How to Identify and Remediate Disclosure Risk

John E Marcotte, PhD

July 11-13, 2018

Simon Fraser University (Burnaby, BC Canada)

Disclosure risk is the possibility of respondents or subjects being identified in data and is of concern to all people involved in collecting, analyzing, and distributing research data. As data for research includes more detailed information, disclosure risk increases.

First, this course will show the importance of public-use data. Public-use data has very low disclosure risk and is often readily available for download. A public-use version of the data provides the widest access for secondary analysis.

Second, the course will demonstrate how respondent confidentiality can be protected in research data. This segment will show how to assess and mitigate disclosure risk. This section will examine elements of a disclosure analysis as will disclosure protection such as statistical disclosure control. This segment will demonstrate measures commonly used to create public-use data files. Examples of public-use files created from restricted-use data, steps that can be taken early in the research process to optimize distribution options, and methods of distributing restricted-use data when public-use files cannot be created will also be covered. Examples of disclosure work from ICPSR will be used to illustrate disclosure risk and protection methods.

A third segment will show how to provide access to non-public-use data. These types of data require security protections and procedures. Although these data are not public-use, summary results must be public-use if published. This segment will discuss how to review summary results such as crosstabs and regression coefficients for disclosure risk.

How to Identify and Remediate Disclosure Risk

I. Session 1: Disclosure and Disclosure Risk
   a. Disclosure Risk
   b. Risk: Harm scale
   c. Making data public-use
   d. International Issues
   e. New types of data

II. Session 2: Assessing Disclosure Risk
   a. Classifying variables
   b. How ICPSR reviews data deposits

III. Session 3: Restricted-use Data
   a. Security Options
   b. Vetting Output For Disclosure

ICPSR

SHARING DATA TO ADVANCE SCIENCE

How to Identify and Remediate Disclosure Risk

John E Marcotte, PhD
ICPSR
University of Michigan
July 2018

ICPSR INSTITUTE FOR SOCIAL RESEARCH UNIVERSITY OF MICHIGAN

John E Marcotte, PhD

- Archivist at ICPSR
- Adjunct Faculty at UMSN
- Data Security Officer
- Primary Vetter of Output

Sessions

I. **Session 1: Disclosure and Disclosure Risk**

II. Session 2: Assessing Disclosure Risk

III. Session 3: Restricted-use Data

What do you want to learn?

- Please ask questions
- Please share your experiences
- I would prefer not to use jargon; however, some terms are unavoidable. Please ask if a term is not clear

ICPSR

SHARING DATA TO ADVANCE SCIENCE

Disclosure and Disclosure Risk

Disclosure and Risk

I. What is disclosure?

II. Types of Data

III. What increases risk

IV. Quantification of Risk

I. What is Disclosure?

- Unauthorized release of information about an individual or organization

- Information that pertains to a specific individual or organization

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 as in ADDEP
- 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

Disclosure Risk

- **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

Direct Identifiers

- Personally Identifiable Information (PII)

- Information uniquely associated with one individual or organization or geographic area

Personally Identifiable Information (PII)

Some examples of PII:

- Full name
- Address and telephone number
- Social Security number (SSN)
- Social Insurance Number (SIN)
- Drivers license number
- Face, fingerprints, or handwriting
- Credit card numbers
- Date of birth and Birthplace

Indirect Identifiers

- Form a profile that allows identification of an individual
- Combination of variables
- Combinations may become PII

Canada Laws

- **Privacy Act** covers the personal information-handling practices of federal government departments and agencies.
- **PIPEDA** (Personal Information Protection and Electronic Documents Act) covers the personal information-handling practices of many businesses.
- **Privacy Commissioner of Canada**
  https://www.priv.gc.ca/en

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
- **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

- **Laws may not be applicable across international boundaries**

- **Respect terms of data collection**

Consequences

- Fines
- Jail
- Notify respondents
- Pay for protections from potential harm

Unintended Disclosure

- Lack of intention to disclosure is not an excuse.

- Accidental disclosure still has ramifications.

II. Types of Data

- Public-use
- Restricted-use
- Sensitive
- Confidential
- Proprietary

Public-use Data

- All direct identifiers have been removed.
- Risk of inferential identification is practically non-existent.
- Terms of use

Public-use Data

- Most widely used because it is the most widely available

- Making a public-use rendition of data is very important

- Analysis of public-use data may lead to more complex analysis with restricted-use data

Public-use Data

- Most mitigations of disclosure risk are for making a public-use rendition

- Mitigations may be still be insufficient for public-use data.

Public-use Data

- Public policy decisions
- Tracking trends

Restricted-use Data

- All direct identifiers have been removed.
- Inferential identification is possible.
- Data may contain sensitive information.
- Data Use Agreements

Terms of Use vs. Data Use Agreement

- While these phrases may not be standard in all organizations and countries, *the concepts are*
- *Terms of Use: Individual* researcher can accept and access data
- *Data Use Agreement:* Understanding between *organizations* so a an individual at the receiving organization can access the data.

Sensitive Data

Information that can cause harm or legal jeopardy; damage reputation

Some examples are:

- Health information
- Drug use
- Criminal record
- School record

Sensitive Data

- Mitigating disclosure risk for sensitive data is particularly important.

- The disclosure risk threshold for data with sensitive information is more risk averse.

- All information about minors is automatically sensitive.

