NIH Policy for Data Management and Sharing

Consultations: Data Management in Grant Proposals

- Takes effect January 25, 2023
- All data-producing research funded by NIH will require a 2-page Data Management and Sharing Plan
- How data will be managed, which data will be shared
- Plans should address privacy of human participants, planning for data sharing in informed consent

NIH Scientific Data Sharing website: https://sharing.nih.gov/
JHU Data Services support for NIH DMS Plans

- A free, online tool that provides funder-specific templates for NIH Data Management & Sharing Plans (DMPs) (and for other funders)
- Leads researchers through the NIH DMSP writing process
  - Expert guidance within the tool
- Button to Request feedback from JHU Data Services
  - We also provide direct consultations

Ways to Share Data

- **Data repositories** manage online access, preservation, citation support
  - **Open Access** repositories
    - publicly published data
    - require full de-identification
  - **Restricted Access** repositories
    - manage researcher approval, Data Use Agreements
    - require removal of some PII/PHI

JHU Data Services can help you find a data repository for sharing data
Find a Data Repository: Discipline-specific, Restricted Access

  - Tend to be for specialized fields and NIH divisions

- A data repository for research in the social, behavioral sciences and public health data
  - With the options of Open or Restricted data repositories
    [https://www.icpsr.umich.edu/icpsrweb/ICPSR/](https://www.icpsr.umich.edu/icpsrweb/ICPSR/)

- Data repository for clinical trial data
  - Fee for deposits and de-identification services
    [https://vivli.org](https://vivli.org)

JHU Data Archive: general open access, de-identified data

- Accepts only fully de-identified data
- Any discipline and data type
- Open access (no user registration.)
- JHU Data Services screens de-identified data for remaining risk

JHU Data Archive information and FAQ [http://dataservices.library.jhu.edu/archiving](http://dataservices.library.jhu.edu/archiving)
How to assess the feasibility of de-identifying and sharing data

Planning for sharing de-identified data

• When preparing the NIH DMS Plan, do you know if de-identifying patient / human participant data is **feasible**?
• What do I need to know to about my data to decide?
• Can de-identified data be shared with **public access** or only **restricted access**?
• What **skills** will my research team need to de-identify data?
Is it feasible to de-identify my data for public sharing?

- NIH does not require de-identification for public sharing
- They encourage sharing what is feasible (restricted or open)
- NIH DMS Plan asks for justification for what cannot be shared openly.
  - Why data cannot be fully identified
  - When and how restricted access is possible

Justification for sharing restricted access only:

“Due to the geographic specificity of participants and potential links to public records, anonymization for public release is not feasible. Selected de-identified datasets will be shared with restricted access at ICPSR”

Personal and Health Identifiers

- Collected information expected to remain private, that can link to individual participants

- **PII:** Personally identifying information
- **PHI:** Protected Health Information
How do I know if de-identification is feasible?

**Basic de-identification:**

- **Remove or replace Direct Identifiers:** *uniquely private information*
  - HIPAA's 18 Direct Identifiers: ID numbers, SSN, MRN, etc.
  - Names
  - Address (within a U.S. state)
  - Phone, email, I.P. addresses

<table>
<thead>
<tr>
<th>HIPAA's 18 Direct Identifiers</th>
<th>ID numbers: SSN, MRN, etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Names</td>
<td>Full dates (D.O.B., Visit date)</td>
</tr>
<tr>
<td>Address (within a U.S. state)</td>
<td>Photos/biometric IDs</td>
</tr>
<tr>
<td>Phone, email, I.P. addresses</td>
<td>Treatment provider locations</td>
</tr>
</tbody>
</table>

**More thorough de-identification:**

- Remove or anonymize Indirect or **Quasi-identifiers:**
  - Demographics, socioeconomic variables, health measures...
  - Could link some records to externally available information

---

**Joe Patient**

<table>
<thead>
<tr>
<th>SubjID</th>
<th>Gend</th>
<th>HIVs</th>
</tr>
</thead>
<tbody>
<tr>
<td>2253</td>
<td>F</td>
<td>Pos</td>
</tr>
<tr>
<td>2254</td>
<td>M</td>
<td>Pos</td>
</tr>
<tr>
<td>2255</td>
<td>F</td>
<td>Neg</td>
</tr>
</tbody>
</table>
De-identifying direct identifiers is feasible, anonymizing quasi-identifiers can be challenging

**Direct Identifiers**: relatively simple to mask, most are not needed for analysis

**Quasi-identifiers**: more challenging to assess their risk and **anonymize** to decrease the probability of re-identification

<table>
<thead>
<tr>
<th>Facility name:</th>
<th>East Baltimore Clinic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age:</td>
<td>92</td>
</tr>
<tr>
<td>Date of onset</td>
<td>2/13/2020</td>
</tr>
</tbody>
</table>

Which variables are risky?
Reducing utility of data?

