HEAL Data2Action Network: Addiction and Chronic Pain Research

Data Type

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.*

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)

Example Answer:

This project includes several types of data:

- **Secondary surveillance data** from Named County Data Commons, including health, service, and community data gathered from the public health department, abstracted into dashboards ("Surveillance data"), to assess geographic burden and service delivery gaps. Data are centrally de-identified by the health department, available through a data use agreement, and will be summarized and visualized at the community level by zip code. Data are secondary data available through data use agreement from the Named County Data Commons and will not be submitted to a repository. Dashboard visualizations will be available through public access via Named County Website.
- **Secondary prescribing data** from IQVIA (National Prescription Audit), to assess prescribing practices in local areas ("prescribing data"). Data are secondary data, purchased through IQVIA with a data use agreement, and will not be submitted to a repository. Summary data will be made available to the scientific community via study publications including summary data tables.
- **Secondary ABC Health System electronic health record data**, de-identified by the health system and analyzed at a system level to assess patient and provider characteristics to inform the recruitment and analysis strategy ("EHR systems data", N = ~ 10,000). Data stored on secured servers in the ABC Health System, available through a data use agreement, and accessible locally only; thus, EHR systems data will not be submitted to arepository. Summary data used to inform recruitment and analysis will be made available to the scientific community via study publications including summary data tables.
- **Primary qualitative data from interviews and focus groups** with providers and patients ("Qualitative data"; N = 50 provider interviews, 50 patient interviews, 5 provider focus groups of 8 participants, 5 patient focus groups of 8 participants). Interviews and focus groups will be recorded and transcribed, and recordings destroyed after transcription. Transcripts will be coded, data entered into NVivo, and summary data will be exported into Excel. Summary data will be de-identified for analysis and Excel files will be submitted to the NAHDAP repository and made available to the scientific community via restricted use access (data use agreement required).
- **Clinical outcome data** from clinical trial patients (including both treatment and control participants, N=250 patients) will be derived from the electronic health record and collected through patient-reported outcomes. This data will be entered into a SAS file, de-identified for analysis, and submitted to the NAHDAP repository.

Guidance:

NIH Guidance

The final DMS Policy has specific definitions for what Scientific Data is, and what proposals are considered to be producing scientific data.

Per the Policy, "Even those scientific data not used to support a publication are considered scientific data and within the final DMS Policy's scope. We understand that a lack of publication does not necessarily mean that the findings are null or negative; however, indicating that scientific data are defined independent of publication is sufficient to cover data underlying null or negative findings."

NIH Genomic Data Sharing (GDS) Policy Considerations

Check if your research is subject to NIH GDS (Genomic Data Sharing) policy usingthis criteria and list those data and the levels of processing here.

Individual NIH Institutes and Centers (IC) may have additional expectations or requirements for genomic data sharing as well. Please check the IC-specific genomic data sharing requirements.

NIH HEAL Initiative Specific Considerations

Researchers should consider their intended choice of data repository when summarizing data types, sizes, and formats in this section. Most HEAL-compliant repositories have guidelines for which data types should be included in submitted data sets. Refer to these guidelines to ensure that the appropriate repository has been selected for the intended data types. Contact the desired repository for specific guidelines and expectations of data submission.

The recommended repository for HEAL projects studying addiction is the National Addiction & HIV Data Archive Program (NAHDAP), and researchers planning to use this archive should budget accordingly. Researchers and NAHDAP staff will collaborate to determine whether data should be stored as a public-use file or a restricted-use file, depending on the presence of confidential information in the data.

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.*

Example Answer:

Primary demographic, medical history, prescribing, and clinical outcome data from clinical trial patients will be abstracted from the electronic health record (derived data) and collected through patient-reported outcomes, including both treatment and control participants ("Clinical outcome data"; N = 250 patients). Data will be at the individual patient level, entered into a SAS file, and de-identified for analysis. Individual level-data in SAS files will be submitted to the NAHDAP repository and made available to the scientific community via restricted access (data use agreement required). Summarized qualitative data from patient and provider interviews and focus groups will be submitted via Excel files. Summary data tables will be submitted to ClinicalTrials.gov.