Confidential Data

Information that has been promised to keep secret

Proprietary Data

- Information that is owned.
- Data for which permission to distribute has not been given.
- May not be sensitive nor confidential

III. What Increases Risk

- Risk Factors
- Inferential Issues
- Hierarchical and Longitudinal Data
- Linkages

Risk Factors

- Sampling frame
  - If sampling fame is known or can be recreated, disclosure risk increases dramatically
- Individual known to be in study
- **Self disclosure** (often via social media)

Risk Factors

- Small and special populations
- Small geographic areas
- Cluster sample

Risk Factors

- Geographic information
- Longitudinal data
- Special purpose studies

Inferential Issues

- Small cell sizes
  - Check if combinations of variables produce unique or nearly unique individuals
  - Variables that could form links to other data
- Low levels of geography
- Hierarchy such as schools and students or hospitals and patients

Inferential Issues

- Sampling clusters often contain imbedded geography or disclosive hierarchy such as health facility or school

- Sampling frame can be reconstructed for small or special populations

- Histories are more likely to be unique for an individual

Inferential Issues

- Variables with high level of detail have higher disclosure risk

- Examples:
  - Exact age
  - Exact income
  - Detailed household structure

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 need to be protected from disclosure too.

Data Linkages

- Links to other data sources may make data disclosive.

- Information that can be used for statistical matching.

- Geography, Demography, History

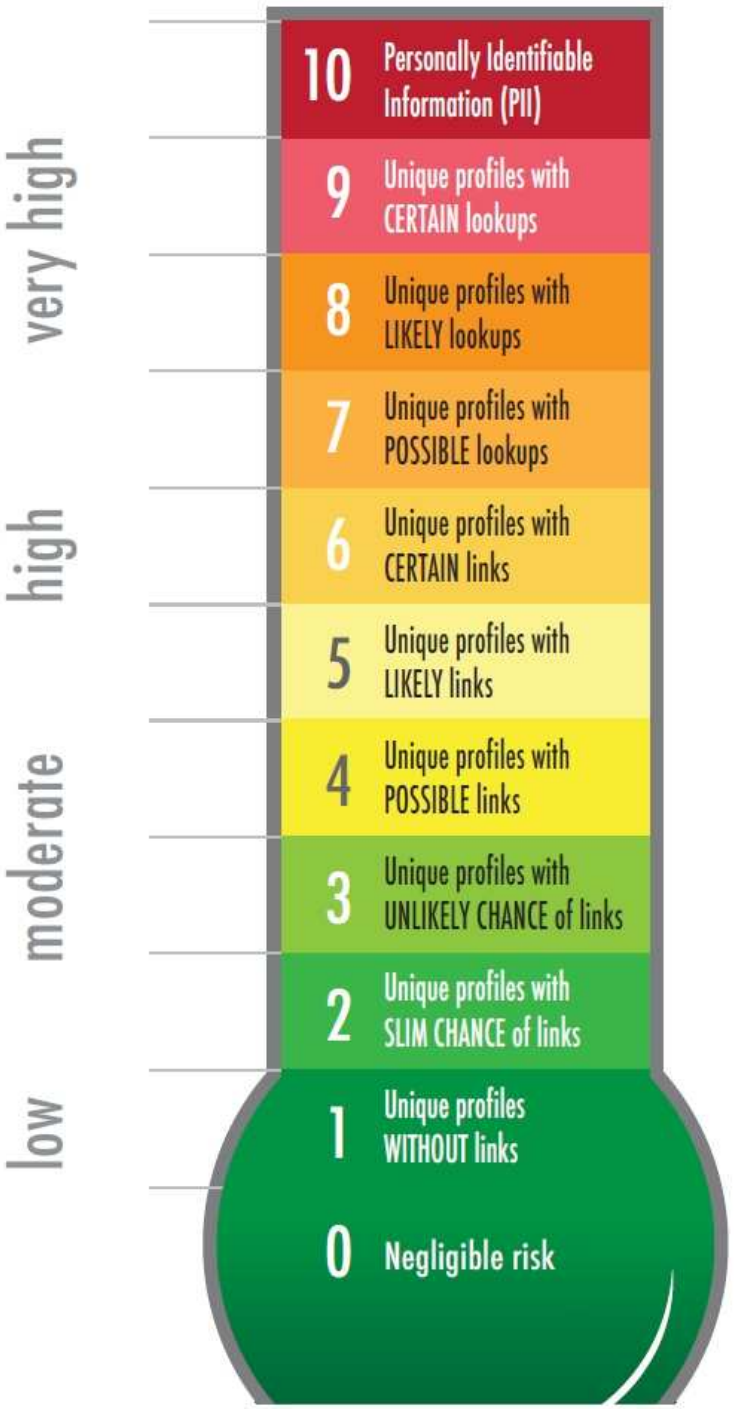
Potential Data Linkages

- Other studies
- Administrative data
- Social media; people self identify as being part of a study
- “Big Data”

IV. Quantification of Risk

- No consensus scale
- Probabilities are subjective
- Categories are fuzzy

Re-identification Risk



Harm

Low	0	No Harm
	1	Little Harm
	2	Humiliation
	3	Reputation Damage
Moderate	4	Financial Loss
	5	Health Threat
	6	Legal Jeopardy
High	7	Prison
	8	Physical Injury/Impairment
	9	Disfigurement
	10	Death

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.

