Sample DMS Plan – Survey/Interview Data Project

DATA MANAGEMENT AND SHARING PLAN

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. 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:

1. **Survey** This study will collect quantitative and qualitative survey data from 3,000 U.S. adults. The survey instrument will include 40-50 fixed-scale and open-ended text items including novel measures, scales identified in the PhenX toolkit, and two proprietary measures from [insert entity]. Raw and recoded data including open-ended text responses and resulting recoded variables will be produced. Open-ended text items will not request personally identifying or sensitive information but will be reviewed for privacy disclosure risks and redacted accordingly.

2. **Interviews** This study will also conduct digitally recorded semi-structured interviews with patients (n=40), healthcare providers (n=40) and industry professionals (n=20). Deidentified, raw transcripts will be generated and coded using methods described in Aim 2. Codebooks will be developed and used for coding transcripts, as described in Aim 2. Codebooks and coding summary files will be shared as detailed below.

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

Survey Data: Except where mentioned in Section 5 below, de-identified individual and aggregate survey data (including raw and recoded data) will be shared. The de-identification process will remove direct and indirect respondent identifiers. Once data are confirmed final, respondent identifiers will be deleted.

Interview Data: Following generation and quality check of raw transcripts from interviews, digital voice recordings will be permanently deleted to protect participant privacy. Respondent identifiers will not be shared. Raw transcripts will be maintained but not shared. Transcripts from interviews with patients and healthcare providers will be de-identified and sensitive content redacted where identification is plausible. These de-identified and redacted transcripts and coding summaries will be shared. Transcripts from industry stakeholder interviews will not be shared to protect participant privacy (see 5A). All shared data sets mentioned above, and metadata (see below) will be made publicly available through the Analysis, Visualization, and Informatics Lab-space (AnVIL).

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

Documentation to be made publicly available to the research community will include PDF documents containing:

- Survey instruments with proprietary measures redacted (Note we will not be at liberty to share proprietary instruments used in the survey but will provide citations and contact information for proper licensure)
- Interview guides
- All data collection protocols including sample and subject selection methods
- Copies of blank, dated, stamped consent forms and IRB approvals, and resulting limitations of data usage
- Survey codebook including question number, question text, variable name, variable label, value labels, codes for missing, non-applicable, "don't know," and refusal values
- Methods used to code open-text survey responses
- Codebook for analyses of interviews, including a list and definition of all codes used, and coding examples
- Steps taken to remove direct and indirect identifiers in the data
- Description of software and analytical methods used in survey and interview data analyses
- R code used in survey data analyses.
- A standard citation and unique identifier to facilitate attribution of data use.

These will be shared in AnVIL. To the extent the context of data collection can be revealed without compromising privacy and identity of research participants, it will be included in study protocols.

Element 2: Related Tools, Software and/or Code

Novel tools and software will not be generated. Proprietary data analysis software such as [insert name] may be needed to analyze transcripts data and must be licensed independently by data users. Other data analyses will be conducted using

the open-source R package. Copies of the R code used in our analyses will be made available in AnVIL.

Element 3: Standards

The study will use standard processing and documentation protocols adopted by the Inter-university Consortium for Political and Social Research (ICPSR) for data formats, dictionaries, variable names, descriptions, and labels. An XML schema using Data Documenting Initiative standards will also be used for codebooks and other metadata as appropriate.

Data Type	Standard/Format
Raw and recoded individual-level survey data	Survey data and coding schemes will be shared in character-delimited ASCII files widely read by most data analysis programs
Aggregate survey data	Character-delimited ASCII files
Voice Recordings	Will not be shared. WAV files will be deleted following generation of transcripts
Deidentified & coded interview transcripts	Text files readily imported by qualitative data analysis programs. Files will include line numbers and time codes.
Metadata	PDFs, DDI-standard XML files and customized (non-standard) files as appropriate

Element 4: Data Preservation, Access, and Associated Timelines

A. Repository where scientific data and metadata will be archived:

All shared study materials, data, and metadata will be made publicly available through the Analysis, Visualization, and Informatics Lab-space (AnVIL). Funds have been requested in this grant budget to support data curation and deposition. Details are found in the budget justification.

B. How scientific data will be findable and identifiable:

All shared data and metadata will be findable and identifiable using the standard data indexing tools in AnVIL which will assign unique persistent identifiers to data and metadata files that can be referenced. Links to the AnVIL workspace(s) where data can be found will be included in all publications, progress reports, and on the PI's academic home page.

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.*

Final data submission and release of data used in publications will occur approximately 8 and 12 months following the end of fieldwork, respectively. Datasets underlying publications will be shared at or prior to initial publication date. Data we do not publish on will be shared before the end of this award. Shared data will be preserved according to AnVIL's data retention policy. Currently, AnVIL has no process for deleting or retiring data sets.

Element 5: Access, Distribution, or Reuse Considerations

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

The subitem questions and data making up the two proprietary scales used in the survey cannot be shared due to licensing restrictions. However, the computed measures resulting from these scales will be shared.

Sharing qualitative data such as transcriptions from interviews can potentially reveal sensitive information and identify individual participants. The potential for identification from interview transcripts is particularly high among industry professionals given the small population from which they are recruited and the inclusion of questions regarding organization-specific practices. The higher level of redaction needed to remove potentially identifying and sensitive information would substantially reduce the utility of transcripts in future analyses. Moreover, our prior experience and the attached letters of support indicate that confidentiality would be required for industry stakeholders' full participation in this study. To ensure this group's participation and honest responses, we will not share transcripts or coding summaries from industry stakeholder interviews.

Research participants will receive information about where and how data from this study will be shared during study enrollment procedures and in informed consent documents. Interview participants will be informed their data will be shared on AnVIL with controlled access for future use (see 5B). Participants who want to withdraw their data from the study prior to de-identification may contact the study team or the university's research administration office (See 5C).

B. Whether access to scientific data will be controlled:

Survey data and metadata will be shared through AnVIL as open access. Interview data will be shared through AnVIL as controlled access to further protect research participants. The NHGRI Data Access Committee (DAC) will manage access to the controlled datasets. Users of the data in AnVIL must register with AnVIL and agree to the Terms of Use. Data users also agree not to share or redistribute any data downloads.

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

Data will be de-identified according to HIPAA and the Common Rule. All direct respondent identifiers (e.g., names, residence, email addresses) will be removed. Interview transcripts will be redacted to remove additional non-standard identifiers. Deidentification will be completed prior to the finalization of shared data files. Digitally recorded interviews will be deleted following transcription.

Participants will have the opportunity to opt out of sharing during informed consent procedures. Once personally identifying information are removed from the data set and deleted, we will not be able to identify data associated with a specific research participant for removal from the dataset. Limited provisions for participants' withdrawal of data from sharing protocols will be outlined in informed consent.

Upon receipt of an NIH Award, the data for this study will be protected by a Certificate of Confidentiality.

Element 6: Oversight of Data Management and Sharing:

The study PI will oversee execution of this Data Management and Sharing Plan. The PI will oversee data curation, redaction, and submission to the AnVIL. Compliance with the plan will be monitored by the PI routinely. The PI will conduct monthly meetings with key study personnel to ensure the timeliness of data entry and will review data to ensure quality of data entry. The PI will ensure data are submitted and shared according to this DMSP. Progress on data sharing will be reported in the Research Performance Progress Report.