Guidance:

NIH Guidance

NIH does not anticipate that researchers will preserve and share all scientific data generated in a study. Researchers should decide which scientific data to preserve and share based on ethical, legal, and technical factors that may affect the extent to which scientific data are preserved and shared. Provide the rationale for these decisions.

NIH GDS Policy Considerations

If you are generating genomic data, follow specific sharing requirements (data submission and release expectation) under the NIH GDS policy (five levels of processing and associated expectations for data submission and release).

NIH HEAL Initiative Specific Considerations

Each HEAL-compliant repository that offers controlled access also has expectations and considerations for data de-identification. For example, according to NAHDAP, researchers should consider the degree of identification contained in their data when determining the extent to which the data may be shared. Any identifiers in the collected data should be examined for their usefulness towards the research analysis so that unnecessary identifiers do not prevent important data from being available. Additionally, some researchers may choose to submit both a public-use and a restricted-use version of their data. This allows some information to be shared broadly, while reserving more detailed information for certain approved researchers in a controlled process.

When constructing their DMP, researchers should provide justification for inclusion or exclusion of specific expected indirect identifiers in their deposited data.

NAHDAP characterizes identifiers in two categories: direct and indirect. Direct identifiers are variables such as names, addresses, and phone numbers, and they must be removed or masked from any data that will be stored in a repository. Indirect identifiers include variables such as exact occupations, exact dates of events, and detailed geographic information. These must be considered carefully, but they may be included in deposited data if excluding them will detract from the data analysis. NAHDAP will examine any identifiers for their safety, and they may recode them for further de-identification. For example, they may convert an exact date to a range of dates.

Additional Guidance from DMPTool

If human subjects data will be collected and only de-identified subsets are to be shared, consider specific de-identification approaches that fit the population and purposes. Guidance on protecting privacy is at NOT-OD-22-213.

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.*

Example Answer:

Metadata: Substantive metadata for the qualitative and clinical outcome data will be provided in

compliance with the Data Documentation Initiative, XML standard. This allows for tagging of content to facilitate preservation and flexibility in display. Codebooks and other study documentation and metadata will be made available on the ICPSR and NAHDAP websites alongside the data. The following type of metadata will be produced and archived:

- Study-level metadata record: A summary DDI-based record will be created for inclusion in the searchable ICPSR/NAHDAP catalog, indexed with terms from the ICPSR Thesaurus to enhance data discovery.
- Data citation with digital object identifier (DOI): A standard citation will be provided to facilitate attribution. The DOI provides permanent identification for the data and ensures that they will always be found at the URL specified.
- Variable-level documentation: ICPSR will tag variable-level information in DDI format for inclusion in ICPSR's Social Science Variables Database (SSVD), which allows users to identify relevant variables and studies of interest.
- Technical documentation: The variable-level files described above will serve as the foundation for the technical documentation or codebook that ICPSR will prepare and deliver.
- Related publications: Resources permitting, ICPSR will periodically search for publications based on the data and provide two-way linkages between data and publications.

Promptly after receiving funding, the researcher will register the study within the HEAL Data Platform and submit the associated study-level metadata through the CEDAR form. Metadata will be updated in the CEDAR form as needed throughout the study.

Guidance:

NIH Guidance

In addition to the documentation examples, consider metadata that will provide additional information intended to make scientific data interpretable and reusable (e.g., date, independent sample and variable construction and description, methodology, data provenance, data transformations, any intermediate or descriptive observational variables).

NIH HEAL Initiative Specific Considerations

All HEAL awardees are required to submit study-level metadata through the CEDAR form after registering the study within the HEAL Platform. Updates must be made at the time of any study data release. Researchers should acknowledge their plan for submitting their metadata in the DMP.

