Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D1J31B

Title: FAIR annotated dataset of stroke MRIs, CTs, and metadata

Creator: Andreia Faria - **ORCID:** 0000-0002-1673-002X

Affiliation: Johns Hopkins University (jhu.edu)

Data Manager: Xin Li, Johnny Hsu, Brenda Johnson

Funder: Federation of American Societies for Experimental Biology (faseb.org)

Template: DataWorks! Data Management and Sharing Plan Challenge

Project abstract:

To extract meaningful and reproducible models of brain function from stroke images, for both clinical and research proposes, is a daunting task severely hindered by the great variability of lesion frequency and patterns. Large datasets are therefore imperative, as well as fully automated image post-processing tools to analyze them. The development of such tools, particularly with artificial intelligence, is highly dependent on the availability of large datasets to model training and testing. We will create and share a public dataset of 4,000 multimodal clinical MRIs and CTs of patients with acute and early subacute stroke, with manual lesion segmentation, and metadata. The dataset provides high quality, large scale, human-supervised knowledge to feed artificial intelligence models and enable further development of tools to automate several tasks that currently rely on human labor, such as lesion segmentation, labeling, calculation of disease-relevant scores. It also represents a valuable training and testing resource for translational research relating lesion features to risk factors, brain functions and patients' outcomes.

Start date: 10-02-2022

End date: 09-30-2027

Last modified: 10-15-2022

Copyright information:

FAIR annotated dataset of stroke MRIs, CTs, and metadata

Data Type

A general summary of the types and estimated amount of scientific data to be generated and/or used in the research. Describe data in general terms that address the type and amount/size of scientific data expected to be collected and used in the project (e.g., 256-channel EEG data and fMRI images from ~50 research participants). Descriptions may indicate the data modality (e.g., imaging, genomic, mobile, survey), level of aggregation (e.g., individual, aggregated, summarized), and/or the degree of data processing that has occurred (i.e., how raw or processed the data will be)

We will organize and share a dataset of 4,000 clinical MRIs and metadata of patients admitted at the Johns Hopkins Comprehensive Stroke Center (CSC) with acute or early subacute stroke, from 2009-2022. The CSC will provide the metadata (basic demographic and clinical information) in standardized format that they regularly record in the admission and discharge of patients, as part of their participation on the national "Get with the Guidelines" (GWTG) stroke program. The images will consist of brain MRIs performed at admission, de-identified and defaced. In addition to the images in the native space, we will offer the images mapped to common coordinates and intensity normalized, since these are common steps required for most of the imaging processing pipelines. We will also share the annotation of the stroke core and a structured radiological description. The data format and organization follow the Brain Imaging Data Structure, BIDS, guidelines facilitating navigation and sharing. The anonymized data will be shared at individual level, following "FAIR" principles. The first part of the collection, 2888 records (~180 GB) are currently deposited and in curation at ICPSR (https://doi.org/10.3886/ICPSR38464), expected release in late 2022.

A description of which scientific data from the project will be preserved and shared.

The dataset provides high quality, large scale, human-supervised knowledge to feed artificial intelligence models and enable further development of tools to automate several tasks that currently rely on human labor, such as lesion segmentation, labeling, calculation of disease-relevant scores, and lesion-based studies relating function to frequency lesion maps. This is the first, largest, clinical dataset of brain MRIs associated with clinical information ever shared. It offers unique possibilities for education, clinical modeling, and research, promoting reproducibility and enabling access to users of diverse expertise. It shows how available information and services can be reused to generate and share new data, and that these data can be shared FAIRly, despite technical and regulatory issues, opening endless opportunities for clinical, translational, and biomedical research.

A brief listing of the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.

