. A., Daniels, N., Pegado, F., & Op de Beeck, H. P. (2017). The Effect of Spatial Smoothing on Representational Similarity in a Simple Motor Paradigm. *Frontiers in Neurology*, 8, 222. https://doi.org/10.3389/fneur.2017.00222

Holdgraf, C., Appelhoff, S., Bickel, S., Bouchard, K., D'Ambrosio, S., David, O., Devinsky, O., Dichter, B., flinker, adeen, Foster, B., Gorgolewski, K. J., Groen, I. I. A., Groppe, D., Gunduz, A., Hamilton, L. S., Honey, C. J., Jas, M., Knight, R., Lachaux, J.-P., … Hermes, D. (2018). *BIDS-iEEG: An extension to the brain imaging data structure (BIDS)*

specification for human intracranial electrophysiology [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/r7vc2

Horien, C., Noble, S., Greene, A. S., Lee, K., Barron, D. S., Gao, S., O'Connor, D., Salehi, M., Dadashkarimi, J., Shen, X., Lake, E. M. R., Constable, R. T., & Scheinost, D. (2021). A hitchhiker's guide to working with large, open-source neuroimaging datasets. *Nature Human Behaviour*, 5(2), 185–193. https://doi.org/10.1038/s41562-020-01005-4

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, 14(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Laird, A. R. (2021). Large, open datasets for human connectomics research: Considerations for reproducible and responsible data use. *NeuroImage*, 244, 118579. https://doi.org/10.1016/j.neuroimage.2021.118579

Lieberman, M. D., Berkman, E. T., & Wager, T. D. (2009). Correlations in Social Neuroscience Aren't Voodoo: Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, 4(3), 299–307. https://doi.org/10.1111/j.1745-6924.2009.01128.x

Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., Hardcastle, N., Wexler, J., Esteban, O., Goncavles, M., Jwa, A., & Poldrack, R. (2021). The OpenNeuro resource for sharing of neuroscience data. *ELife*, 10, e71774. https://doi.org/10.7554/eLife.71774

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., Kriegeskorte, N., Milham, M. P., Poldrack, R. A., Poline, J.-B., Proal, E., Thirion, B., Van Essen, D. C., White, T., & Yeo, B. T. T. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience*, 20(3), 299–303. https://doi.org/10.1038/nn.4500

NIMH Data Archive. (2021). *About Us*. NIMH Data Archive. https://nda.nih.gov/about/about-us.html

Niso, G., Gorgolewski, K. J., Bock, E., Brooks, T. L., Flandin, G., Gramfort, A., Henson, R. N., Jas, M., Litvak, V., T. Moreau, J., Oostenveld, R., Schoffelen, J.-M., Tadel, F., Wexler, J., & Baillet, S. (2018). MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. *Scientific Data*, 5(1), 180110. https://doi.org/10.1038/sdata.2018.110

NITRC. (2007). *The fMRI Data Center*. NeuroImaging Tools & Resources Collaboratory. https://www.nitrc.org/projects/fmridatacenter/

Notter, M. P., Costa, S. D., & Murray, M. M. (2019). *Decoding of multisensory semantics and memories in low-level visual cortex* [Dataset]. Openneuro. https://doi.org/10.18112/OPENNEURO.DS001345.V1.0.0

Pernet, C., Belin, P., McAleer, P., Gorgolewski, K., Valdes-Sosa, M., Latinus, M., Charest, I., Bestelmeyer, P., Watson, R., & Fleming, D. (2019). *The human Voice Areas: Spatial organisation and inter-individual variability in temporal and extra-temporal cortices* [Dataset]. Openneuro. https://doi.org/10.18112/OPENNEURO.DS000158.V1.0.0

Pernet, C., & Poline, J.-B. (2015). Improving functional magnetic resonance imaging reproducibility. *GigaScience*, 4(1), 15. https://doi.org/10.1186/s13742-015-0055-8

Pernet, C. R., Appelhoff, S., Flandin, G., Phillips, C., Delorme, A., & Oostenveld, R. (2018). BIDS-EEG: *An extension to the Brain Imaging Data Structure (BIDS) Specification for electroencephalography* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/63a4y

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2), 115–126. https://doi.org/10.1038/nrn.2016.167

Poldrack, R. A., & Gorgolewski, K. J. (2014). Making big data open: Data sharing in neuroimaging. *Nature Neuroscience*, 17(11), 1510–1517. https://doi.org/10.1038/nn.3818

Poldrack, R. A., & Gorgolewski, K. J. (2017). OpenfMRI: Open sharing of task fMRI data. *NeuroImage*, 144, 259–261. https://doi.org/10.1016/j.neuroimage.2015.05.073

Poldrack, R. A., & Mumford, J. A. (2009). Independence in ROI analysis: Where is the voodoo? *Social Cognitive and Affective Neuroscience*, 4(2), 208–213. https://doi.org/10.1093/scan/nsp011

Robbins, K., Truong, D., Jones, A., Callanan, I., & Makeig, S. (2020). *Building FAIR functionality: Annotating events in time series data using Hierarchical Event Descriptors* (HED) [Preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/5fg73

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Sadoun, A., Chauhan, T., Mameri, S., Zhang, Y. F., Barone, P., Deguine, O., & Strelnikov, K. (2020). Stimulus-specific information is represented as local activity patterns across the brain. *NeuroImage*, 223, 117326. https://doi.org/10.1016/j.neuroimage.2020.117326

Scargle, J. R. (2000). Publication bias: The "file-drawer" problem in scientific inference. *Journal of Scientific Exploration*, 14(1), 91–106.

Sengupta, A., Kaule, F. R., Guntupalli, J. S., Hoffmann, M. B., Häusler, C., Stadler, J., & Hanke, M. (2016). A studyforrest extension, retinotopic mapping and localization of higher visual areas. *Scientific Data*, 3(1), 160093. https://doi.org/10.1038/sdata.2016.93

Stanford Center for Reproducible Neuroscience. (2021). *OpenNeuro: FAQ*. OpenNeuro. https://openneuro.org/

Stanford CRN. (2021). *About the Center*. Stanford Center for Reproducible Neuroscience. https://reproducibility.stanford.edu/about-us/

Staniscuaski, F., Kmetzsch, L., Soletti, R. C., Reichert, F., Zandonà, E., Ludwig, Z. M. C., Lima, E. F., Neumann, A., Schwartz, I. V. D., Mello-Carpes, P. B., Tamajusuku, A. S. K., Werneck, F. P., Ricachenevsky, F. K., Infanger, C., Seixas, A., Staats, C. C., & de Oliveira, L. (2021). Gender, Race and Parenthood Impact Academic Productivity During the COVID-19 Pandemic: From Survey to Action. *Frontiers in Psychology*, 12, 663252. https://doi.org/10.3389/fpsyg.2021.663252

Tagliazucchi, E., & Laufs, H. (2014). Decoding Wakefulness Levels from Typical fMRI Resting-State Data Reveals Reliable Drifts between Wakefulness and Sleep. *Neuron*, 82(3), 695–708. https://doi.org/10.1016/j.neuron.2014.03.020

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, 4(3), 274–290. https://doi.org/10.1111/j.1745-6924.2009.01125.x

Walz, J. M., Goldman, R. I., Muraskin, J., Conroy, B., Brown, T. R., & Sadja, P. (2018). *Auditory and Visual Oddball EEG-fMRI* [Dataset]. Openneuro.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18