Guidance on completing the CEDAR form can be found here. Details on the HEAL metadata can be found here.

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.

Example Answer:

Software packages to enable analysis of qualitative and clinical outcome data include Excel, SAS, SPSS, and Stata. Code for manipulating secondary EHR system data will be published on GitHub and linked in publications.

Guidance:

NIH HEAL Initiative Specific Considerations

To promote collaboration and information sharing, the HEAL Initiative encourages HEAL researchers to use tools, software, and/or code that can be made broadly accessible. Researchers should prioritize the most accessible software that will meet their data analysis requirements. Sharing of code and other digital artifacts is highly encouraged, and plans to do so can be included in DMPs as relevant.

Additional Guidance from DMPTool

Tool(s) and software should be identified, then plans should specify how the 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). When known, the longevity or period of time for which custom or proprietary tools will be available should be addressed.

In addition, file formats in which data are saved in a digital format can be divided into two general categories.

Proprietary - The specification of the data encoding format is not released or restricted in some way. Proprietary formats can only be easily opened and manipulated by particular software tools.
Open - The specification of the data encoding format which can be used and implemented by

anyone. Open formats can often be easily opened and manipulated by a large number of software tools.

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.

Example Answer:

The data and documentation will be submitted to ICPSR in recommended formats. ICPSR will make the data files available in several widely used formats, including ASCII, tab-delimited (for use with Excel), SAS, SPSS, Stata, and R. Documentation will be provided as PDF. Substantive metadata for the qualitative and clinical outcome data will be provided in compliance with the Data Documentation Initiative, XML standard. Questionnaires and case report forms that include pain-related outcomes will capture pain Common Data Elements (CDEs) and will be submitted to the HEAL Clinical Data Elements (CDE) Program. This allows for tagging of content to facilitate preservation and flexibility in display. NAHDAP also uses controlled vocabularies to ensure consistency across studies.

Guidance:

NIH Guidance

While many scientific fields have developed and adopted common data standards, others have not. In such cases, the Plan may indicate that no consensus data standards exist for the scientific data and metadata to be generated, preserved, and shared.

NIH HEAL Initiative Specific Considerations

According to the NIH HEAL Initiative, "new HEAL clinical pain studies are required to submit their case-report forms/questionnaires to the HEAL Clinical Data Elements (CDE) Program. The program creates CDE files containing standardized variable names, responses, coding, and other information. HEAL Initiative clinical studies using copyrighted questionnaires must obtain use licenses, and share them with the HEAL CDE team and program officer prior to using copyrighted material or collecting data."

Details on the core HEAL CDEs (required for use by HEAL Initiative clinical pain research studies) and supplemental CDEs (to use when appropriate) can be found here. The nine core CDEs are pain intensity, pain interference, physical functioning/quality of life, sleep, pain catastrophizing, depression, anxiety, global satisfaction with treatment, and substance use screener.

A statement on the standard HEAL study-level metadata should also be included in this section to ensure that the appropriate HEAL metadata requirements are met.

All HEAL studies collecting human subjects data and planning to use CDEs (even studies outside the clinical pain research portfolio) are strongly encouraged to search for applicable CDEs within the HEAL database, and use questionnaires from this database if possible. Studies using CDEs, regardless of whether they are part of the HEAL repository, will be required to report which questionnaires are being used.

Additionally, consider your intended data repositories when outlining CDEs in your DMP.

Additional Guidance from DMPTool

A *standard* specifies how exactly data and related materials should be stored, organized, and described. In the context of research data, the term typically refers to the use of specific and well-defined formats, schemas, vocabularies, and ontologies in the description and organization of data. However, for researchers within a community where more formal standards have not been well established, it can also be interpreted more broadly to refer to the adoption of the same (or similar) data management-related activities or strategies by different researchers and across different projects.

It is possible that your work will employ multiple formal standards or a mix of formal standards and other data management strategies. You should be as specific as possible when describing the standards used for each type of data included in your proposal.