Chances of Link

- **Certain:** links with other data are available but not re-identification ($P > 0.75$)

- **Likely:** links are probably available ($0.5 < P < 0.75$)
- **Possible:** links may exist ($0.10 < P < 0.50$)

- **Unlikely:** links are improbable ($0.05 < P < 0.10$)

- **Slim:** links are highly improbable ($P < 0.05$)

Chances of Lookup

- **Certain:** lookup will return identity from information supplied ($P > 0.75$)
- **Likely:** lookup will probably return identity but not certain ($0.5 < P < 0.75$)
- **Possible:** lookup might be able to return identity but not likely ($P < 0.5$)

Samples and Populations

- National samples
- Small area samples
- Small populations

Samples and Populations

- Numerators
- Denominators

Unique Combinations

- Unique combinations of variables that form a profile

- Unique

ICPSR

SHARING DATA TO ADVANCE SCIENCE

Data Modifications to Reduce Disclosure Risk

ICPSR INSTITUTE FOR SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN

Data Modifications

- Modify data sufficiently reduce risk to make public-use
- Modify data for restricted-use access
- Depend on the purpose of the data

Data Modifications

- Suppress variables
- Replace values with random numbers
- Collapse categories, coarsen coding, top and bottom limits
- Perturb variables by adding random noise
- Swap records

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

Sampling Variables

- Sampling variables such as Strata and Cluster my increase disclosure risk because of embedded information

- These variable are necessary to compute standard errors for data from complex samples.

Suppress Variables

- Almost all data for research have variables that have been suppressed

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

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

Replace Values

- Examples:
  - PII
  - Respondent ID
  - Household ID
  - Sampling Stratum and Cluster

- Inferential re-identification still possible

Coarsen Coding

- Coarsening coding is primary method for keeping variables in the data

- Collapse categories

- Loss of information

- Reduces the level of measurement

Coarsen Coding

- Measurement in survey data is often not as good as some variables imply
- Researchers often want to do their own collapses of variables
- Coarsening may be impractical

Coarsen Coding

- Top and Bottom coding

- Examples of variable that typically are coarsened: Age, Education, Income

Perturb Variables

- Add some random noise
- Mean unaffected
- Increases variance
- Attenuates correlations

Perturb Variables

- Maintains level of measurement
- Variable can be used in linear regression
- Attenuates effects

Perturb Variables

- Examples of Variable that are candidates for being perturbed: Age in months, Height, Weight

Swap Records

- Maintains most univariate statistics
- Maintains some correlations
- Deniability if someone claims re-identification

Swap Records

- Swapped records should only match on non-disclosive variables
- If match is based on too many variables, the disclosure protection is attenuated
- Matched variables are the highest analytic value
- Maintains some correlations

Swap Records

- Swapping records between geographic areas is most common

- Data with sensitive information may swap records

- 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.

Data Sharing for Demographic Research

Data Linkages & Disclosure Risk

John Marcotte, PhD, and Sarah Rush, MSW
ICPSR, University of Michigan

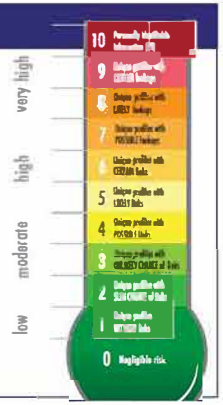
Background

A great variety of data sources are now available to researchers. Analysts may wish to study outcomes from one dataset with predictors from another data source. Combining data from multiple sources can enrich research and increase analytic potential. At the same time, linking data can increase the risk of re-identification and disclosure.

Measuring Disclosure Risk

Degree of cisclosure risk is related to presence of:

- Unique Profiles: Set of variables that when combined together can be used to link data to other sources
- Links: Pieces of information that can be used to connect data (e.g., external ID)
- Lookups: Information that translates profiles into identities



Ways to Reduce Disclosure Risk

- Linked data can be made available as restricted-use, requiring an application and approval process for access
  - Linked data with high or very-high risk can be made available via enclave only
- Strategies can be implemented to mitigate risk
  - Swapping records
  - Minimum cell and sub-sample sizes
  - Suppressing link variables
- Tighter security and enhanced monitoring of researchers performing own linkages
  - Masking of internal IDs when possible

Types of Linkages

Data files are linked using identifiers (linking variables). Identifiers differ based on the type of linkage and matching used.

TYPE OF LINKAGE	LINKS OCCUR BY	EXAMPLES	MATCH TYPE
Between waves	Internal IDs	Case ID, unique to study	Exact
Overlapping cross-sections	Internal IDs	Case ID, unique to study	Exact
Separate sources	External IDs	PII: SSN, Health ID	Exact
Contextual linking (incl. geography)	Geocodes, Org codes	Zipcode, Tract Number, County Code, School Name, Hospital Name	Exact
matching	Common variables	Age, Sex, Ethnicity, Education	Probabilistic

Sources of Disclosure Risk in Linked Data

Disclosure Risk varies based on the type of data linked and the combined information contained in the linked files.

SOURCE OF DISCLOSURE RISK	TYPE OF LINKAGE: MATCHING (Common Variables)	PANEL WAVES (Internal IDs)	OVERLAPPING CROSS-SECTIONS (Internal IDs)	CONTEXTUAL LINKING (incl. geography) (Geocodes, Org Codes)	SEPARATE SOURCES (External IDs)
LINKAGE VARIABLE Examples: Case IDs, PII, PHI, FIPS	6	7	8	9	10
RESPONDENT'S HISTORY Examples: Employment History, Marital History	5	6	4	2	3
OTHER RESPONDENT INFORMATION Examples: Household Structure, Employment Status, Area Characteristics	3	4	6	8	7

Summary

Linkages provide richer data for researchers, but disclosure risk may increase substantially. While access to linked data con be provided through restricted-use data agreements and security plans, results from these combined datasets must still be reviewed for disclosure risk. Even if IDs and other linking variables are removed after data are combined, histories and additional information still increase disclosure risk.

dsdr.icpsr.umich.edu

Eight Questions

True of False

1. Data with serious disclosure risk are easy to recognize.
2. Unintended or accidental disclosure is not culpable.
3. Data without direct identifiers such as names, addresses or SSNs may still have serious disclosure risk.
4. Two variables may be sufficient for inferential disclosure.

