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), HIPPA 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.

Preparing Your Data

**Dryad does not publish any direct identifiers.** A direct identifier is information that is sufficient, on its own, to disclose the identity of a research participant. Examples include: name, address, postal code, telephone number, voice, video, or photograph.

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 (either collected as part of your research, or available elsewhere). Examples include: institutional affiliations, occupation, geographic region, unique values or characteristics (outliers).

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 pages 2-3. Because this is not a definitive list of potentially concerning identifiers, we recommend referencing the table to guide your understanding of the type of variables that can be identifying and help you recognize and categorize 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.
**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 page 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 controlled-access for human subject data.

<table>
<thead>
<tr>
<th>Direct Identifiers (none allowed)</th>
<th>Indirect Identifiers (3 maximum)</th>
</tr>
</thead>
<tbody>
<tr>
<td>➢ Name or initials</td>
<td>➢ Year of birth or exact age</td>
</tr>
<tr>
<td>➢ Names of relatives</td>
<td>➢ Ethnicity, race, indigenous status</td>
</tr>
<tr>
<td>➢ Participant IDs that are assigned using a combination of characters or numbers linked to an individual (e.g. initials, birth year, etc.)</td>
<td>➢ Gender or sex</td>
</tr>
<tr>
<td>➢ Exact dates related to an individual (e.g., birth, hospitalization dates)</td>
<td>➢ Criminal record</td>
</tr>
<tr>
<td>➢ Unique identifying numbers (e.g., social security number, individual digital identifier, medical record number)</td>
<td>➢ Place of birth, treatment, residence or geographic location</td>
</tr>
<tr>
<td>➢ Address, including full or partial postal code</td>
<td>➢ The inclusion of minors (under 18 years old) in the study</td>
</tr>
<tr>
<td>➢ Telephone or fax numbers, email address</td>
<td>➢ Socioeconomic data (e.g., occupation, job title, place of work, income, education)</td>
</tr>
<tr>
<td>➢ Vehicle identifiers (e.g., license plate, VIN number)</td>
<td>➢ Household or family composition</td>
</tr>
<tr>
<td>➢ Medical device identifiers (e.g., serial numbers, manufacturer info)</td>
<td>➢ One or more pregnancies, fertilization methods, pregnancy/birth outcomes</td>
</tr>
<tr>
<td>➢ Web or internet protocol (IP) address</td>
<td>➢ Sexual attitudes, practices, or orientation</td>
</tr>
<tr>
<td>➢ Fingerprints, retina scans</td>
<td>➢ Organizational membership or affiliation (e.g., religious, political, trade group)</td>
</tr>
<tr>
<td></td>
<td>➢ Anthropometric measures (e.g., height, weight)</td>
</tr>
</tbody>
</table>
### Best Practices: Sharing Human Subjects Data

**Direct Identifiers (none allowed)**
- Facial photograph or comparable image such as fMRI data showing facial structures
- Audiotapes or videos with participant voice

**Indirect Identifiers (3 maximum)**
- Information regarding an individual’s psychological well-being or mental health; including family history (e.g., alcoholism, genetic conditions)
- 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
- Biometric data (including BMI)

### De-Identification Techniques

There are many de-identification techniques that can be used to modify the variables in your data package. The table on the next page lists some of the more common techniques used to de-identify data. Additional methods and guidance for preparing human subject data for publication are available on page 5.
<table>
<thead>
<tr>
<th>Technique</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate</td>
<td>Place data in ranges</td>
<td>Use age-ranges instead of date of birth or exact age (e.g. 10-20, 90 or older); report group average vs individual values</td>
</tr>
<tr>
<td>Collapse and/or combine variables</td>
<td>Merge the concepts embodied in two or more variables by creating a new summary variable</td>
<td>Change exact dates to time interval between events (e.g., duration of time spent on a survey)</td>
</tr>
<tr>
<td>Eliminate outliers</td>
<td>Restrict the upper or lower ranges of a continuous variable to avoid outliers</td>
<td>Group income or age into broader categories. A 72-year old → grouped in “people in their 70s” or “senior citizens”</td>
</tr>
<tr>
<td>Generalize</td>
<td>Adjust precision of data</td>
<td>Specific professional position → occupation or area of expertise Date of birth → year or decade Town → region</td>
</tr>
<tr>
<td>Limit number of indirect identifiers to three (3) maximum</td>
<td>Be cautious when using small subgroups or small areas; avoid submitting tables with small cell sizes (i.e., cells with fewer than 5 respondents)</td>
<td>☒ Age + sex + job title + # kids ☒ Age (aggregated) + sex + occupation</td>
</tr>
<tr>
<td>Remove/Suppress</td>
<td>Determine the data necessary for reproducibility and remove the rest</td>
<td>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</td>
</tr>
</tbody>
</table>
Best Practices: Sharing Human Subjects Data

Tools

**Amnesia**: Used to remove and transform identifiers in a dataset: [https://www.openaire.eu/item/amnesia](https://www.openaire.eu/item/amnesia)

**pydeface and mri_defaceddress**: Used to remove/obscure facial structures in fMRI data: [https://pypi.org/project/pydeface/](https://pypi.org/project/pydeface/)

**ARX De-Identifier**: [https://arx.deidentifier.org/](https://arx.deidentifier.org/)

**Tool for creating unique identifiers for study participants**: [https://fitbir.nih.gov/content/global-unique-identifier](https://fitbir.nih.gov/content/global-unique-identifier)


**Applications to De-Identify Human Subject Research (Johns Hopkins University)**: [https://dataservices.library.jhu.edu/resources/applications-to-assist-in-de-identification-of-human-subjects-research-data/#image](https://dataservices.library.jhu.edu/resources/applications-to-assist-in-de-identification-of-human-subjects-research-data/#image)

Resources

**HIPAA privacy rule**: [https://www.hhs.gov/hipaa/for-professionals/privacy/index.html](https://www.hhs.gov/hipaa/for-professionals/privacy/index.html)

**HIPAA Safe Harbor Identifier List**: [https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#safeharborguidance](https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#safeharborguidance)

**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](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](https://uwaterloo.ca/research/sites/ca.research/files/uploads/files/direct_and_indirect_identifiers_access_check_done_0.pdf)

**Qualitative Data Repository (Syracuse University)**: [https://qdr.syr.edu/guidance/human-participants/deidentification](https://qdr.syr.edu/guidance/human-participants/deidentification)

**Personally Identifiable Information Guide**: a list of PII examples: [https://matomo.org/personally-identifiable-information-guide-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](https://irb.ucsf.edu/definitions#indirectly_identifiable)

**IRB Table of De-Identification Techniques**: [https://www.sisu.edu/research/docs/irb-deidentification-techniques-table.pdf](https://www.sisu.edu/research/docs/irb-deidentification-techniques-table.pdf)
Best Practices: Sharing Human Subjects Data

References


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