Data Preservation, Access, and Associated Timelines

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.*

Example Answer:

Qualitative and clinical outcome data will be archived in the National Addiction & HIV Data Archive

Program (NAHDAP) with the Inter-university Consortium for Political and Social research (ICPSR) at the University of Michigan. NAHDAP is one of NIH's designated repositories and permanently archives deposited files. NAHDAP offers unique persistent identifiers, long term sustainability, metadata, curation and quality assurance, free and easy access, broad and measured reuse, clear use guidance, security and integrity, confidentiality, common format, provenance, and a retention policy. Researchers interested in using the data can search for and request the data on the ICPSR and NAHDAP websites.

Guidance:

NIH HEAL Initiative Specific Considerations

HEAL-funded data must be deposited in a HEAL-compliant repository. This should be done no later than the time of an associated publication or the end of an award period, and ideally will be done as soon as possible.

The HEAL data platform does not host data, but there are established long-term data storage repositories that have been selected for use by the HEAL Data Ecosystem. The HEAL Data Stewards compiled the list of HEAL-compliant repositories by choosing repositories from the NLM Data Sharing Resources that satisfy the HEAL data repository selection principles.

HEAL investigators should have access to a repository managed by their administering NIH Institute or Center, and this repository is generally the preferred option.

Refer to the HEAL Repository Selection Guide for assistance in selecting the best repository for your data and a complete list of HEAL-compliant repositories.

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.*

Example Answer:

NAHDAP will assign the study a DOI, a unique persistent identifier to ensure the data are permanently identifiable. To enhance data findability, NAHDAP creates metadata records in DDI format. This includes study-level metadata and can also include variable-level metadata to further enhance findability.

Guidance:

NIH Guidance

Unique Persistent Identifiers: The repository assigns datasets a citable, unique persistent identifier, such as a digital object identifier (DOI) or accession number, to support data discovery, reporting, and research assessment. The identifier points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.

NIH HEAL Initiative Specific Considerations

Much of this information can be found in FAIRsharing or re3data entries linked in the far right column of the HEAL-Compliant Repository List. It will be specific to the individual repository chosen for the DMP.

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.*

Example Answer:

De-identified data will be shared with NAHDAP as soon as possible by the Principal Investigator, upon completion of quality control procedures, analysis, and at the time of associated publication or end of performance period. Data will be released by NAHDAP at the time of receipt and will be available indefinitely.

Guidance:

NIH Guidance

NIH encourages scientific data be shared as soon as possible, and no later than time of an associated publication or end of the performance period, whichever comes first. Researchers are encouraged to consider relevant requirements and expectations (e.g., data repository policies, award record retention requirements, journal policies) as guidance for the minimum time frame scientific data should be made available. NIH encourages researchers to make scientific data available for as long as they anticipate it being useful for the larger research community, institutions, and/or the broader public. Identify any differences in timelines for different subsets of scientific data to be shared.

Genomic data has further guidance on release expectations and timelines.

NIH HEAL Initiative Specific Considerations

HEAL aligns with NIH guidance on timelines for how long data will be made available.

Access, Distribution, or Reuse Considerations

Factors affecting subsequent access, distribution, or reuse 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.

Example Answer:

Secondary surveillance data, secondary EHR systems data (other than derived outcome data from clinical trial participants), and secondary prescribing data will not be made available through a repository due to restrictions imposed by agreements for data use by the data owners. However, dashboard visualizations of secondary surveillance data will be available through public access via Named County Website.

Guidance:

NIH HEAL Initiative Specific Considerations

The HEAL Initiative recognizes that much of its data will contain sensitive personal information and thus will need to be protected. Datasets containing this type of information is expected to be de-identified and stored in controlled-access repositories. Researchers should describe in detail their plans for addressing concerns about protecting sensitive information in order to safely meet the HEAL Initiative's data sharing goals. If any data should be withheld from repositories for privacy concerns, researchers should provide thorough justification for this in their DMP.