Eight Questions

True of False

5. Geographic variables are always recognizable in data.
6. Sampling frames can be reconstructed.
7. Only data collected explicitly for research have disclosure issues.
8. Disclosure of non-sensitive information is acceptable.

Eight Answers

1. Data with serious disclosure risk are easy to recognize. *False. Inferential disclosure risk is often subtle.*
2. Unintended or accidental disclosure is not culpable. *False. Accidents and ignorance are not excuses.*
3. Data without direct identifiers such as names, addresses or SSNs may still have serious disclosure risk. *True. Inferential disclosure is usually the issue.*
4. Two variables may be sufficient for inferential disclosure. *True. Birth dates, income, family characteristics combined with geography.*

Eight Answers

5. Geographic variables are always recognizable in data. *False. Geographic data is often imbedded in sampling clusters.*
6. Sampling frames can be reconstructed. *True. Particularly for small or specialized populations.*
7. Only data collected explicitly for research have disclosure issues. *False. Administrative data are usually highly disclosive.*
8. Disclosure of non-sensitive information is acceptable. *Usually false, but depends on the terms of collection.*

(1) How would you assess the risk?

- Opinion poll on college students' favourite holiday with variable containing respondent names.

(2) How would you assess the risk?

- Opinion poll on an upcoming election with respondent sex, age, and postal code

(3) How would you assess the risk?

- Survey of sexual behaviour with geographic variable at census region level, and respondent age and religion.

(4) How would you assess the risk?

- Survey of hospital discharges with dates of intake and discharge.

(6) How would you assess the risk?

- Administrative records of nursing staffing levels by hospital units

(5) How would you assess the risk?

- Survey of elementary school children and transportation method to school

ICPSR

SHARING DATA TO ADVANCE SCIENCE

Assessing Disclosure Risk

How ICPSR review data deposits

- ICPSR assesses every data deposit for disclosure risk.

- This assessment can take from 2 hours to 3 days depending on the complexity of the study and the data

ICPSR Disclosure Assessment

1. When a deposit disclosure review begins, the first step involves gathering general background information about the study and methods. Information to look for includes:

- Was confidentiality promised to respondents by PI/organization?
- Examine the sensitivity of the type of data being gathered (Would disclosure of respondent/subject identity inflict a high level of harm?)
- Could subjects of study/survey be classified as sensitive or vulnerable (e.g. children, inmates, etc.)
- Do other geographic/situational circumstances increase the negative consequences for disclosure (for example, a study where citizens give survey responses in a country with a repressive government)?

Methods:

Generally, this is done through a review of the documentation provided with studies. Studies are often explained in greater

detail on associated websites, so internet research is often involved as well.

2. The deposit is next reviewed for direct identifiers. Methods:

String variables are often the most common location of direct identifiers, so browsing strings within frequency tables is generally a first step. UNIX tools are also available that assist in searching for names (fname, lname, mname, ICPSR Anonymizer) within strings and qualitative data.

3. Next, we look for indirect identifiers. The first variables examined are generally the geographic variables, since they dictate the level of disclosure risk present in every other variable. If, for example, the only variable is country or state, those are large enough areas where disclosure risk is much less of a concern. In smaller areas, such as towns or districts, or when an area only has a population of approximately 8,000-10,000 or less, the risk of disclosure is more likely.

After that, other indirect identifiers are In particular, direct identifiers with low observation counts (such as 10 or less in a small geographic area) are the areas of most relevance.

ID variables are examined as well, since they are often not randomized and are often sequenced based on possible identifier variables such as geography, name, or organization.

Methods:

Generating frequency tables to view the values provided with possible indirect identifiers. Any possible indirect identifying variables are listed, then a decision is made on which variables pose a risk. Cross-tabulations are also utilized, particularly crosstabs of possible indirect identifiers with the provided geographic variables. This helps determine whether there are a small number of observations within a certain geographic area. Populations are also checked online to determine population density of certain geographic areas.

4. After reviewing the study itself, curators may need to check for any possible linkages with previous studies or outside information. For example, if a new wave of study contains sensitive information that previous waves did not contain, subjects in the study could be identified from earlier parts of the series.

Methods:

Checking if study is connected to series on ICPSR website. If ID variables are similar across waves or parts, examine if subjects in current study may be disclosed by any variables in previous parts or waves.

I. Direct Identifiers:

These are variables that point explicitly to particular individuals or units. They may have been collected in the process of survey administration and are usually easily recognized. Any variable that functions as an explicit name can be a direct identifier.

Categories:

1. Names
2. Unique identifying numbers
    a. SSN
    b. Account numbers
    c. Certificate or license numbers
    d. Vehicle identifiers
    e. License plate numbers
    f. Serial numbers
    g. Institution ID
    h. Badge Numbers
3. Telephone numbers and facsimile numbers
4. E-mail addresses
5. Web Universal Resource Locators (URLs)
6. Internet Protocol (IP) address numbers
7. Biometric identifiers
    a. Fingerprints
    b. Voiceprints
    c. Specimens
8. Full-face photographic images or any comparable images

Also see the 18 identifiers which are part of the Health Insurance Portability and Accountability Act of 1996 (HIPAA). In addition to the above, this includes

1. Address (all geographic subdivisions smaller than state, including street address, city county, and zip code)
2. All elements (except years) of dates related to an individual (including birth date, admission date, discharge date, date of death, and exact age if over 89)
3. Medical record number
4. Health plan beneficiary number
5. Any other characteristic that could uniquely identify the individual

The preferred method of handling direct identifiers would be to completely mask them. However, it may be that direct identifiers could be completely removed if they hold no analytic value. Ideally, the PI would be consulted regarding the analytic value of such variables before they were to be removed.