The metadata includes demographic information, basic clinical profile (NIHSS scores, hospitalization duration, biometric screening at hospital admission and discharge, and associated health conditions), and expert, standardized description of the infarct in the MRIs. It is archived in tab-separated values (TSV) format, following the BIDS recommendation and is accompanied by a data dictionary. Free text fields were not included. A "readme" file, a pre-print (https://www.researchsquare.com/article/rs-1705779/v1) submitted to publication in Nat Sci Data, with all the possible details about the dataset creation and organization will be linked to the repository. The "change" file will dynamically describe all the future changes, updates, and version numbers

Related Tools, Software and/or Code

An indication of whether specialized tools are needed to access or manipulate shared scientific data to support replication or reuse, and name(s) of the needed tool(s) and software.

The data can be accessed by direct download from ICPSR. To manipulate the raw image data, researchers might use public and free software commonly used for image analysis, such as MRIcron, MRIStudio, FSL, etc. Sharing these data as "computational data objects" (CDO) in BIDS format facilitates code and batch processing

If applicable, specify how needed tools can be accessed, (e.g., open-source and freely available, generally available for a fee in the marketplace, available only from the research team) and, if known, whether such tools are likely to remain available for as long as the scientific data remain available.

Although no specific tools are needed, a variety of public tools for image manipulation are widely available online. Because this dataset is suited to researchers in a variety of fields, including clinical researchers, epidemiologists, and not exclusively image experts, our group developed several free, public, user-friendly tools to generate the computational image objects that can be readily used by people with diverse expertise. These include the first 3D digital atlas of arterial territories (https://www.nitrc.org/projects/arterialatlas) and the "Acute-stroke Detection and Segmentation" tool (ADS). Future tools for calculation of disease relevant scores, prognostic modeling, and engines for image searching are being developed and will be available using the same free and public model

Standards

An indication of what standards will be applied to the scientific data and associated metadata (i.e., data formats, data dictionaries, data identifiers, definitions, unique identifiers, and other data documentation).

The images will be shared in native subject space and in standard space (Montreal Neurological Institute, MNI), in Nfiti compressed format. The data format and organization follow the Brain Imaging Data Structure, BIDS, guidelines, which will enable navigation and sharing. Additionally, this format is compatible for conversion to newly developed semantic standards. This allows generating computable data objects (CDOs) readily used by AI community, and facilitate non-expert data analysis. The CDO's also

allow integration with open science efforts, analytical pipelines, APIs, and modules for quality control, harmonization, indexing and management engines

Data Preservation, Access, and Associated Timelines

The name of the repository(ies) where scientific data and metadata arising from the project will be archived.

The dataset will be shared at the Inter-University Consortium for Political and Social Research, ICPSR, a professionally managed data repository operating over 60 years and among NIH's recommended repositories

How the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools.

ICPSR provides metadata and persistent identifiers using DOI and citations. Their expanding collections in public health and imagery enhance findability for our dataset. A publication (https://www.researchsquare.com/article/rs-1705779/v1), submitted to Nat Sci Data, will be associated to the dataset. All our scientific productions mentioning the dataset will refer to the repository link and DOI. The availability of this dataset will be advertised in Nitrc and other neuroscience and neuroimaging repositories and resources.

When the scientific data will be made available to other users (i.e., the larger research community, institutions, and/or the broader public) and for how long.

The first part of these data (2,888 annotated multimodal clinical MRIs and metadata) is in curation at ICPSR expected release by 2023 (https://doi.org/10.3886/ICPSR38464). The second set of data will be shared as soon as it is ready; there will be no postponement. There are no plans to remove or restrict access in the future.

Access, Distribution, or Reuse Considerations

Describe any applicable factors affecting subsequent access, distribution, or reuse of scientific data related to whether access to scientific data derived from humans will be controlled (i.e., made available by a data repository only after approval).

