Plan Overview

A Data Management Plan created using DMPTool

Title: Disparities in Substance Use and Mental Health for Women by Sexual Orientation: Identifying Modifiable Risk and Protective Factors during the COVID-19 Pandemic Using Multiple Sources of Health Data

Creator: Kristie Seelman - **ORCID:** [0000-0002-4064-2927](https://orcid.org/0000-0002-4064-2927)

Affiliation: Georgia State University (gsu.edu)

Principal Investigator: Kristie Seelman, Grace Eau

Data Manager: Grace Eau, Blake McGee

Project Administrator: Kristie Seelman

Funder: National Institutes of Health (nih.gov)

Funding opportunity number: RFA-PM-23-002

Template: NIH-Default DMSP

Start date: 12-01-2023

End date: 11-30-2025

Last modified: 02-14-2023

Copyright information:

Disparities in Substance Use and Mental Health for Women by Sexual Orientation: Identifying Modifiable Risk and Protective Factors during the COVID-19 Pandemic Using Multiple Sources of Health Data

This project involves secondary data analysis of health data from the All of Us research study; no additional data are generated. The data utilized include survey data and electronic health records. The All of Us data are stored in a common cloud (online) environment, accessible to approved researchers who undergo training about privacy rules and other topics. The present study uses the controlled tier data, which include the most sensitive information that may increase the risk for possible identification of individuals, including sexual orientation information. Researchers are not able to download copies of the data - therefore, the researchers on this team are unable to prepare the actual data for public sharing, but others can access the data through the All of Us Research Hub.

The researchers will manage, preserve, and share:

(a) A project-specific codebook of variable names, descriptions, and values used in our project from the All of Us dataset;

(b) A copy of R/Python syntax used for data analysis for each resulting publication.

Because researchers cannot download copies of the full All of Us dataset, all analysis occurs in the cloud environment, and access to the data is controlled by All of Us, our research team cannot preserve or share actual copies of the data.

We will preserve and publicly share a list of variables used for this study in a project-specific codebook so that others may potentially replicate or build from our analysis in future research. Each variable in the codebook will include a brief description of the item along with the question number and question text from the relevant survey (if applicable), variable name, variable label, value labels, and standard codes for missing values—including codes for non-applicable,
"don't know," and refusal. Documentation will be provided in portable document format (PDF).

We will preserve and share R and/or Python syntax used for our analyses so that others may potentially replicate or build from our analysis in future research. We will note which versions/packages of R/Python were used for analysis.

Study protocol for the All of Us research project and copies of surveys used with participants are already made publicly available through the All of Us Research Hub: https://www.researchallofus.org/.

Documentation to be made publicly available by our research team for the research community will include a project-specific codebook with univariate statistics for each variable used in our analysis, and study-level metadata following the Data Documentation Initiative specification. Each variable in the codebook will include a brief description of
the item along with the question number and question text from the relevant survey (if applicable), variable name, variable label, value labels, and standard codes for missing values—including codes for non-applicable,
"don't know," and refusal. Documentation will be provided in portable document format (PDF).

We will manage, preserve, and share R and Python syntax used for analyzing the All of Us data for this research. We will publicly share information about the version/packages of R and/or Python that we used for analysis so that procedures can be replicated by others.

The All of Us data are accessed using the cloud environment available to all approved researchers, including

Jupyter Notebooks. Additional tools or software are not needed to replicate our study.

We will use open file formats (e.g., PDF, TXT, HTML, etc.) to preserve and share our project-specific codebook and R/Python syntax.

Our project-specific codebook with variable names and descriptors and R/Python syntax will be deposited in the Open Science Framework registry.

Open Science Framework provides searchable study-level metadata for dataset discovery. Open Science Framework assigns DOIs as persistent identifiers, and has a robust preservation plan to ensure long-term access. Data will be discoverable online through standard web search of the study-level metadata as well as the persistent pointer from the DOI to the dataset.

All data generated from this project will be made available as soon as possible, and no later than the time of publication or the end of the funding period, whichever comes first. The duration of preservation and sharing of the data will be a minimum of 10 years after the funding period.

Because we are not sharing the underlying data, but only a project-specific codebook and R syntax, there are not limitations requiring approval before we publicly share this information.

The All of Us study has its own limitations on availability of data and privacy of participants. The Public Tier of data contains only aggregate data with identifying information removed and is open to the general public for use.

For our project, we are utilizing the Controlled Tier of data, which contains genomic data in the form of whole genome sequencing (WGS) and genotyping arrays, previously suppressed demographic data fields from EHRs and surveys, and unshifted dates of events. According to the All of Us protocol, these data are accessible only to approved, trained researchers. Further, researchers are not able to download the actual data for analysis. These limitations prevent our team from distributing the actual dataset used for our analyses.

Controlled access will not be used for access to our project-specific codebook or R/Python syntax. This information will be shared by unrestricted download.

All of Us data are available by controlled access only. Any individual may receive access to All of Us data resources (i.e., become an "authorized user") if they follow the appropriate process. An authorized data user is a person who is authorized to access and/or work with registered or controlled tier data from the All of Us Research Program. Authorized users receive a data passport that allows them to create workspaces and conduct research. Initially, a data user's institution must enter into an institutional data use agreement with the All of Us Research Program for an individual to become an authorized user. Once their institution has entered into the agreement, the individuals must take the following steps to become authorized data users: (1) Provide their identity to the All of Us Research Program; (2) Provide consent for public display of their name and affiliations along with plain language descriptions of their research project; (3) Provide consent for public release of name and affiliation if the RAB finds that they have violated the DUCC. (5) Complete the All of Us Responsible Conduct of Research Training, including modules on data security and participant privacy awareness, and renew this training on an annual basis. (5) Provide a signature that codifies that the user has read, understood, and agrees to abide by the DUCC, and has completed the requisite training.

N/A - The only data being managed and publicly shared by the research team are not human subjects data (i.e., the project does not involve interactions with human subjects, and the data being analyzed are not identifiable). This

project will adhere to the All of Us Research program's Data Protection and Privacy standards as well as their code of conduct for responsible data use. All members of the research team involved in direct analysis of study data have undergone the training and registration required by All of Us for access to the controlled tier of data.

PI Kristie Seelman will be responsible for the day-to-day oversight of project management activities and data sharing, while PI Grace Eau will track our final variables list and store copies of our R/Python syntax. Broader issues of the DMS Plan compliance oversight and reporting will be handled by the PIs and Co-Investigators as part of Georgia State University's stewardship, reporting, and compliance processes.