ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

Disclosure Risk Analysis

ICSPR Official Representative (OR) Meeting
Tuesday, October 6, 2009
2:00 p.m.

JoAnne McFarland O'Rourke

Learning Objectives

Participants attending this session will be able to:

1. Define and cite examples of disclosure;
2. Describe methods used to protect social science research subjects from disclosure risk in public-use releases of microdata, such as microaggregation and data swapping;
3. Name 5 essential steps for disclosure risk reduction;
4. Summarize ICPSR's training, implementation, goals, and certification process for disclosure risk assessment.

Reasons for Increased Concern

- Easy access via the Internet
- Explosion of information available
- Increased sophistication of databases used for matching
- Extremely low cost of hard disk space
- Numerous files available from single sources

Disclosure Risk Training Content

- Disclosure risk framework (overview)
- Overview of social science research process
- Research designs, sampling plans, and data collection: impact on disclosure risk
- Relationship between disclosure analysis and the confidentiality requirements of ethical research
- Dissemination of findings: impact on risk
- Examples, group exercises

Training (con't.)

- Web-based certification test – auto-scored, plus short answer
- 20 questions randomly selected for each test (person)
- 40 Collection Development staff members trained in 3 sessions, October and December 2008, February 2009.
- 33 staff passed test (80% required to pass).

Importance of Best Practices

- Set consistent standard across all ICPSR projects
- Extend staff training
- Improve awareness of disclosure risk issues; set organizational mindset of disclosure protection
- Ensure that confidentiality of respondents is maintained throughout lifecycle of data, including confidentiality promised to survey respondents during data collection.
- Ensure continued participation of survey respondents (provide confidence regarding data protection).
- Improve consistency in ICPSR data releases
- Assist other organizations with these issues
- Brief history: Link to disclosure poster

Disclosure Definition

Disclosure relates to inappropriate attribution of information to a data subject, whether an individual or an organization.

- Disclosure occurs when a data subject is identified from a released file (identity disclosure),
- Sensitive information about a data subject is revealed through the released file (attribute disclosure), or
- Released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible (inferential disclosure).

Citation: Confidentiality and Data Access Committee (CDAC), Federal Committee on Statistical Methodology. Statistical Policy Working Paper 22, Report on Statistical Disclosure Limitation Methodology (2nd version, 2005), http://www.fcsm.gov/working-papers/spwp22.html.

Disclosure Definition (con't.)

- If you already know it, it is not disclosure.
- If you find your own record, it is not disclosure.
- Knowing a person participated in a study is not disclosure **as long as** you cannot find that person's record with a high level of certainty; **exceptions include** when participation alone reveals something private.
- Confidentiality v anonymity.
- Most scenarios discussed are possible though unlikely as long as reasonable disclosure protections are added to released data. However:
  - People represented by the data have been assured confidentiality;
  - Many data subjects are in vulnerable positions;
  - Researchers, archivists, information scientists, and others have legal and ethical responsibilities to uphold throughout the data lifecycle, including data dissemination.

Types of Disclosure

- Identity disclosure (or re-identification)
  - Study records match an external file on overlapping data;
  - Matching data could include race, ethnicity, DOB, gender, and city;
  - Binary outcome: match or no match for each record.
  - ✓ Example: Record number 7237 is Mary Jones' record.
- Attribute (or predictive) disclosure
  - More subtle; uses indirect identifiers; degree of disclosure.
  - Occurs if you can determine from microdata alone that an individual or group has a distribution on a sensitive variable that is significantly different than the population's.
  - Can result from empty cells in tables.
  - Re-identification is not necessary.

Types of Disclosure (con't.)

- Attribute disclosure (continued)
  - ✓ Example: The researchers concluded that no one who participated in the study contracted lung disease from environmental factors; Joy Smith was in the study; therefore, Joy Smith did not contract the lung disease from environmental factors.
  - See: http://www.fcsm.gov/99papers/massell.pdf.
- Inferential disclosure
  - Typically concerns aggregate data;
  - Occurs when information can be inferred with high confidence from statistical properties of the released data.
  - ✓ Example: Data show that income and home purchase price are correlated; purchase price of home is public information; therefore, income can be inferred for a data subject by a third party.
  - See: http://stats.oecd.org/glossary/detail.asp?ID=6932.