The images are fully de-identified by removing all HIPAA-protected direct and indirect identifiers. The original DICOM files are converted to compressed ".nii.gz" / ".json" using dcm2niix (https://github.com/rordenlab/dcm2niix) with the anonymization option according BIDS guidelines. The ".json" preserves the technical information from the image header. All the high resolution T1-WI MPRAGE, and the low resolution T1-WIs and FLAIR with full head coverage are defaced using FSL (https://surfer.nmr.mgh.harvard.edu/fswiki/mrideface). Another round of visual quality control was preformed to secure complete anonymization of possible face recognition by 3D reconstruction. In consultation with the Johns Hopkins Data Services librarians (partners in this project), we removed sensitive information and possible remaining indirect identifiers, making the data compatible with HIPAA "Safe Harbor" regulations.

Johns Hopkins IRB and JHM Data Trust committee reviewed and approved the de-identification and sharing plan. Because these data were originally assembled under a waiver of consent Johns Hopkins policy requires restricted access for approved research. ICSPR will release the data as a restricted-use collection under a Data Use Agreement (DUA). Sharing is for research use, restricted to biomedical research in academic institutions, requiring that the source is cited in future publications. ICPSR's researcher approval process is described at https://www.icpsr.umich.edu/web/pages/ICPSR/access/restricted/. Researchers from institutions without an ICPSR membership must pay an access fee for the data. This project will assess funding to cover access fees for requests as feasible. ICPSR employs industry standard security and stringent access requirements. ICPSR will monitor data usage and citations that reference or reuse the collection.

Oversight of Data Management and Sharing

Indicate how compliance with the Plan will be monitored and managed, frequency of oversight, and by whom (e.g., titles, roles).

I, Dr. Andreia Faria, will be responsible for implementing the DMP, and ensuring it is reviewed and revised. The Johns Hopkins Stroke Center (represented by Dr. Victor Urrutia, Center Director, and Ms. Brenda Johnson, co-Director) will be responsible for data capturing and metadata generation. Supported by the Department of Radiology and the Johns Hopkins IT, my research assistant, Mr. Johnny Hsu, will be responsible by automatically capture and transfer the images. My research assistant, Ms. Xin Li, will be responsible by anonymizing and archiving under secure firewall. My research assistants, Ms. Xin Xu and Mr. Hsu and I will be responsible by the annotation. The Johns Hopkins Data Service advises on regulatory issues and assists with data deposit in ICPCR. ICSPR will monitor usage and DUA compliance.

Planned Research Outputs

Dataset - "ANNOTATED CLINICAL MRIS AND LINKED METADATA OF PATIENTS WITH ACUTE STROKE, BALTIMORE, MARYLAND, 2009-2019"

A dataset with 2,888 annotated multimodal clinical MRIs and metadata, deposited in ICPSR (https://doi.org/10.3886/ICPSR38464). The public release, by curation completion, is expected by the end of 2022

Software - "ACUTE-STROKE DETECTION AND SEGMENTATION"

This a free, public, user-friend tool developed with this dataset, to automatically detect, segment and quantify acute ischemic strokes. It converts images in CDO (computable data objects) enabling direct AI modeling. It runs in real time, in regular local computers, with a single command line, facilitating the use by non-experts and researchers in diverse fields and enabling easy reproducible and replicable research. Recently added features are the output of radiological reports and the stroke score ASPECTS. Currently this resource has more than 400 downloads in Nitrc. Accessible at https://www.nitrc.org/projects/ads

Interactive resource - "ARTERIAL ATLAS"

This is the first digital 3D atlas of brain arterial territories, developed with this dataset, currently with more than 700 dowloads all around the world. Accessible at https://www.nitrc.org/projects/arterialatlas

Data paper - "A LARGE PUBLIC DATASET OF ANNOTATED CLINICAL MRIS OF PATIENTS WITH ACUTE STROKE AND LINKED METADATA"

Description of the StrokeFAIR dataset. Pre-print available in Research Square https://www.researchsquare.com/article/rs-1705779/v1. Submitted to Nature Scientific Data.

