ICPSR

SHARING DATA TO ADVANCE SCIENCE

Deidentifying Data:
A Primer on Disclosure Risk

John E Marcotte, PhD
ICPSR
University of Michigan

16 February 2022

ICPSR INSTITUTE FOR SOCIAL RESEARCH UNIVERSITY OF MICHIGAN

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.

IIHI

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

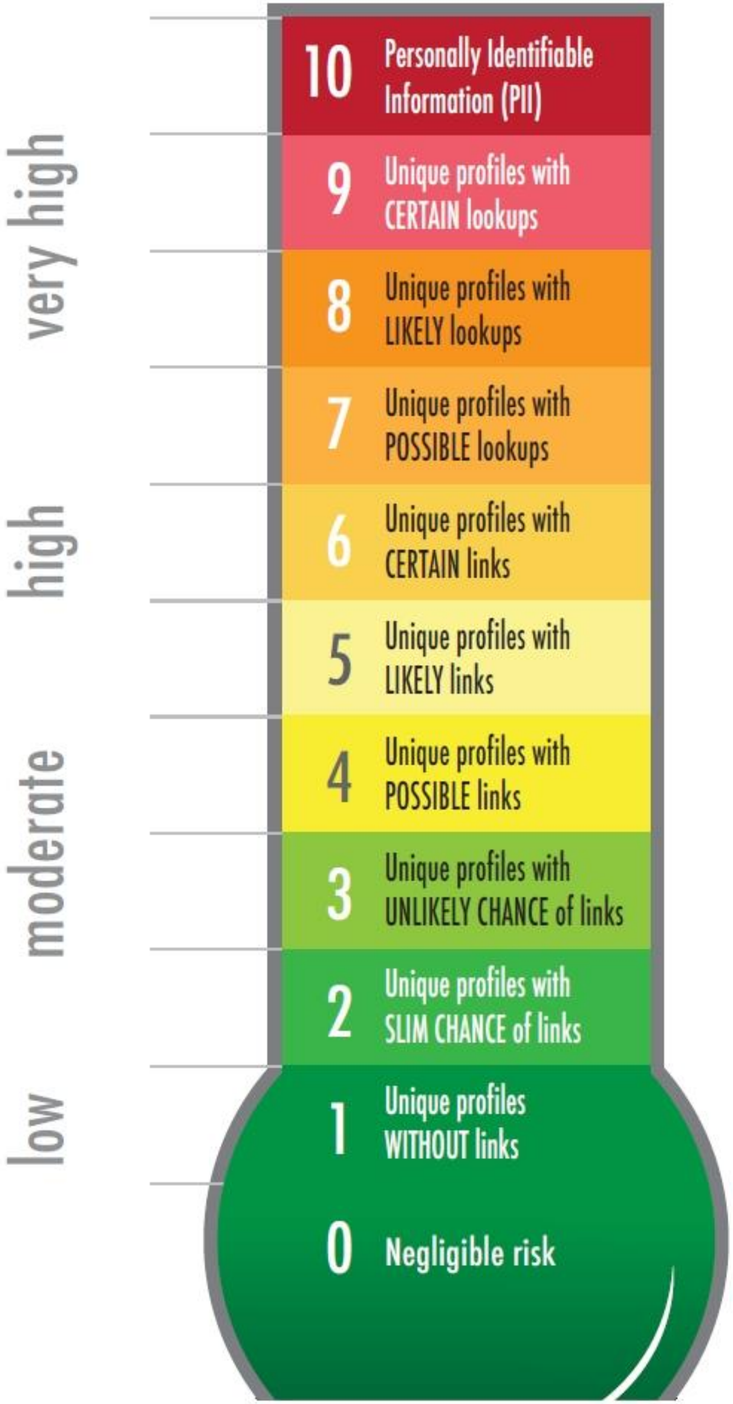
Potential Data Linkages

- 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



very high

high

moderate

low

10 Personally Identifiable Information (PII)

9 Unique profiles with CERTAIN lookups

8 Unique profiles with LIKELY lookups

7 Unique profiles with POSSIBLE lookups

6 Unique profiles with CERTAIN links

5 Unique profiles with LIKELY links

4 Unique profiles with POSSIBLE links

3 Unique profiles with UNLIKELY CHANCE of links

2 Unique profiles with SLIM CHANCE of links

1 Unique profiles WITHOUT links

0 Negligible risk

Harm

Low	0	No Harm
	1	Little Harm
	2	Humiliation
	3	Reputation Damage
Moderate	4	Emotional Distress
	5	Financial Loss
	6	Legal Jeopardy
High	7	Temporary Harm, Health Threat
	8	Permanent Harm, Impairment
	9	Severe Permanent Harm, Disfigurement
	10	Death

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

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

Questions