II. Indirect Identifiers:

Unlike direct identifiers, indirect identifiers usually require more work to identify, and are often present in many types of variables. As a practice, indirect identifiers are less likely to require removal of values or responses (unlike direct identifiers).

Indirect identifiers should first be addressed with a re-coding of values to combine responses in order to make specific respondents less visible (such as top-coding or other grouping of respondents). When re-coding is not enough to ensure disclosure remediation, the next step is to make the data file (and documentation) restricted use.

Some of these standards were taken from other disclosure related documentation at ICPSR (and list what source or document) and others are general guidelines we think are reasonable to reduce disclosure risk.

The first variables to examine are geographic variables because disclosure risk in these variables often affects disclosure remediation in other variables (because larger geographic units usually provide a lower risk of disclosure).

Geographic variables could be examined with the following standards:

- If state is the smallest geographic unit - are there variables with fewer than 3 observations?

- If smaller geographic units exist - are there variables with fewer than 10 observations? [Taken from Disclosure Risk Assessment Tool.docx]

- Population size: Does town, city, or geographic unit have a population of 10,000 or less?

If any of the following are true, the variables may need to be regrouped into larger geographic regions (for example, towns or states could be re-coded into regions in order to create a larger geographic unit).

The next step after assessing geographic variables is to assess possible indirect identifier variables for low observation counts or outliers.

The following is a list of other indirect identifiers typically found within studies, as well as suggested standards and practices in identifying and addressing them:

<u>a. Race, Ethnicity, and Other Demographic Variables</u>

- If a race or ethnicity variable contains 10 or fewer observations for a given demographic, groups should likely be collapsed in order to remediate disclosure.

For example if there are two race or ethnicity groups (say Hispanic and Asian-Pacific) with a low number of observations, they can be combined into one group to remediate disclosure (one “Hispanic or Asian-Pacific” value). This can be done as many times as necessary until disclosure is properly remediated.

b. Income Variables

- Is exact income listed in the variable? If income is the only indirect identifier included in the data it is usually not very disclosive, but if other disclosive variables are present income information should be re-coded into income bracket groups. Top coding especially high or outlying incomes or salaries should be done as well. Re-coding is generally sufficient for this type of variable, and shouldn't usually require restriction or masking.

c. Physical Feature/Characteristics Variables

- Are exact height, weight, BMI, disability, or other variables describing physical appearance or characteristics present? Only one of these being present may not cause much of a disclosure risk, but more than one being present may require collapsing values into groups.

d. Medical History Variables

- Possible information includes illnesses, medical treatments, doctors' offices or hospitals visited, prescriptions, or other variables related to medical treatment. If other types of disclosive variables present, may need to consider data restriction.

- Some of this information could be considered sensitive as well, especially if it involves younger respondents.

e. Family Characteristic Variables

- Examples include size of household, number of children, etc.

- If household size is 8 or larger, or number of children is 6 or larger, observations variables with these observations should be top-coded appropriately (e.g. all values 8 or more become “8 or more”).

- Besides outliers, family information is otherwise not too disclosive, since most families usually have fairly similar characteristics.

f. Sensitive Behavior Variables

- Examples include drug use, sexual history, illegal behavior, etc.

- If data includes sensitive behavior by minors, data will likely need to be restricted

- Should likely use the most stringent standards for DR, since identity discovery could potentially be harmful

g. Age Variables

- Younger respondents (18 or younger) usually require more disclosure remediation

- Are there observations listed as age 90 or older? Should be top-coded to “Age 90 or older”

- Are exact ages listed in variable, or are ages grouped into fewer values? Consider collapsing age into groups if other disclosive variables are present.

h. Personal History and Characteristic Variables

- Some are often related to geographic variables such as school or university attended, previous or current place of employment, place of birth, or hometown. One variable by itself is usually fine, but multiple variables of this type could be disclosive.

- Others include variables on relationships (married or divorced, relationships with family members or peers, sexual orientation) or variables on beliefs or opinions (including religion, political beliefs, etc.). Sexual orientation is usually the most sensitive information in these types of variables, and a small number of values for one category (e.g. gay, bisexual, or other) should likely be strongly considered for DR.

If a respondent is a notable outlier in one variable, crosstabs with other indirect identifier variables should be performed in order to ensure observations are not outliers in multiple variables (since this may pose a higher level of disclosure risk).

HIPAA 18 Identifiers

1. **Names**
2. **Geographic** subdivisions smaller than a state (except the first three digits of a zip code if the geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people and the initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000)
3. All elements of **dates** (except year) for dates directly related to an individual, including birth date, admission date, discharge date, and date of death and all ages over 89 and all elements of dates (including year) indicative of such age (except that such ages and elements may be aggregated into a single category of age 90 or older)
4. **Telephone** numbers
5. **Fax** numbers
6. **Electronic mail** addresses
7. **Social security** numbers
8. **Medical record** numbers
9. **Health plan** beneficiary numbers
10. **Account** numbers
11. **Certificate/license** numbers
12. **Vehicle identifiers** and serial numbers, including license plate numbers
13. **Device identifiers** and serial numbers
14. **Web Universal Resource Locators** (URLs)
15. **Internet Protocol** (IP) address numbers
16. **Biometric identifiers**, including finger and voice prints
17. **Full face photographic images** and any comparable images
18. Any other **unique identifying number**, characteristic, or code (excluding a random identifier code for the subject that is not related to or derived from any existing identifier)

