DATA MANAGEMENT AND SHARING PLAN

An example from an application focusing on secondary data analysis on data from human subjects.

If any of the proposed research in the application involves the generation of scientific data, this application is subject to the NIH Policy for Data Management and Sharing and requires submission of a Data Management and Sharing Plan. If the proposed research in the application will generate large-scale genomic data, the Genomic Data Sharing Policy also applies and should be addressed in this Plan. Refer to the detailed instructions in the application guide for developing this plan as well as to additional guidance on sharing.nih.gov. The Plan is recommended not to exceed two pages. Text in italics should be deleted (but this has not been done in the sample below). There is no "form page" for the Data Management and Sharing Plan. The DMS Plan may be provided in the *format* shown below.

Element 1: Data Type

A. Types and amount of scientific data expected to be generated in the project:

Summarize the types and estimated amount of scientific data expected to be generated in the project.

The data to be shared will include MRI images and clinical assessments from human research participants. This application is focused on secondary data analysis from existing data but will also deposit privately held data to a public repository. The existing data is available from the NIMH Data Archive (NDA) in collections 2134 (148 subjects) and 2433 (47 subjects). In addition, we have data from a previous study involving 155 research participants with major depressive disorder that have not yet been shared with the research community but will be uploaded to NDA during the second quarter of the first year of funding. As discussed in the application, structural MRI scans are available for time points before and after treatment along with relevant clinical data.

B. Scientific data that will be preserved and shared, and the rationale for doing so:

Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.

This is a secondary data analysis application, so new data is not being measured. Much of the data is already available through NDA. Clinical and imaging data from 155 new subjects will be shared.

C. Metadata, other relevant data, and associated documentation:

Briefly list 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.

Preparation for submitting existing data to NDA is largely complete. Within the first six months following the award, we will submit the Data Submission Agreement to NDA and will create the Data Expected list (see Standards section) in our new NDA Collection. The policies of our institution mandate that exact dates will not be shared (see Access section).

Element 2: Related Tools, Software and/or Code:

State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed.

The basic statistical analyses described in the application will be done using R. We plan to use the MRI data analysis tools in the FMRIB Software Library (FSL) for multi-level modeling of group effects. BrainVoyager software will be used for anatomical segmentation to isolate regions of interest within individual subjects, and the AI-powered analyses described in the application will use custom code written with the PyTorch library for Python. R, FSL, Python, and PyTorch are all freely available to the research community. BrainVoyager is commercial software, with licenses available for purchase from https://www.brainvoyager.com/.

All R and Python code (including trained model weights) will be available on our lab Bitbucket page (located at https://www.thelab.edu/) no later than when publications are submitted. The Bitbucket page is publicly assessable and will be hosted for at least 5 years after the grant award ends.

Element 3: Standards:

State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by

the research proposed in this project. If applicable, indicate that no consensus standards exist.

The data that will be used for some of the proposed secondary data analysis is already in NDA and is formatted using NDA data dictionaries. The new data we will deposit will also use existing NDA data dictionaries. Since the data set to be deposited into NDA was collected prior to the publication of NOT-MH-20-067, not all of the common data elements expected by NIMH are available. However, we will transform some existing demographic and clinical data into the formats expected for:

1) Age (ndar_subject01)
2) Sex at Birth (ndar_subject01)
3) Patient Health Questionnaire-9 (PHQ-9, cde_phq901 NDA data dictionary).

In addition, information from the Beck Depression Inventory will be deposited for all 155 research participants using the NDA bdi01 data dictionary. Deposited images will use the NDA image03 data dictionary. Data derived from the MRI images will be deposited into NDA using the imagingcollection01 data dictionary.

Element 4: Data Preservation, Access, and Associated Timelines

A. Repository where scientific data and metadata will be archived:
Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived; see Selecting a Data Repository).

All previously unshared data will be deposited to NDA no later than 12 months after the award begins.

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

Data will be findable for the research community through the NDA collection that will be established when this application is funded. For all publications, an NDA study will be created, and the data relevant to that publication will be shared immediately. Each of those studies is assigned a digital object identifier (DOI). This data DOI will be referenced in the publication to allow the research community easy access to the exact data used in the publication.

C. When and how long the scientific data will be made available:
Describe when the scientific data will be made available to other users (i.e., no later than time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available

The research community will have access to the previously unshared data at the end of the grant award. Researchers will request data using the standard processes at NDA, and the NDA data access committee will decide which requests to grant. The standard NDA data access process allows access for one year and is renewable. Once the data are submitted to NDA, that archive will control the long-term persistence of the data set.

Element 5: Access, Distribution, or Reuse Considerations

A. Factors affecting subsequent access, distribution, or reuse of scientific data:

NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing. See Frequently Asked Questions for examples of justifiable reasons for limiting sharing of data.

The two existing data sets from NDA used consents that allow broad data sharing. The new dataset to be uploaded to NDA also was collected using informed consent terms that allow broad data sharing. Access to data housed by the NDA requires the completion of a Data Use Certification (https://nda.nih.gov/faq.html#dac.3), which prohibits any redistribution or attempts to re-identify research participants.

B. Whether access to scientific data will be controlled:

State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).

To request access of the data, researchers will use the standard processes at NDA, and the NDA Data Access Committee will decide which requests to grant. The standard NDA data access process allows access for one year and is renewable. Once the data are submitted to NDA, that archive will control the long-term persistence of the data set. Currently, NDA has no process for deleting or retiring data sets.

C. Protections for privacy, rights, and confidentiality of human research participants:

If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de-identification, Certificates of Confidentiality, and other protective measures).

The NDA GUID tool allows researchers to aggregate data from the same research participant without different laboratories having to share personally identifiable information about that research participant. The NDA data dictionaries do not permit personally identifiable information to be shared. NDA maintains a Certificate of Confidentiality.

For the 155 participants from our previous study, exact dates have been obscured via the Shift and Truncate method [1], which preserves within-case temporal relations.

Element 6: Oversight of Data Management and Sharing:

Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by whom at your institution (e.g., titles, roles).

The Office of Sponsored Programs at University X has created a data management and sharing plan compliance system as part of their process for submitting the annual NIH progress report. That Office is collecting information related to the number of research participants that are deposited each reporting year. For this award, all of the data will be uploaded in the first year, so the data deposition oversight will end then. The Office of Sponsored Programs will look for the NDA data DOIs when papers are published and will include that information in the annual progress report.

Validation Schedule (this section is required by NIMH)

Since this is a secondary data analysis application, validation of newly collected data will not occur. The new data to be deposited to NDA will go through their validation tool when the data are initially uploaded.

Reference

1. Hripcsak, G., Mirhaji, P., Low, A. F., & Malin, B. A. (2016). Preserving temporal relations in clinical data while

maintaining privacy. *Journal of the American Medical Informatics Association*, *23*(6), 1040-1045.