Deidentifying Data: A Primer on Disclosure Risk

John E Marcotte, PhD

ICPSR

University of Michigan

16 February 2022
International LOVE DATA WEEK
A Primer on Disclosure Risk

A. Disclosure and Risk
B. Evaluating risk
C. Remediation Options
A. Disclosure and Risk

• Unauthorized release of information about an individual or organization

• Information that pertains to a specific individual
Disclosure

• Identification of specific individuals or organizations in a study

• *Disclosive*

  Disclosive data may lead to the identification of a specific individual or organization.
Disclosure Risk

• More studies have detailed individual information and histories

• Studies of special populations

• Rich research possibilities

• Increased disclosure risk
Disclosure vs. Risk

• Protect against disclosure by reducing risk of disclosure

• While disclosure is rare with research data, risk of disclosure is increasing as studies include more details
Responsibility

• Data providers, data disseminators, data stewards and researchers have a responsibility to protect the identity of respondents

• Disclosure may violate laws

• Disclosure hurts all research
(Re-)Identification

• Direct identifiers

• Indirect or inferential identification

• Personal Identifiable Information (PII)

• Protected Health Information (PHI)
PII and PHI

- Name
- Address (all geographic subdivisions smaller than state, including street address, city county, and zip code)
- All elements (except years) of dates related to an individual (including birthdate, admission date, discharge date, date of death, and exact age if over 89)
- Telephone numbers
- Fax number
- Email address
- Social Security Number
PII and PHI

• Medical record number
• Health plan beneficiary number
• Account number
• Certificate or license number
• Vehicle identifiers and serial numbers, including license plate numbers
• Device identifiers and serial numbers
• Web URL
• Internet Protocol (IP) Address
• Finger or voice print
• Photographic image - Photographic images are not limited to images of the face.
Individually Identifiable Health Information (IIHI)

- Information that is a subset of health information, including demographic information collected from an individual
- Is created or received by a healthcare provider, health plan, employer, or healthcare clearinghouse
- Relates to the past, present, or future physical or mental health or condition of an individual
- Reasonable basis to believe the information can be used to identify the individual.
Indirect Identifiers

• Form a profile that allows identification of an individual

• Combination of variables

• Combinations may become PII
United States Laws

- **CIPSEA** Confidential Information Protection and Statistical Efficiency Act
- **HIPAA** Health Insurance Portability and Accountability Act
- **FISMA** Federal Information Security Management Act of 2002 Non-US
- **FERPA** Family Educational Rights and Privacy Act
- **Privacy Act** Requires the government and its agents to protect personal information it collects and maintains on private citizens
- **Workforce Investment Act** Prohibits the disclosure of data collected for statistical purposes
- **Trade Secrets Act** Prohibits disclosure of confidential business information collected and maintained by the government
Cross-national Issues

- International Laws
  
  *Europe has its own privacy laws*
  
  EU General Data Protection Regulation *(GDPR)*

- Laws may not be applicable across international boundaries

- Respect terms of data collection
Consequences

• Grants revoked
• Fines
• Jail
• Notify respondents
Unintended Disclosure

• Lack of intention to disclosure is not an excuse.

• Accidental disclosure still has ramifications.
Data Nomenclature

- Public-use or Public Access
- Controlled access
- Restricted-use
- Sensitive
- Confidential
- Limited
- Proprietary
Public-use Data

• All direct identifiers have been removed.

• Risk of inferential identification is practically non-existent.

• Terms of use

• Also called *Public Access*
Controlled Access

• Data that require an application or permission to access

• Data that are not readily available for download from a website

• Restricted-use is a subset of controlled access
Restricted-use Data

• All direct identifiers have been removed.

• Inferential identification is possible.

• Data may contain sensitive information.

• Data Use Agreements
Sensitive Data

Information that can cause harm or legal jeopardy; damage reputation

Some examples are:

- Health information
- Drug use
- Criminal record
- School record
- Information about minors
Confidential Data

Information that has been promised to keep secret
Limited Data

• PHI and PII have been removed or masked.

• May still have risk of inferential disclosure

• HIPAA designation
Proprietary Data

- Information that is owned.

- Data for which permission to distribute has not been given.

- May not be sensitive nor confidential
B. Evaluating Risk

• Check for PII, PHI or direct identifiers

• Check for sensitive information

• Are data confidential or proprietary?

• Check for inferential risk
Inferential Risk

- Low-levels of geography (for some data even State is too low)
- Special populations
- Histories
- Extreme or outlier values
- Highly detailed variable coding
- Unique profiles or typologies
Profiles

• **Unique profile**: Set of variables when combined together form a profile which can be used to link data from different sources

• Profiles may be for an individual, a family, a geographic area or an organization

• Unique profiles increase the risk of re-identification
Links and Lookups

• **Links:** Other sources of information that can be linked to data. Links increase the chances of re-identification and may enable the formation of a profile for lookup.