Disclosure Risk

- Disclosure risk occurs if (1) an unacceptably narrow estimation of a respondent's confidential information is possible or (2) exact disclosure is possible with a high level of confidence (see citation).
- People who may do damage referred to as "data intruders", "data snoopers." Also, an analyst may inadvertently find the record of someone s/he knows.
- Cannot protect against insider who wants to do harm through disclosure protections. Data security measures can help reduce risk:
  - Enforce access controls (e.g., physical access, passwords for electronic access);
  - Create an environment of data protection and awareness about data security through formal staff training, discussion, and established protocols.
  - Document procedures.

Citation: Statistics Netherlands, Statistics Canada, Germany FSO, University of Manchester, 2005, Glossary of Statistical Disclosure Control, incorporated in paper presented at Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, 9-11 October 2005. See: http://stats.oecd.org/glossary/detail.asp?ID=6917.

Common factors related to risk

- Geography
- Publications
- Outliers
- Unique combinations of variables
- File level concerns (matching databases)

Disclosure Review and Analysis

- Disclosure review
  - Examination of **overall** factors related to risk
  - Includes ICPSR's current confidentiality review
- Disclosure analysis
  - **Detailed** examination of
  - Indirect identifiers (e.g., combinations of sensitive variables)
  - External files that may match data under review
  - Typically involves internal staff effort and external committee members, such as the SAMHDA Disclosure Review Committee (DRC).

Disclosure Analysis

- Examine risk to confidentiality
  - Indirect identifiers (e.g., age, gender, race, location – taken together)
  - Focus is characteristics **that can be known** (demographics) not beliefs or attitudes
  - External files;
- Goal: create a public use (or shared) file of high utility that protects respondents
  - Risk/utility balance
  - Must look at study as a whole
  - Introduce uncertainty / provide deniability
  - Due diligence
- Art and Science
  - What is acceptable risk?
  - Approaches vary

Addressing Risk

- Distribution options for files with unacceptable risk:
  - Public-use
    - Apply disclosure protection techniques to the data in order to alter the data yet provide high utility;
    - Data protection also called Statistical Disclosure Control (SDC), Disclosure Limitation (DL);
  - Restricted-use
    - Secure Survey Documentation and Analysis (SSDA)
    - Remote access (under consideration at ICPSR)
    - User contracts
    - Data enclave

Disclosure Risk and Analysis Examples

SAMHDA Disclosure Review Committee

- JoAnne McFarland O'Rourke, ISR/ICSPR
- Steven G. Heeringa, ISR SRC
- Stephen F. Roehrig, Carnegie Mellon University
- Margaret A. Overcashier, ISR/ICPSR
- William C. Birdsall, U-M School of Social Work
- Beth Glover Reed, U-M School of Social Work

- Disclosure work funded through SAMHDA by the Office of Applied Studies (OAS), Substance Abuse and Mental Health Services Administration (SAMHSA).

Example 1:
Treatment Episode Data Set (TEDS)

- National substance abuse treatment admissions
- State-level data
- Geography also includes PMSA, CBSA
- ~1.9 million records
- Annual

Disclosure risk: unique records

- Indirect identifiers combine to create unique records
- Detailed geography increases risk
- Sensitive study content

Gender	Female
Age	30-34
Race	American Indian
Ethnicity	Not Hispanic
Veteran Status	Yes
Pregnant	Yes
Methadone planned?	Yes
CBSA	Kokomo, Indiana

Indirect IDs: Unique Combinations

- Combination of multiple variables can lead to isolated cases
- Issue: Unique in sample v unique in population
  - Sample uniques are typically assumed to be population uniques because the denominator (pop uniques) is typically unknown – this is not always the case
  - *Guidelines*:
    - Run cross-tabs of key variables to assess cell size and disclosure risk
    - Collapse categories with small frequencies
    - Extract a sub-sample, swap records, recode, or add random noise to the problem variables
    - Citation: CDAC (see previous note).

