Best Practices: Sharing Human Subjects Data

Dryad Requirements

Researchers are responsible for ensuring that all contents of their data package do not contain information that can be used alone, or in aggregate, to identify any individual.

Dryad's policies on human subjects data are in accordance with accepted international standards for de-identifying data from such trusted sources as General Data Protection Regulation (GDPR), HIIPA privacy rules, the Act on the Protection of Personal Information (APPI), and the Personal Information Protection Law (PIPL). Dryad will uphold the policies and publication requirements set in keeping with our responsibility to protect human participants and maintain the integrity of the research data we publish – regardless of whether the data submitted is or will be openly available elsewhere.

These guidelines apply to human subjects data for both living and deceased participants.

Preparing Your Data

Dryad does not publish any direct identifiers, such as an individual's name, initials, email address, or specific dates related to the individual.

Datasets may contain **no more than three indirect identifiers**, such as demographic, biological, and geographic data, that could lead to re-identification if combined with other available data.

To properly de-identify your data, consider direct and indirect identifiers and evaluate whether the combination of identifiers could lead to re-identification. For example, the age of participants, uncommon characteristics of the individual (e.g., rare health condition, number of children), geographic/regional location, named facility and/or service provider, and highly visible characteristics of the individual (e.g., ethnicity, race).

A partial listing of common **direct and indirect identifiers** is provided in the table on **page 2**. Because Dryad hosts a wide-range of multidisciplinary research data, it is not possible to provide a comprehensive list of potentially concerning identifiers. We recommend referencing this table to guide your understanding of the variables that can be identifying and to assist you in recognizing and categorizing other direct or indirect identifiers in your data.

A detailed description of your process for de-identifying data should be included in your README file. Additionally, if your research funder requires a Data Management Plan (DMP), statements of protections for privacy, rights, and confidentiality of human research participants will be required. Click here to view an example of a publicly available DMP.

Best Practices: Sharing Human Subjects Data

De-Identifying Your Data

Reducing, eliminating, or modifying your data can be challenging in scenarios when identifiers included in a dataset are essential for full analysis and to facilitate reuse. Therefore, Dryad provides mechanisms for authors to assess potentially re-identifying data and techniques for de-identifying the data (**see pages 3-4**).

To minimize the risk of disclosure, work with your institutional review boards and/or directly with study participants. Whenever possible, obtain informed consent to release participant-level data at the time of data collection and preparation. If you are unable to preserve critical data points in order to meet Dryad guidelines for publication, consider an alternate repository that offers managed access for human subject data.

Direct Identifiers (none allowed)	Indirect Identifiers (3 maximum)
➢ Name or initials	➢ Year of birth or age
➢ Names of relatives	➢ Ethnicity, race, indigenous status
➢ Participant IDs that are assigned using a combination of characters or numbers linked to an individual (e.g. initials, birth year, etc.)	➢ Gender or sex
➢ Dates related to an individual (e.g., birthdate, hospitalization dates)	➢ Criminal record
➢ Unique identifying numbers (e.g., social security number, social media IDs)	➢ Place of birth, treatment, residence or geographic location
➢ Address, including full or partial postal code	➢ The inclusion of minors (under 18 years old) in the study
➢ Telephone or fax numbers, electronic mail address	➢ Socioeconomic data (e.g., occupation, job title, place of work, income, education)
➢ Vehicle identifiers (e.g., license plate, VIN number)	➢ Household or family composition
➢ Medical device identifiers (e.g., serial numbers, manufacturer info)	➢ One or more pregnancies, fertilization methods, pregnancy/birth outcomes
➢ Web or internet protocol (IP) address	➢ Sexual attitudes, practices, or orientation
➢ Biometric data (including BMI)	➢ Organizational membership or affiliation (e.g., religious, political, trade group)
➢ Facial photograph or comparable image such as fMRI	➢ Information regarding an individual's psychological well-being or mental health; including family

Best Practices: Sharing Human Subjects Data

data showing facial structures ➢ Fingerprints, retina scans ➢ Audiotapes or videos with participant voice	history (e.g., alcoholism, genetic conditions) ➢ Anthropometric measure (e.g., height, weight) ➢ Rare disease or treatment (defined as <200,000 in the United States, <4.85M globally, or the equivalent proportion of population in the country of study) ➢ Name of health professional or facility responsible for care ➢ Sensitive data and/or stigmatized condition (e.g., illicit drug use, HIV/AIDS, vaccination status) ➢ Small population size — less than 100 (not to be conflated with small *sample* size) ➢ Consumer habits, privileged ownership of or access to uncommon or scarce tangible items ➢ Verbatim responses or transcripts