• **Lookups:** Information that translates profiles into identities.
Potential Data Linkages

• Other studies

• Administrative data

• Social media; people self identify as being part of a study

• “Big Data”

• *Potential linkages are growing*
Re-identification and Harm

- Chances of re-identification
- Possible harm *if* re-identified
- Both aspects must be considered
Re-identification Risk
<table>
<thead>
<tr>
<th>Harm Level</th>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0</td>
<td>No Harm</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Little Harm</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Humiliation</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Reputation Damage</td>
</tr>
<tr>
<td>Moderate</td>
<td>4</td>
<td>Emotional Distress</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Financial Loss</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Legal Jeopardy</td>
</tr>
<tr>
<td>High</td>
<td>7</td>
<td>Temporary Harm, Health Threat</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Permanent Harm, Impairment</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Severe Permanent Harm, Disfigurement</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Death</td>
</tr>
</tbody>
</table>
Sensitive Data

• Mitigating disclosure risk for sensitive data is particularly important.

• The disclosure risk threshold for data with sensitive information is lower (more risk averse).

• All information about minors is automatically sensitive.
Hierarchical Data

• Disclosure of higher levels in a hierarchy may lead to disclosure at lower levels.

• Identifying school and class will make the identification of students extremely probable.

• Sometimes organizations such as schools and health facilities need to be protected from disclosure too.
C. Remediation Options

• Remove or obscure identifying variables

• Remove or obscure sensitive variables

• Make data restricted-use
Data Modifications

- Suppress variables
- Replace variables
- Collapse categories, coarsen coding, top and bottom limits
- Perturb variables by adding random noise
- Swap records
- Aggregate to higher unit of observation (only release tables)
Data Modifications

• Suppressing or changing data can reduce the analytic value of data

• Some data cannot be modified sufficiently to mitigate disclosure risk

• Making data restricted-use decreases analysis based on the data
Suppress Variable

• Variable removed from data release and codebook

• Retain a restricted-use version with the variable

• Analysis must still be possible without the variable.
Suppress Variables

Some variables can be removed with no reduction in analytic value

- Personal identifiers are usually removed; however, suppressing identifiers will make linking harder

- Clusters are needed to compute standard errors
Replace Values

- New values are substituted for current values
- New values can be random but unique
- Prevents external linking of data
- Prevents direct re-identification
Coarsen Coding

• Recode so all categories have sufficient number of cases

• Recoded categories should have analytic validity

• Retain a restricted-use version with the original variable
Perturb Variables

• Maintain moments (Mean and Variance)

• Maintain order statistics (Median)

• Maintain one covariance if possible

• Retain a restricted-use version with the original variable
Swap Records

• Match records on variables that must be maintained.

• Univariate statistics should be very close before and after match.

• Multivariate statistics will vary more, but patterns of relationships should remain intact.

• Retain a restricted-use version of the original data
Swap Records

• Swapping records between geographic areas is most common

• Swapping is most often used when only public-use data can be made available

• Swapping is used in data that are used to report incidence and prevalence

• Deniability if individual claims record
# Public-use v. Restricted-use Data

<table>
<thead>
<tr>
<th></th>
<th>Public-use</th>
<th>Restricted-use</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purpose</strong></td>
<td>• Research Only</td>
<td>No attempt to identify respondents</td>
</tr>
<tr>
<td><strong>Request Data</strong></td>
<td>No application</td>
<td>Application</td>
</tr>
<tr>
<td><strong>Understanding</strong></td>
<td>Terms of Use</td>
<td>Data Use Agreement</td>
</tr>
<tr>
<td><strong>IRB</strong></td>
<td>Exempt</td>
<td>Possible Review</td>
</tr>
<tr>
<td><strong>Disclosure Risk</strong></td>
<td>Data: Very Low</td>
<td>Results: Very Low</td>
</tr>
<tr>
<td><strong>Security</strong></td>
<td>No security requirements</td>
<td>Security Plan</td>
</tr>
<tr>
<td><strong>Access</strong></td>
<td>Download from website</td>
<td>• Encrypted Download</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Online enclave</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Guarded cold room</td>
</tr>
</tbody>
</table>
Questions

Q & A
Events

- Randall Library and DEeL Virtual Coffee Hour – Introduction to Data Literacy (Mon 2/14 at 11:00)
- Citizen Science: Engaging Community Participation in Data Collection (Mon 2/14 at 4:00)
- Research Computing Updates at UNCW (Tues 2/15 at 3:00)
- Deidentifying Data: A Primer on Disclosure Risk (Wed 2/16 at 1:00)
- Who “Knows” Me?: Data Privacy and Online Behavior (Thurs 2/17 at 4:00)
- Tips for Writing an Effective Data Management Plan (Fri 2/18 at 1:00)

Register to attend at https://library.uncw.edu/love_data_week