TEDS Goals

1. Protect confidentiality of treatment clients
2. Leave as much detail as possible on files
3. Provide level of detail that is adequate for the states
4. Create disclosure protection procedures that could be easily repeated

DP Methods for TEDS

- Removed high risk / low utility variables (variables with both detailed codes and high % missing)
- Re-coded and categorized some variables
- Used data swapping technique to protect unique (at-risk) records
- Advantages of data swapping
  - All data remain in the file
  - Good choice for data requiring geographic identifiers, files with a large number of records

Data Swapping Basics

- ID at-risk records (e.g., unique records) based on a set of key variables (uniques key)
- Match unique records with other records in the dataset on a set of pre-determined variables with other records in the data set (swapping key)
- Switch matched records
- Only a sample of at-risk records are swapped, based on "Swapping *p*" (proportion of unique records that are swapped); p value is held confidentially.

Data Swapping (con't.)

- Definition of unique record
  - Based on characteristics that can be known
  - Single record that is unique within geographic area
  - Variables that, *when taken together* could ID a person in the file
  - Missing data within uniques key
- Without first re-coding and categorization, many records would be unique (e.g., if age was continuous)
  - Could categorize, swap, then un-categorize if categorizing was not important to the disclosure protection plan
- Unique records remain in the file, cannot be certain from where they originated

Uniques and swapping keys

Uniques key: Variables used to ID unique records
Swapping key: Variables used to ID matching records, acceptable for swapping with unique records

Example: Record 001 is unique record; record 287 is swappable with record 001.

ID	Age	Gender	Meth	Race	Eth	Preg	Vet	PMSA	Prim Sub
001	30-34	Female	Y	White	Mexican	Y	Y	Portland	Heroin
287	30-34	Female	Y	White	Mexican	Y	N	Seattle	Heroin

Example 2:
Alcohol & Drug Services Study (ADSS)

- Facility data (services, revenue, staffing, client characteristics)
- Client record abstracts (urine screen results, services, dates, tx history, health)
- Initial client interviews (drugs, illegal acts, health, income)
- Follow-up client interviews (similar to initial interview)
- National sample

Disclosure risk: matching external file

	Total Revenue	Geo. IDs
ADSS	726,093 (1)	None (2)
Matching File 1	726,093 (3)	PMSA (4)

Matching File 2 included same variables as Matching File 1, plus County FIPS Codes

File Level Concerns

- Is the data file linked (or linkable) to other resources which identify elements of the file?
- Are the data longitudinal?
- Are the data part of a larger data system?
  - *Guidelines:*
    - Remove linkage keys
    - Coarsen, mask, or recode the data
    - Citation: CDAC (see previous note).

ADSS Disclosure Protection Goals

- Retain facility and client file linkages within ADSS
- Release all or nearly all the data
  - Restricted-use data agreement not an option
- Retain utility
  - Categorizing problem variables would eliminate the possibility of determining averages and ratios

DP methods: Microaggregation

- Involves averaging the values of a set of records and applying the average to the record set
  - Identify problem variable
  - Sort data on the problem variable
  - Group records by 3
  - Calculate mean for each grouping
  - Apply mean to each record within each grouping
- Advantages; works well with
  - Data that are correlated
  - Data that can be logically averaged (counts, revenue, etc.)
  - Retains ability to use measures of central tendency (this ability is precluded by top- and bottom-coding or categorizing)

Microaggregation Example

Total Client Count
ORIGINAL, Column Means

	In-patient	Out-patient	Intensive OP
Facility A	250	480	52
Facility B	261	500	27
Facility C	270	650	38
COLUMN MEAN	260	543	39

Total Client Count
RECODED

	In-patient	Out-patient	Intensive OP
Facility A	260	543	39
Facility B	260	543	39
Facility C	260	543	39

Five Essential Steps for Risk Reduction

1. Review and remove direct identifiers
2. Review specific dates and remove and recode them as needed
3. Review specific geography and remove or recode it as needed,
4. Review possible links to external files
5. Re-number cases*

Notes:
* (1) Exceptions are made for files that require linkage retention (e.g., linkage (a) has an analytic purpose (b) is required by archive to provide.
(2) Some of these steps were done as ICPSR's confidentiality checks.

Decision Tree

[handout]

Questions?

Thank you!

jmcfar@umich.edu