De-Identification Techniques

Technique	Description	Examples
Aggregating	Place data in ranges	Use age-ranges instead of date of birth or exact age (e.g. 10-20, 90 or older); report group average vs individual values
Bracketing/Blurring	Reducing precision of data by combining one or more data elements	Modify or obscure visual information (e.g., blurring faces in a photograph/video)
Collapsing and/or combining variables	Merge the concepts embodied in two or more variables by creating a new summary variable	Change exact dates to time interval between events (e.g., duration of time spent on a survey)
Eliminate outliers	Restrict the upper or lower ranges of a continuous variable to avoid outliers	Sort your data sheet for each variable and look for unusually high or low values. Or, use histograms, scatterplots, or boxplots to identify outliers.
Limit number of indirect identifiers to three (3) maximum	Be cautious when using small subgroups or small areas; avoid submitting tables with small cell	🚫 Age + sex + job title + family composition ✅ Age (ranged) + sex + occupation

Best Practices: Sharing Human Subjects Data

Technique	Description	Examples
	sizes (i.e., cells with fewer than 5 respondents)	
Perturbation	Replace sensitive data with realistic but inauthentic data or modify original data based on predetermined masking rules (which may include randomization). This method reduces the likelihood of reverse identification.	Use an algorithm that replaces the subjects' date of birth
Reducing	Adjust precision of data	Remove day and month from year of birth; use total time "x months" instead of exact date ranges; use county instead of city; add or subtract a small, randomly chosen number
Removal/Suppression	Determine the data necessary for reproducibility and remove the rest	Eliminate unique identifying numbers, characteristics, or codes including: telephone/fax, email addresses, postal code, medical record numbers, IP addresses, URLs, certificate/license numbers, device identifiers and serial numbers, health plan beneficiary numbers, account numbers, vehicle identifiers and serial numbers
Sampling	Reduce the amount of data reported	Rather than providing all of the original data, release a random subsample of sufficient size to yield reasonable inferences
Summarizing	Present only the total ranges, for example, mean, median, mode, etc. for a given group	Summary tables that reflect but are not themselves raw data
Swapping/Shuffling	Data for one or more variables are switched with another record, so that the data user does not know whether the real data values correspond to certain records (i.e., all the values in the data set are real, but are assigned to the wrong people)	Replace one data element with either a random or made-up value, or with another value in the data set; can be done manually or by using an algorithm

Tools

Amnesia - Used to remove and transform identifiers in a dataset: https://www.openaire.eu/item/amnesia

pydeface and mri_defaceddress - Used to remove/obscure facial structures in fMRI data: https://pypi.org/project/pydeface/

Best Practices: Sharing Human Subjects Data

Resources

HIIPA privacy rule: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html

Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule:
https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard

Direct and Indirect Identifiers:
https://uwaterloo.ca/research/sites/ca.research/files/uploads/files/direct_and_indirect_identifiers_access_check_done_0.pdf

Personally Identifiable Information Guide: a list of PII examples:
https://matomo.org/personally-identifiable-information-guide-list-of-pii-examples/

Human Research Protection Program (HRPP): https://irb.ucsf.edu/definitions#indirectly_identifiable

IRB Table of De-Identification Techniques:
https://www.sjsu.edu/research/docs/irb-deidentification-techniques-table.pdf

References

Canadian Institutes of Health Research (2005, September) CIHR Best Practices for Protecting Privacy in Health Research. Ottawa, Ontario: Public Works and Government Services Canada. Retrieved January 23, 2017 from http://www.cihr-irsc.gc.ca/e/29072.html

Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG (2010) Preparing raw clinical data for publication: Guidance for journal editors, authors, and peer reviewers. BMJ 340: c181 http://dx.doi.org/10.1136/bmj.c181

National Institutes of Health (2003, March 5) NIH Data Sharing Policy and Implementation Guidance Retrieved March 13, 2018 from https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#hs

Olesen S, Australian National Data Service (2014) ANDS Guide to Publishing and sharing sensitive data. Australian National Data Service. Retrieved February 01, 2017 from http://ands.org.au/guides/sensitivedata.pdf

The Inter-University Consortium for Political and Social Research (2012) Guide to social science data preparation and archiving: Best practice throughout the data life cycle, 5th ed. Ann Arbor, MI: Institute for Social Research, University of Michigan. Retrieved January 30, 2017 from
http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

The UK Data Service (n.d.) Anonymisation. Retrieved February 1, 2017 from
https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation

United States Department of Health and Human Services (2007, February 2) HIPAA privacy rule: Information for researchers. Limited data set and data use agreement. Retrieved January 20, 2017 from
https://privacyruleandresearch.nih.gov/pr_08.asp#8d

Best Practices: Sharing Human Subjects Data

Van den Eynden V, Corti L, Woollard M, Bishop L, Horton L (2011) Managing and Sharing Data: Best Practice for Researchers, 3rd ed. Essex, UK: UK Data Archive. Retrieved February 01, 2017 from http://www.data-archive.ac.uk/media/2894/managingsharing.pdf