ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

A PARTNER IN SOCIAL SCIENCE RESEARCH

ICPSR Disclosure Risk Review Training

Training Purpose and Overview
Disclosure Risk Framework

October, 2008

JoAnne McFarland O'Rourke

Training purpose

- Increase understanding of disclosure risk
- Implement five essential steps for disclosure risk reduction
- Complete Disclosure Risk Assessment (disclosure review)
- Be able to include DRA & procedures in processing plans
- Recognize need for disclosure analysis, but not to conduct the analysis (at this point)
- Improve consistency in data releases at ICPSR
- Improve respondent confidentiality protection

Training purpose (con't.)

- Inaugural training (especially), feedback on:
  - Format
  - Content
  - Length
  - Other suggestions

Training Overview

- Training is intended for Collection Development staff, current staff and new hires.
- Training distinguishes three steps
  - Five essential steps for disclosure protection
  - Usability processing
  - Disclosure review
- Prerequisite: PEERRs "*Human Research, Social and Behavioral Sciences*" module http://my.research.umich.edu/peerrs
- Next step will be training on disclosure analysis

Training includes …

- Sessions that provide fundamental groundwork for assessing risk:
  - Informed consent, research ethics, IRB role
  - Research and sampling designs, data collection methods
  - Data dissemination
- Sessions on common risks and methods for protecting against them
  - Disclosure risk examples
  - Essential Steps for Disclosure Risk Reduction
- Training on how to conduct a disclosure risk assessment
  - Decision tree components and flow
  - Disclosure Risk Assessment, class exercise
  - Small group exercise
  - Homework review, class discussion
- Certification

Importance of Best Practices

- Set consistent standard across all ICPSR projects
- 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 in terms of disclosure risk.
- Assist other organizations with these issues

Disclosure Risk Framework

ISR Privacy Pledge

- What does the ISR Privacy Pledge include?
- Why do all employees need to sign it?

ISR Privacy Pledge

- I will not reveal the name, address, telephone number, or other identifying information of any respondent (or family member of a respondent or other informant) to any person other than a member of the research staff directly connected to the study in which the respondent is participating.
- I will not reveal the contents or substance of the responses of any identifiable respondent or informant to any person other than a member of the staff directly connected to the study in which the respondent is participating, except as authorized by the project director or authorized designate.
- I will not contact any respondent (or family member, employer, other person connected to a respondent or informant) except as authorized by a member of the staff directly connected to the project in which the respondent is participating.
- I will not release a dataset (including for unrestricted public use or for other, unrestricted, uses) except in accordance with policies and procedures established by ISR and the Center with which I am affiliated.

ISR Privacy Pledge (con't.)

- I agree that compliance with this pledge and the underlying policy is: 1) a condition of my employment (if I am an employee of ISR), and/or 2) a condition of continuing collaboration and association with ISR (if I am not an employee of ISR, such as a student, visiting scholar, or outside project director or co-principal investigator).

- If I supervise non-ISR employees who have access to ISR respondent data (other than unrestricted public release datasets), I will ensure that those employees adhere to the same standards of protection of ISR respondent **privacy**, **anonymity**, and **confidentiality**, as required by this pledge and the associated policy.

What is "Disclosure"?

- "Disclosure takes place if tabulations or microdata make it possible to determine the value of some characteristic of an individual (including organizations) more accurately than would otherwise have been possible." (Federal Committee on Statistical Methodology)
  - If you already know it, it is not disclosure
  - If you find your own record, it is not disclosure
  - In most cases, knowing a person participated in a study is not disclosure as long as you cannot find that person's record with certainty; exceptions are when participation alone reveals something private
- Confidentiality v anonymity
- Impossible to protect against an insider who wants to do harm
- Most scenarios discussed are possible but unlikely; however, people represented by the data have been assured confidentiality, some are in vulnerable positions; legal and ethical responsibilities.
- People who may do damage referred to as "data intruders", "data snoopers", may be people who didn't mean to stumble across their neighbor's record.

Types of Disclosure

- Identity or Re-identification
  - Matching study data to external file(s) using overlapping data; binary outcome
- Attribute
  - More subtle; indirect identifiers; degree of disclosure
  - Occurs if you can determine from microdata that an individual or group has a distribution on a sensitive variable that is significantly different than the population's, including estimation of the group's distribution from the sample. Also can result from empty cells in tables. Examples:
    - (Non-sensitive data): Women in a study, based on a scientific random sample, ate 20% more ice cream than men in the study. Therefore, all women eat 20% more ice cream than men.
    - No one received an A in the class. John Smith was in the class; therefore, John Smith did not receive an A.
    - No one in the study contracted the illness from environmental factors; therefore, Joy Smith did not contract the illness from environmental factors.
  - Also called predictive disclosure
  - Sources: http://www.fcsm.gov/99papers/massell.pdf; http://stats.oecd.org/glossary/detail.asp?ID=6886.