Data paper - "DEEP LEARNING-BASED DETECTION AND SEGMENTATION OF DIFFUSION ABNORMALITIES IN ACUTE ISCHEMIC STROKE"

Liu, CF., Hsu, J., Xu, X. et al. Deep learning-based detection and segmentation of diffusion abnormalities in acute ischemic stroke. Commun Med 1, 61 (2021). https://doi.org/10.1038/s43856-021-00062-8. https://www.nature.com/articles/s43856-021-00062-8. This paper describes a tool, created with this dataset, for automated detection, segmentation, and quantification of acute strokes. It outputs the 3D mask of the stroke core and the ratio of brain regions affected by the stroke in two parcelation schemes (classical anatomy and arterial territories). This toll is free and publicly available, runs in local computers, is user-friendly to non-experts and people in diverse field, facilitating reproducible and replicable research

Data paper - "DIGITAL 3D BRAIN MRI ARTERIAL TERRITORIES ATLAS"

Liu CF, Hsu J, Xu X, Kim NG, Sheppard SM, Meier EL, Miller M, Hillis AE, Faria AV. Digital 3D Brain MRI Arterial Territories Atlas. https://www.biorxiv.org/content/10.1101/2021.05.03.442478v2.full.pdf. In Press in Sci Data. This paper describes the fisrt digital 3D atlas of brian arterial territories, created with this dataset

Data paper - "AUTOMATIC COMPREHENSIVE ASPECTS REPORTS IN CLINICAL ACUTE STROKE MRIS"

Faria, A., Liu, C. F., Li, A., Kim, G., Miller, M., & Hillis, A. (2022). Automatic Comprehensive Aspects Reports in Clinical Acute Stroke MRIs. Pre-print in Research Square. This paper describes our software to calculate ASPCTS in patients with ischemic strokes, using this dataset. Under revision in Sci Reports

Data paper - "AUTOMATIC COMPREHENSIVE RADIOLOGICAL REPORTS FOR CLINICAL ACUTE STROKE MRIS"

Liu, C. F., Zhao, Y., Miller, M. I., Hillis, A. E., & Faria, A. (2022). Automatic comprehensive radiological reports for clinical acute stroke MRIs. Available at Research Square. This manuscript describes our free and public software to generate automated reports, developed with this dataset. Under revision in Nat Comm Med

Software - "DETECTION AND SEGMENTATION OF PERFUSION DEFICITS IN CLINICAL ACUTE STROKE MRIS"

This tool will detect, segment and quantify perfusion deficits in perfusion-weighted images. It will output not only volumetric information but also the 3D masks of perfusion deficits, to be used on clinical research, particurly to correlate with function or to predcit outcomes

Software - "MRI-BASED PREDICTION OF TIME TO STROKE ONSET"

This tool will automatically calculate time to stroke onset based on brian MRIs. The time to onset is a crucial informaiton for acute treatment, but is unknown in about 25% of patients with ischemic stroke, which prevents them to be treated. As our other tools, this will be free and publicly available, and accessible to expert and on-expert users

Software - "CONTENT-BASED IMAGE RETRIEVAL (CBIR) FOR ACUTE STROKE MRIs"

Using our "ADS" system already in place, this tool will automatically calculate transform the original data in a computational data object and search in our dataset (which will be then a library of thousands of computational data objects) for the clusterof similar cases. Then, it will output frequence reports of relevant information in this cluster, for instance, 90-days follow up scores and response to acute treatment. It will also be useful to identify specific risk-factors, as it is linked to the metadata. This will enable population stratification and personalized medical approach. As our other tools, it will be free and publicly available, and accessible to expert and on-expert users

Software - "SYNTHETIC CTs FOR QUANTIFICATION OF ACUTE ISCHEMIC STROKE BASED ON DIFFUSION MRISSYNTHETIC CTs FOR QUANTIFICATION OF ACUTE ISCHEMIC STROKE BASED ON DIFFUSION

MRIS"