---

Quasi-identifiers can pose risk of linking to publicly available data

Example: database link re-identifies facility

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Government Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinic Name</td>
<td>Rosewood Rehab Center</td>
</tr>
<tr>
<td>Clinic Revenue</td>
<td>$824,209</td>
</tr>
<tr>
<td>Location</td>
<td>Baltimore, Maryland</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Female</th>
<th>Pregnant</th>
<th>Veteran</th>
<th>Baltimore, MD</th>
</tr>
</thead>
</table>

Unique “outlier” cases have higher probability of matching
De-identifying direct identifiers is feasible, anonymizing quasi-identifiers can be challenging.

De-identification techniques vary in difficulty.

<table>
<thead>
<tr>
<th>Date Shift</th>
<th>Age Range</th>
<th>Truncate zip code</th>
</tr>
</thead>
<tbody>
<tr>
<td>+/- 180 days</td>
<td>20-25</td>
<td>21219</td>
</tr>
</tbody>
</table>

Some de-identification techniques are relatively simple.

Advanced anonymization methods: unfamiliar & time consuming.

K-anonymity risk probability calculations

\[
\frac{1}{n} \sum_{i \in \mathcal{D}} f_i \times \left( \frac{1}{\min(f_i)} \right) \leq \frac{1}{\min(f_i)}
\]

Use HIPAA’s 3 levels of disclosure protection to decide between restricted vs. open access.

- **Limited Data Set**: Direct identifiers removed or masked, Zip code-level location, Full dates can be retained.
  - Direct collaboration, secure data enclaves

- **“Safe Harbor”**: Direct & Quasi-identifiers that “knowingly” re-identify, 3-digit Zip code truncation, Year only dates, ages above 90.
  - Restricted access repositories

- **Expert/Statistical Determination = “De-identified”**: De-identification performed with appropriate knowledge, accepted statistical techniques, Data assessed for remaining disclosure risk, Documented methods and results of analysis.
  - Public/open access repository
**Rule-of-thumb approach to restricted vs. public data release**

Assess a prospective codebook/data dictionary to anticipate quasi-identifiers

<table>
<thead>
<tr>
<th>Variable</th>
<th>Codes</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2cluster</td>
<td>ID</td>
<td>Cluster ID</td>
</tr>
<tr>
<td>P2district</td>
<td>Baltimore, DC, Fredrick, Towson</td>
<td>VDC code</td>
</tr>
<tr>
<td>P2surveydate</td>
<td>Survey date</td>
<td>Survey date</td>
</tr>
<tr>
<td>P2womana</td>
<td>age in years</td>
<td>Woman's age</td>
</tr>
<tr>
<td>P2womansch</td>
<td>Years of schooling</td>
<td>Completed years schooling</td>
</tr>
<tr>
<td>P2childbenefits</td>
<td>Benefit category</td>
<td>Child benefits</td>
</tr>
</tbody>
</table>

Could more than 10 participants match some combination of these variables given relevant public information?

What statistical anonymization techniques might be required? Can we reasonably budget for time and expertise?

Would benefits of open access outweigh difficulties/costs of restricted access?

If **Restricted Access** is the feasible choice, consider data selections that could be public

**Proposals could state:**

“We will consider feasibility of public data release.”

---

**A de-identification plan may be needed for reviews by IRB and JHM Data Trust (after grants are awarded)**

For sharing data from JH Medical records & systems:

**JHM Data Trust Research Subcouncil** must approve sharing data with external collaborators, restricted and public access repositories

- CCDA reviews de-identification protocols for datasets
- Your IRB form should include:
  - Plans to use a data repository
  - State the planned de-identification level

SOM IRB will request the Data Trust review

---

Welch Grand Rounds talk
Learning de-identification and risk assessment techniques

- Study teams can learn basic “Safe Harbor” techniques for restricted access. (PI’s should budget for learning)
- Large-scale public access datasets: consider a statistician (who may still need to learn anonymization techniques)

JHU Data Services provides:

<table>
<thead>
<tr>
<th>Training</th>
<th>Protecting &amp; removing human participant identifiers</th>
<th>Webinars &amp; Self-paced online training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consulting</td>
<td>1-1 meeting to provide guidance* for PHI/PII protection &amp; anonymization</td>
<td></td>
</tr>
<tr>
<td>Archiving</td>
<td>Screening for identifiers in JHU Data Archive deposits</td>
<td>* NOT de-identifying data for researchers</td>
</tr>
</tbody>
</table>
How can Data Services help researchers?

- Write your DMP using DMPTool and send to us for feedback (dataservices@jhu.edu)
- Contact us for consultations on DMSP preparation and help selecting a data sharing repository
- Writing Your NIH DMSP webinar, this January (recording online soon)
- Self-paced online training (link)

Guides for data management and sharing topics

- Protecting Identifiers in Human Subjects
- Data Management & Sharing
- Documenting Research Data

Contact Data Services via dataservices@jhu.edu