Types of Disclosure (con't.)

- Inferential
  - Typically concerns aggregate data
  - Example: Area average home purchase price infers income

Types of Disclosure (con't.)

- Type I Disclosure: An intruder has knowledge that a given person (or organization) is included in the survey and the intruder attempts to find this record.
- Type II Disclosure: Occurs when an intruder does not know the identity ahead of time and uses externally available resources (linking databases) to attempt to find survey respondents.
- Source: http://www.icpsr.umich.edu/org/publications/bulletin/fall03.pdf.

Disclosure Review and Analysis

- Disclosure review
  - Examination of overall factors related to risk
  - Includes ICPSR's current confidentiality review
- Disclosure analysis is a 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 (e.g., SAMHDA DRC).
- Focus of initial training is disclosure review
- Later training will cover disclosure analysis

Disclosure Review and Analysis (con't.)

- Examine risk to confidentiality
  - Indirect identifiers (e.g., age, gender, race – taken together)
  - External files;
  - Focus is characteristics that can be known (demographics) not beliefs or attitudes
- 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

Other key terms and definitions

Privacy

- Has to do with personal choice or interest.
- According to Warren and Brandeis (1890) privacy is simply "the right to be left alone." More broadly, privacy can be thought of as describing conditions of limited accessibility to various aspects of an individual -- both physical and informational.
- Definitions of privacy include the capacity to be physically alone (solitude); to be free from physical interference, threat or unwanted touching (assault, battery); or to avoid being seen or overheard in particular contexts. Privacy also refers to the capacity to control when, how and to what degree information about oneself is communicated.
- A broad range of social and legal institutions set the terms of privacy, and societies lay out the boundaries very differently. New technologies can challenge old understandings of privacy.
- Citations:
  - Source: http://privacy.med.miami.edu/glossary/xd_privacy_basicdef.htm; extracted 10/17/08.
  - S. Warren and L. Brandeis, The Right to Privacy, Harvard Law Review, Vol. IV December 15, 1890 No. 5.
- An individual's interest in limiting who has access to personal health care information (Health Information Privacy and Accountability Act, HIPAA).

Confidentiality

- *Confidentiality* is closely related to privacy, but not identical. It refers to the obligations of individuals and institutions to use information under their control appropriately once it has been disclosed to them. One observes rules of confidentiality out of respect for, and to protect and preserve, the privacy of others.
- The ethical principal of autonomy or self determination requires respect for each individual's choices about uses and disclosures of his/her own information, as it does for privacy generally.
- Individual control must obviously be weighed against other goals achievable only by limits on privacy.
- Source: http://privacy.med.miami.edu/glossary/xd_privacy_basicdef.htm.

Anonymity / Anonymized [data]

- Previously identifiable data that have been de-identified and for which a code or other link no longer exists. An investigator would not be able to link anonymized information back to a specific individual (Source: http://healthcare.partners.org/phsirb/hipaaglos.htm).
- A record from which direct identifiers have been removed (Source: http://stats.oecd.org/glossary/detail.asp?ID=6883) .
- Some authors refer to anonymized data as those having undergone disclosure review / analysis. Others refer to anonymized data as only having direct identifiers removed. The same is true for “de-identified data.”

Usability Processing

- Steps may include
  - Add value and variable labels
  - Standardized missing values
  - Insert question text
  - Correct errors
  - Use “toolbox” for efficiency
  - Quality control checks
  - Utilize Hermes
- This is the stage at which you may notice disclosure risks
- Process of disclosure review / analysis is not linear

Disclosure Risk

- Risk is determined by careful review of the data
- Disclosure risk is addressed through
  - Assessment
    - Disclosure review
    - Disclosure analysis
  - When unacceptable risk is discovered
    - Applying disclosure protection techniques to alter the data
    - Using other mechanisms to protect respondents (e.g., secure online analysis, restricted-use agreements, data enclave)

Geography

- A key factor in identification
  - As geographic units become smaller, unique values on key variables become more visible
  - Reason sampling design is important
    - Were data selected to represent nation? States? Smaller units?
- When possible, remove references to geography
  - Thank you's to cooperating agencies, facilities, staff in study sites
  - Numeric coding that could reveal geography (records sorted by PSU)

Geography (con't.)