By cost reasons, computed tomography (CT) is still the first image acquired for most of patients with acute stroke. However, if it less sensitive than MRIs in the hyper acute stage, which leads to misdiagnosis or underestimation of the stroke volume. Using our dataset, we will create an artificial intelligence tool (likely based on deep-learning) to estimate the volume of the ischemic stroke in CTs, based on the information provided by the diffusion weighted images. By submitting a CT showing a questionable stroke core, the user will receive what the MRI would look-like and the predicted area and volume of the stroke core in the CT. As our other tools, this will be free and publicly available, and accessible to expert and on-expert users.

Data paper - "IMAGE AND NON-IMAGE DETERTMINANTS OF NIHSS IN ACUTE ISCHEMIC STROKES"

This paper is an analysis of image and metadata in our database to identify the factors that most influence the calculation of NIHSS scores in patients with acute stroke, an index of stroke severity that is relevant for treatment and outcome prediction.

Data paper - "VARIOUS TESTS OF LEFT NEGLECT ARE ASSOCIATED WITH DISTINCT TERRITORIES OF HYPOPERFUSION IN ACUTE STROKE"

Stein, C., Bunker, L., Chu, B., Leigh, R., Faria, A., & Hillis, A. E. (2022). Various tests of left neglect are associated with distinct territories of hypoperfusion in acute stroke. Brain communications, 4(2), fcac064. This paper exemplifies the use of our dataset, and of the tools for automated analysis derived from it, to stablish anatomic-functional correlation in stroke patients.

Data paper - "LEFT HEMISPHERE BIAS OF NIH STROKE SCALE IS MOST SEVERE FOR MIDDLE CEREBRAL ARTERY STROKES"

https://www.frontiersin.org/articles/10.3389/fneur.2022.912782/full. Using this dataset, we found a hemisphere-related bias in a largely used clinical indice for strokes, the NIHSS.

Data paper - "ASSOCIATION OF INFERIOR DIVISION MCA STROKE LOCATION WITH POPULATIONS WITH ATRIAL FIBRILLATION INCIDENCE"

Kim, G., Vitti, E., Stockbridge, M. D., Hillis, A. E., & Faria, A. V. (2021). Association of inferior division MCA stroke location with populations with atrial fibrillation incidence. Available at https://www.medrxiv.org/content/10.1101/2021.12.06.21267371v1. This paper describes a demographic association of the stroke location in our dataset. Under revision in Heliyon

Data paper - "POOR GLYCEMIC CONTROL IS ASSOCIATED WITH WORSE BLOOD-BRAIN BARRIER DISRUPTION IN ISCHEMIC STROKE PATIENTS"

We describe the association between metabolic profile and an MRI indice of blood perfusion, found in an external small dataset and confirmed in our dataset. This proves its value as an external test set for replication studies. The abstract will be presetned in the ANA Oct 2022, Chicago

Data paper - "WORSE BLOOD-BRAIN BARRIER DISRUPTION IN ISCHEMIC STROKE IS ASSOCIATED WITH LONGER HOSPITAL LENGTH OF STAY"

This manuscript describes an association between metabolic profile and an MRI indice of blood perfusion, found in an external small dataset and confirmed in our dataset. This proves its value as an external test set for replication studies. The abstract will be presented in the WSC Oct 2022, Singapore

Software - "AUTOMATIC RADIOLOGICAL REPORTS FOR ACUTE ISCHEMIC STROKES"

https://www.nitrc.org/projects/ads. The second version of "ADS" (ADSv1) additionally offers imaging mapping to different standard spaces, the generation of automated radiological reports. It outputs the feature vectors used and comprehensive reports of the classification results. It runs in real time, in regular local computers, with a single command line, facilitating the use by non-experts and researchers in diverse fields and enabling easy reproducible and replicable research.