ICPSR

SHARING DATA TO ADVANCE SCIENCE

Restricted-use Data and Security Options

Restricted-use Data

- Data for which researchers must apply

Restricted-use Data

- Institutional Review Board (IRB) or Ethics Panel approve research plan
- Data Security Plan
- Research Plan
- Confidentiality Pledges

Restricted-use Data

- *Access* to restricted-use data is provided for a limited time
- Only with approved application
- Requires Data Use Agreement between universities or institutions

Restricted-use Data

- Data may be restricted use for a variety of reasons:
  - Re-identification risk
  - Harm risk
  - Proprietary
  - Data provider requirement

Disclosure Risk

- Restricted-use data may have had disclosure remediations
- Disclosure Risk may remain
- Sensitivity of data

Restricted-use Data

- Applicants must be covered by rules that cover research misconduct
  - Faculty
  - Eligible to be Principal Investigator
  - Doctorate or terminal degree
  - Graduate students and staff are *NOT* eligible and require faculty sponsorship
- Some data require a grant

Principal Investigator

- Must be eligible to lead research projects and grants

- Universities and research organizations have internal requirements

IRB

- The Institutional Review Board (IRB) is a panel that reviews research protocols for appropriate protections for research subjects

- IRB may be called other names; regardless of name, the panel must review protections for research subjects

Violations

- In the vent of a violation of research and security protocols, sanctions may occur

Potential Sanctions

- Request receiving organization to render research misconduct sanctions
- Report researcher and receiving organization to grant agencies
- Researchers at receiving organizations may lose eligibility to obtain restricted-use data

Data Collectors

- Study directors
- Data collectors
- Data assemblers

Data Disseminators

- Libraries and Archives
- examples: *ICPSR, Dataverse, IPUMS*
- Provide access to data

Determining Appropriate Security

- Level of disclosure risk
- Sensitivity of data
- Requirements of data provider

Requirements

	Data Collector Restricted-use	Data Collector Public-use
Data Disseminator Restricted-use	Restricted-use	
Data Disseminator Public-use	Restricted-use	Public-use

Restricted-use v. Public-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

End of DUA and Access

- Copy of restricted-use data destroyed or access revoked
- Data extract must also be destroyed
- Any summary results that do not adhere to disclosure protection rules

Retained Items

- Syntax files for statistical software as long as the files do contain unique values for individual respondents

- Summary tables that meet disclosure protection rules for results

- Notes that do include data or specific values

AFFIDAVIT OF DESTRUCTION OF UNAUTHORIZED INFORMATION

I *<name>*, *<title>*, at *<university, school, department>* am the person responsible for analysis and other use of information obtained under an agreement between *<university>* and the University of Michigan on behalf of the Inter-university Consortium for Political and Social Research, dated *<date of agreement>*.

I hereby attest that I have:

1. Shredded all printouts of unauthorized information removed from the VDE;
2. Erased from my computer using a program that overwrites at least 7 times any files that contain unauthorized information;
3. Deleted all copies of the unauthorized information from email on my computer.

Date:_____ *<signature>*___
<name>
<title>
<department>
<university>

Subscribed and sworn to before me on this _____day of _____, 2016

_____
Notary Signature

ICPSR

SHARING DATA TO ADVANCE SCIENCE

Security Options for Restricted-use Data

Security Plan

- Restricted-use data require extra security to protect them from unauthorized access.

- Restricted-use data require extra security to prevent disclosure of individuals or organizations.

Security Layers

- Encryption in transport and at rest
- Blocked Internet: prevent copying files to services such as *Box*, *DropBox* and *Google Drive* (or other Internet sites)
- Vetted output for disclosure risk
- Monitored when accessing data

Determining Appropriate Security

- Level of disclosure risk
- Sensitivity of data
- Requirements of data provider

Security Tension

- Researcher agrees to follow protocols

or

- Technical restrictions

Data Security

The required security for data depends on their sensitivity and identifiableness. Data with more sensitive information and greater potential identifiableness require extra security precautions. Each level adds another restriction layer. The levels are (0) Public-use (0.5) Private-use (1) Restricted-use 1 (2) Restricted-use 2 and (3) Restricted-use 3. Responsibility to protect respondent identities *and* their information

At level 0, public-use data do not require security.

At level 0.5, private-use data require encryption and approval.

At level 1, data must be encrypted at rest and in transmission. An additional security requirements is blocked Internet. The rooms that house the client and server must be lockable. *Data sent to researcher.*

At level 2, in addition to level 1 protections, nothing including output, data and data extracts can be removed from the computing system until vetted for disclosure risk by trained and authorized personnel. Furthermore, files cannot be added without security review. *Researcher comes to data virtually.*

At level 3, in addition to level 2 protections, processing must be monitored by trained personnel. Notes may not be taken. Moreover, all items such as backpacks and briefcases must be inspected for disallowed materials after a processing session ends. *Researcher comes to data in person.* This table summarizes the restrictions:

	Encryption	**Internet**	**Output**	**Processing**	**Distribution**	**Approval**
Public-use 0	Not encrypted	Allowed	Not vetted	Not monitored	Web	Terms of Use
Private-use 0.5	Encrypted	Allowed	Not vetted	Not monitored	Authorized download	Terms of Use with approval
Restricted-use 1 *Data sent to requestor*	Encrypted	Blocked	Self-vetted	Not monitored	CD, DVD or secure download	Data Use Agreement IRB approval
Restricted-use 2 *Requestor comes to data electronically*	Encrypted	Blocked	Vetted	Not monitored	Terminal Server with extra security	Data Use Agreement IRB approval
Restricted-use 3 *Requestor comes to data in person*	Encrypted	Blocked	Vetted	Monitored	Guarded "Cold" Room	Data Use Agreement IRB approval

Data Security

Encryption

Encrypted. Data files, output files, temporary files and other project files must be encrypted *at rest* and *in transport*. Real-time or "on the fly" encryption must be used so that any files placed on the volume are automatically encrypted. Encryption software must be currently maintained. Encryption must meet AES standards.