- *Guidelines*:
  - Ensure that geographic units are of sufficient size (Census uses between 100,000 and 250,000 population per unit)
  - Depends on sensitivity of data; higher thresholds may be used
  - To add uncertainty to potentially unique values in smaller geographic areas:
    - **Coarsen the dataset** (e.g., group response categories)
    - **Apply disclosure protection techniques** (e.g., swap records)
  - Review variables that provide geographic context to protect against inadvertent geographic disclosure (e.g., strata variables, PSUs, average local temperature, facilities or organizations with very large populations, which are likely to be located in the most densely populated areas, or that are identifiable by organization size alone)

Publications

- Review publications to determine whether publications
  - Reveal site names
  - Reveal sample sizes per site or other sample characteristic
  - Could narrow site identities
  - Could otherwise undo disclosure risk protections
- Ask PI if such data have been published

Indirect Identifiers: Outliers

- Values at either extreme of a distribution are most visible
  - Includes variables such as income, age, revenue, etc.
  - SAMHDA practice: for sensitive variables, flag records 2 or 3 standard deviations from the mean to ID potential problems
  - *Guidelines*
    - Top- & bottom-code continuous variables with outliers
    - Recode the entire distribution to be a categorical, ordinal, or interval variable

Indirect IDs: Unique Combinations

- Combination of multiple variables can lead to unanticipated isolated cases
- Called deductive disclosure
- 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

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

Disclosure protection (DP) methods: Examples of coarsening

- Change continuous variables to categories (*age* 12-90 is transformed to 5 year increments)
- Top- and bottom-code data (*age* 12-90 becomes 17 and under; continuous age 18-65; 66 and older)
- Combine categories (individually listed health conditions are combined into broader categories: *cocaine abuse*, *marijuana abuse*, and *methamphetamine abuse* are combined into *drug abuse*)
- Delete variables of low utility and high risk (*past year medical condition* has 95% missing data)

DP methods: Data swapping

- Unique, or at-risk, records are identified
- Random sample of at-risk records are swapped based on logical characteristic, such as geographic identifier
- Proportion of records swapped is held confidentially
- Unique records remain in the data, but it is uncertain from where they originated
- Advantages:
  - All records remain in the dataset
  - Good choice when geographic variables are present (and required)

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 is precluded by top- and bottom-coding or categorizing)

DP methods: Random data extract

- Randomly selecting a subset of records and removing them from the public-use release of the dataset.
- All records have an equal chance of being excluded from public-use file; therefore, re-identification could never be with certainty
- Advantages
  - Relatively simple procedure
  - Works well for studies with a large number of records, for which the extract will not significantly affect statistical precision or estimates.
  - It may be possible (depending on the characteristics of the study) that the data remaining in the public-use file can be less altered than with other DP methods.

DP methods: Blank (edit) & impute

- Selectively blank data for
  - Randomly selected at-risk records
  - Selected variables
- Impute missing data
- Apply imputed data to blanked cells
- Advantages
  - Minimizes distortion on inter-relatedness of variables and records
  - Can be used when small set of records require disclosure protection (perhaps not enough records for data swapping)
  - Program at ISR uses multivariate sequential regression approach to imputing item missing values, or in this case, values randomly selected for editing

Release Summary of Procedures

- Important to summarize what has been changed from the original data
  - Replication
  - Variances
- Yet, do not compromise disclosure protection procedures

Considerations – The Data

- What are intended analytic uses of the file?
  - Data type (survey, administrative)?
  - Disclosure problem (unique records, matching files)?
  - Solutions must match the problems and important uses
- How old are the data?
- How much missing data?
- How much data for each person or organization (longitudinal files generally more risky than cross-sectional)?
- How detailed are the data? How sensitive (opinions v behavior)?
- How much of the data are subject to recall, measurement error?
- What in the data can be observed or discovered (ethnicity v political party, police record v illegal activities)?
- Can the sampling frame be readily identified?
- How will information lost impact interpretation (e.g., loss of variance)?
- Can lost information be released in tabular or other form?

Considerations – Personal

- Must get comfortable with the fact that data will be altered
- Personal factors: experience releasing data, control?
- Review disclosure literature or have assistance of someone knowledgeable about disclosure issues
- Try to decide up front what you will release; think through implications:
  - Once something is made public there is no going back
  - Releasing additional variables after a public release could compromise the disclosure plan for the initial public release
  - Informed consent statements should match your actual intent regarding data release

Considerations – Team

- Must know the content area of the data – not only a statistical or programming procedure
- People with various expertise helpful (for disclosure analysis)
  - Content
  - Statistical
  - Programming
  - Disclosure
  - Intended secondary users or those familiar with secondary data use (and the challenges)