Software - "AUTOMATIC CALCULATION OF ASPECTS FOR ACUTE ISCHEMIC STROKES IN BRIAN MRIs"

https://www.nitrc.org/projects/ads. The second version of "ADS" (ADSv1) additionally offers the calculation of "ASPECTS" for brain MRIs, which is an important index for acute treatment and prognosis. It outputs the feature vectors used and comprehensive reports of the classification results. It runs in real time, in regular local computers, with a single command line, facilitating the use by non-experts and researchers in diverse fields and enabling easy reproducible and replicable research.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
ANNOTATED CLINICAL MRIS AND LINKED METADATA OF PAT ...	Dataset	2022-12-04	Restricted	Inter-university Consortium for Political and Social Research	180 GB	None specified	None specified	No	No
ACUTE-STROKE DETECTION AND SEGMENTATION	Software	2021-05-04	Open	NITRC		None specified	None specified	No	No
ARTERIAL ATLAS	Interactive resource	2021-05-04	Open	NITRC		None specified	None specified	No	No
A LARGE PUBLIC DATASET OF ANNOTATED CLINICAL MRIS ...	Data paper	2023-01-13	Open	None specified		None specified	None specified	No	No
DEEP LEARNING-BASED DETECTION AND SEGMENTATION OF ...	Data paper	2021-12-05	Open	None specified		None specified	None specified	No	No
DIGITAL 3D BRAIN MRI ARTERIAL TERRITORIES ATLAS	Data paper	2022-12-01	Open	None specified		None specified	None specified	No	No
AUTOMATIC COMPREHENSIVE ASPECTS REPORTS IN CLINICA ...	Data paper	2022-12-14	Open	None specified		None specified	None specified	No	No
AUTOMATIC COMPREHENSIVE RADIOLOGICAL REPORTS FOR C ...	Data paper	2022-12-13	Open	None specified		None specified	None specified	No	No
DETECTION AND SEGMENTATION OF PERFUSION DEFICITS I ...	Software	2021-05-20	Open	NITRC		None specified	None specified	No	No
MRI-BASED PREDICTION OF TIME TO STROKE ONSET	Software	2023-06-13	Open	NITRC		None specified	None specified	No	No
CONTENT-BASED IMAGE RETRIEVAL (CBIR) FOR ACUTE STR ...	Software	2023-09-13	Open	NITRC		None specified	None specified	No	No
SYNTHETIC CTs FOR QUANTIFICATION OF ACUTE ISCHEMIC ...	Software	2023-12-13	Open	NITRC		None specified	None specified	No	No
IMAGE AND NON-IMAGE DETERTMINANTS OF NIHSS IN ACUT ...	Data paper	2023-06-13	Open	None specified		None specified	None specified	No	No
VARIOUS TESTS OF LEFT NEGLECT ARE ASSOCIATED WITH ...	Data paper	2022-08-13	Open	None specified		None specified	None specified	No	No
LEFT HEMISPHERE BIAS OF NIH STROKE SCALE IS MOST S ...	Data paper	2022-08-13	Open	None specified		None specified	None specified	No	No
ASSOCIATION OF INFERIOR DIVISION MCA STROKE LOCATI ...	Data paper	2022-12-13	Open	None specified		None specified	None specified	No	No
POOR GLYCEMIC CONTROL IS ASSOCIATED WITH WORSE BLO ...	Data paper	2022-10-27	Open	None specified		None specified	None specified	No	No
WORSE BLOOD-BRAIN BARRIER DISRUPTION IN ISCHEMIC S ...	Data paper	2022-10-27	Open	None specified		None specified	None specified	No	No
AUTOMATIC RADIOLOGICAL REPORTS FOR ACUTE ISCHEMIC ...	Software	2022-08-11	Open	NITRC		None specified	None specified	No	No
AUTOMATIC CALCULATION OF ASPECTS FOR ACUTE ISCHEMI ...	Software	2022-08-11	Open	NITRC		None specified	None specified	No	No