Internet

Blocked. System must prevent all files from being copied to the Internet. Access to any systems with restricted-use files must not be directly accessible from the Internet. Access must be from designated locations only. Access through a VPN or private address space is acceptable; however, split tunneling is not allowed.

Output

Vetted. Results (tables, regressions, etc.) are checked for disclosure risk by authorized and trained personnel (not the researcher). Output includes data extracts. By requiring vetting, data cannot be copied. Inputs must all also be checked but this vetting is not as stringent.

Processing

Monitored: Data analysis can only occur in the presence of authorized and trained personnel.

Distribution

Web: Data are available on the ICPSR website

CD or download: An encrypted version of the data are sent on a CD or available for download

Terminal Server with extra security The ICPSR Virtual Data-Enclave (VDE) meets this requirement. After installing software and obtaining login credentials, a researcher may connect to the VDE to analyze data. Data and results cannot be downloaded.

Guarded "cold" room: The ICPSR Physical Data-Enclave meets this requirement. Researcher must come to the ICPSR (Perry) building to analyze data. Analysis times are restricted to ICPSR operating hours.

Approval

Terms of Use: Researcher must agree to terms of use before downloading data. Some data require approval.

Data Use Agreement. Researcher and researcher's institution must enter into a data use agreement with the University of Michigan/ICPSR.

Pledge of Confidentiality

By virtue of my affiliation with this research project I have access to Confidential Data identified in this Agreement. I understand that access to this Confidential Data carries with it a responsibility to guard against unauthorized use and to abide by the Data Security Plan. To treat information as confidential means to not divulge it to anyone who is not a party to the Agreement for the Use of Confidential Data, or cause it to be accessible to anyone who is not a party to that Agreement.

I agree to fulfill my responsibilities on this research project in accordance with the following guidelines:

1. I agree not to permit Confidential Data access to anyone not a party to the Agreement for the Use of Confidential Data, in either electronic or paper copy.
2. I agree to not attempt to identify private persons as defined in the Agreement for the Use of Confidential Data.
3. I agree that in the event an identity of any private person is discovered inadvertently, I will (a) make no use of this knowledge, (b) report the incident to ICPSR, (c) safeguard or destroy the information after consultation with ICPSR, and (d) not inform any other person of the discovered identity.

DSDR Data Sharing for Demographic Research

Alternatives to "Cold Rooms" for Restricted-use Data

John Marcotte, PhD, and Sarah Rush, MSW
ICPSR, University of Michigan

Introduction

- More and more studies are producing data with disclosive and sensitive variables.
- As the risk of re-identification and harm increases, so does the security level needed to ensure protection to participants.
- Modifying data for public use may reduce the analytic value tremendously.
- Challenge is to provide appropriate data protections while making data as accessible as possible.

Options for Providing Access to Restricted-Use Data

- Physical enclaves or "cold rooms" are often used to make data having increased risk of re-identification and potential for harm available.
- Virtual analysis systems and batch systems with synthetic data provide the same level of protections as cold rooms and are more accessible to researchers.

COLD ROOMS & THEIR ALTERNATIVES

Traditional Cold Rooms

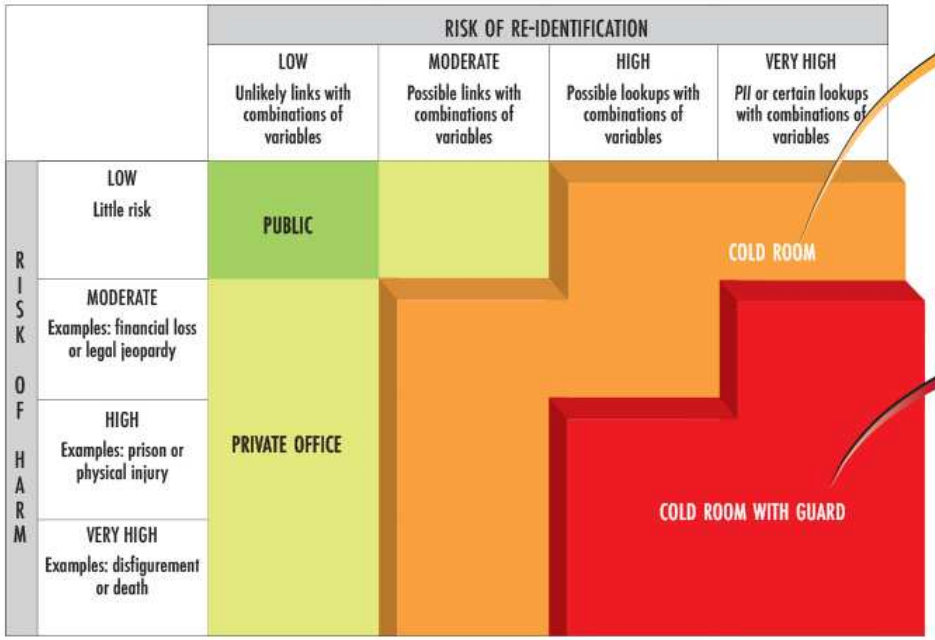
- Types: Unguarded and Guarded.
- Computers not connected to public network or printers.
- USB ports for connecting flash drives are disabled and inaccessible.
- Researchers must submit requests to remove output.
- Output released to researcher only after passing disclosure review.

Benefits

+ Data cannot be copied.
+ In guarded room, data manipulation is monitored and notes may not be removed from the room.