In an effort to make data as shareable as possible, the HEAL Initiative encourages researchers to include an explanation of their intentions for data sharing in their informed consent forms.

Additional DMPTool Guidance

Some data may require extra preparation before they can be shared. This is the section to describe what legal, ethical, or technical issues will require limiting the sharing of your data. Examples may include existing legal limits such as data licenses or use agreements, issues of proprietary IP development, technical limits about the size or structure of the data, or ethical issues for human subjects privacy.

Key issues in justification of human subjects data specifically may be informed consent (e.g., disease-specific limitations, particular communities' concerns) or privacy and confidentiality protections (i.e., de-identification, Certificates of Confidentiality, and other protective measures). Specific steps for human subjects data preparation can be addressed in the protections for privacy subquestion below.

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).*

Example Answer:

NAHDAP conducts disclosure risk review of all data, the results of which inform the appropriate restrictions to place upon access to the data. Descriptive information about care clinics may need to be masked to minimize disclosure risk. NAHDAP will make data available from this study via restricted-use files which requires a signed Restricted Use Agreement between the requesting institution and the University of Michigan, as well as an application to request access.

Guidance:

NIH HEAL Initiative Specific Considerations

To accommodate these data, select repositories on the HEAL-Compliant Repository List that are controlled-access. This information can be found in FAIRsharing or re3data entries linked in the far right column of the list. Type of access to data may vary by factors such as type of data and credentials of the person accessing the data within a repository.

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).*

Example Answer:

Data sharing protocols will be made explicit during the informed consent process to ensure participants understand how their qualitative and clinical trial outcome data will be managed and shared. Data will be de-identified prior to analysis, and a Certificate of Confidentiality serves as a protective measure.

Guidance:

NIH HEAL Initiative Specific Considerations

If data generated by HEAL projects is considered Protected Health Information (PHI), researchers should work with repository staff to determine how the PHI will be de-identified so that it can be disseminated in a safe manner. Details on the extent to which the PHI will be masked should be included in the DMP. Additionally, advanced planning with repository staff will be beneficial for obtaining informed consent from study participants.

Additional DMPTool Guidance

Certain kinds of data, especially human subjects data, require extra preparation before they can be shared to ensure participant privacy. In this section, you will describe your approach to preparing human subjects data for sharing and note any additional restrictions or policies that will impact access to your data. If you are working with human subjects you should also describe how you will address data management and sharing in your informed consent process. You will also need to describe your methods for ensuring privacy and confidentiality, including how you will de-identify your data. If you have decided that a controlled access repository (where researchers must apply to access data) is a better fit for your data than an open repository, you should describe the repository's access procedures. Finally, if there are any other laws, policies, or existing agreements that impact your ability to share your data, they should be described here.

Issues to consider:

- Any restrictions imposed by federal, Tribal, or state laws, regulations, or policies, or existing or anticipated agreements (e.g., with third-party funders, with partners, with Health Insurance Portability and Accountability Act (HIPAA) covered entities that provide Protected Health Information under a data use agreement, through licensing limitations attached to materials needed to conduct the research).
- Any other considerations that may limit the extent of data sharing.

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).

Example Answer:

Compliance with the DMS Plan will be monitored by the study Principal Investigator. The DMS Plan will be reviewed and updated annually by the Principal Investigator and Project Manager. The Principal Investigator will be responsible for submitting data to the NAHDAP repository.

Guidance:

NIH Guidance

This element refers to oversight by the funded institution, rather than by NIH. The DMS Policy does not create any expectations about who will be responsible for Plan oversight at the institution.

Additional DMPTool Guidance

Describe how and by whom compliance with this Plan will be managed. If oversight and roles will include the addition of study personnel for oversight of data management and sharing, describe reasonable, allowable personnel costs in the budget justification rather than the DMS Plan.