Disadvantages

– Inconvenient; often requires travel.
– Personnel to maintain and guard rooms makes them costly to operate.
– Space is often a premium resource at universities and other research organizations.



Alternatives to Cold Rooms

Virtual Analysis System

- Encrypted connection to online server for data analysis while blocking connections to the Internet.
- Files can't be copied from the server.
- Researchers submit requests to remove output.

Benefits

+ Researchers connect to the server from a specified, private location (e.g., office).
+ Facilitates collaboration; each member of project team can connect to server from an approved location.
+ Server may offer more processing power than a desktop computer.

Disadvantages

– Like Cold Rooms, note taking isn't prevented.

Batch System & Synthetic Data

- Submitted data analysis programs run on a server with the restricted-use data.
- No access to restricted-use data.
- Output is only provided to researcher after passing disclosure review.

Benefits

+ No travel required to submit programs.
+ Less costly than a guarded cold room.
+ Prevents unauthorized note taking.

Disadvantages

– Researchers must wait to receive output.
– Programming mistakes may take days to detect.
  - To address this, synthetic data can be used to debug programs. Fully tested programs can then be submitted to run on the restricted-use data.

ICPSR

SHARING DATA TO ADVANCE SCIENCE

Vetting Output for Compliance

John E Marcotte, PhD
ICPSR
University of Michigan
July 2018

Vetting Output

- Checking if summary results comply with disclosure protection rules

- NOT checking for validity of the analysis or the scientific value

Disclosure Protection Rules

Disclosure protection rules define what results from analysis from restricted-use data may be presented or published. These rules prevent the indirect re-identification of respondents and organization.

Rule	**Description**	**Values**
PII or PHI	Personally Identifiable Information such as names, addresses and respondent ID cannot be reported.	
Suppressed Variables	While these variables can be included in analysis, coefficients and tables for them cannot be reported	
Suppressed combinations of variables	While these variables can be reported separately, they may not be used together in tables or interactions	
Minimum cell sizes	For tables, minimum allowed cell sizes. Cells below this value require rows or columns to be combined. Redaction of the individual cell is insufficient.	*10*
Minimum sample and sub-sample size	Minimum number of valid observations (excluding missing data) for regression analysis	*50*
Disallowed sub-samples	Sub-samples that are not allowed even if the sub-sample meets sample size requirements	
Dummy variables	Dummy variables for which coefficients cannot be reported	
Organizations and Groups	Organizations and Groups for which results cannot be presented separately	
Nested tables	Tables that can be combined into one table	*Should be presented as single table*
Saturated or near saturated models	Models that reproduce the data exactly	*Maximum R-squared* *Minimum df remaining*
List cases including predicted values	An individual case or roster of cases cannot be reported.	*Micro data will not be released*
Weights	Do results have to be weighted?	*Unweighted counts for table totals only*
Visualizations		*Maps must obscure exact locations*

Study: Sample of 433 adults from all over Canada

Linear Regression: BMI (Body Mass Index) on Education, Marital Status, Type of Place

Y-variate = BMI

R-squared = 0.77; df = 20

	Coefficient	Standard Error
Constant	22.43	10.61
Education	-0.94	0.67
Married	1.12	0.89
Urban	-1.48	0.77

Study: Sample of 433 adults from all over Canada

Linear Regression: BMI (Body Mass Index) on Education, Marital Status, Type of Place, BMI (previous year)

Y-variate = BMI

R-squared = 0.82; df = 431

	Coefficient	Standard Error
Constant	21.34	9.16
Education	-0.91	0.54
Married	0.88	0.41
Urban	-1.11	0.51
BMI previous year	-3.14	0.83

Study: Sample of 413 patients from four hospitals in Ontario. Data are provided under three conditions: (1) The identities of patients and hospital staff are kept confidential; (2) The names of hospitals will remain confidential; (3) The data will not be used to rank hospitals

Linear Regression: CAUTI (Catheter-Associated Urinary Tract Infections) on Time of Insertion, Nursing level of Inserter, Hospital

Y-variate = CAUTI (yes/no)

R-squared = 0.14; df = 218

	Coefficient	Standard Error
Constant	0.0837	0.0341
Day Insertion	-0.0224	0.0082
RN	-0.0148	0.0091
Hospital 2	0.0114	0.0062
Hospital 3	-0.0101	0.0047
Hospital 4	0.0251	0.0019

Study: National Sample of 1274 adults about smoking and drinking behaviour.

Tobacco Product User by Race

	White	Black	Other
Cigarettes	143 16.0%	30 19.1%	22 31.0 %
Cigar or Pipe	52 5.8%	11 7.0%	2 2.8%
Snuff or other	11 1.2%	1 0.6%	1 1.4%
None	689 77.0%	115 73.3%	46 64.8%
Column Total	895 100.0%	157 100.0%	71 100.0%

Education Level by Place Type

	Urban	Suburban	Rural
Less than High School	1519 10.0%	1369 6.0%	1462 12.0%
High School	8376 55.0%	10283 45.0%	6702 55.0%
College	4570 30.0%	8682 38.0%	3656 30.0
Post College	764 5.0%	2513 11.0%	366 3.0%
Column Total	15231 100.00	22847 100.0%	12186 100.0%

Education Level by Place Type for Whites and Blacks

	Urban	Suburban	Rural
Less than High School	1158 8.0%	886 4.0%	1206 11.0%
High School	7856 54.3%	9910 44.7%	6004 55.0%
College	4630 32.0%	8864 40.0%	3400 31.0%
Post College	825 5.7%	2501 11.3%	357 3.0%
Column Total	14469 100.00	22161 100.0%	10967 100.0%