Attention to Confidentiality in CAI Studies

Tom Krenzke, Westat

March 19, 2014

FedCASIC Conference, Washington D.C.

Objectives

- To provide some insights on, and attention to confidentiality in CAI studies
- To share some experiences from
  - The viewpoint of a sampling statistician
    - Statistical disclosure control (SDC)
    - Data dissemination
  - A recent Institutional Review Board (IRB) member

Motivation for Confidentiality

- Ethical principals, guidelines, rules, laws
  - Belmont Report (1979)
    - Ethical principals and guidelines for protecting human subjects
  - Common Rule (1991)
    - Provisions for IRBs, informed consent, assurances of compliance
  - Several other laws, include
    - Privacy Act of 1974 (Section 552a), Office of Management and Budget (OMB, 1997), Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)
    - HIPAA for patient privacy protections (OCR, 2012)

Motivation for Confidentiality (2)

- What if there is a breach?
  - Trust and response rates may plummet
  - Harm
- Risk scenarios
  - Prosecutor – Looking for a specific person (El Emam et al. , 2009)
  - Journalist – Not looking for a specific person, just breaking a story (El Emam et al. , 2009)
  - Graduate students

Motivation for Confidentiality (3)

- Examples of breaches
  - Group Insurance Commission (Sweeney, 2002)
  - Netflix (Jiang, X. et al., 2013)
  - AOL (Jiang, X. et al., 2013)
  - Target – December 2013
  - Several universities in past year – Southern Maryland Gazette March 1, 2014 article
  - Laptops

Select General Risk Factors

- Modes and access levels of dissemination
  - Public use file (PUF)
  - Restricted use file (RUF)
  - Remote access to RUF (e.g., NCHS)
    - Agency analysts review output, and provide results
  - Real-time on-line analytic system (OAS) from a RUF
  - OAS from a PUF
    - Census Bureau’s DataFerrett
  - OAS from static tables
    - Census Bureau’s American FactFinder static tables
    - Tables in reports

Select General Risk Factors (2)

- Sampling fraction
  - How likely are sample uniques actually population uniques?
- Sensitive questions
  - Attracts attention and curiosity
  - Can a breach potentially harm the respondent?
    - Suicidal thoughts
    - Crimes
    - Sexual abuse
    - Income and taxes
  - Are such questions necessary? Sometimes added by one of several researchers and not the main focus of the study

Disclosure Risks Within Dataset

- Personal Identifying Information (PII)
  - HIPAA Rule
    - Safe Harbor
      - 18 data elements (names, addresses, dates, etc)
      - Safe?
    - Statistical expert review
  - General safeguard
    - Dis-associate PII from data in transport or even storage

Disclosure Risks Within Dataset (2)

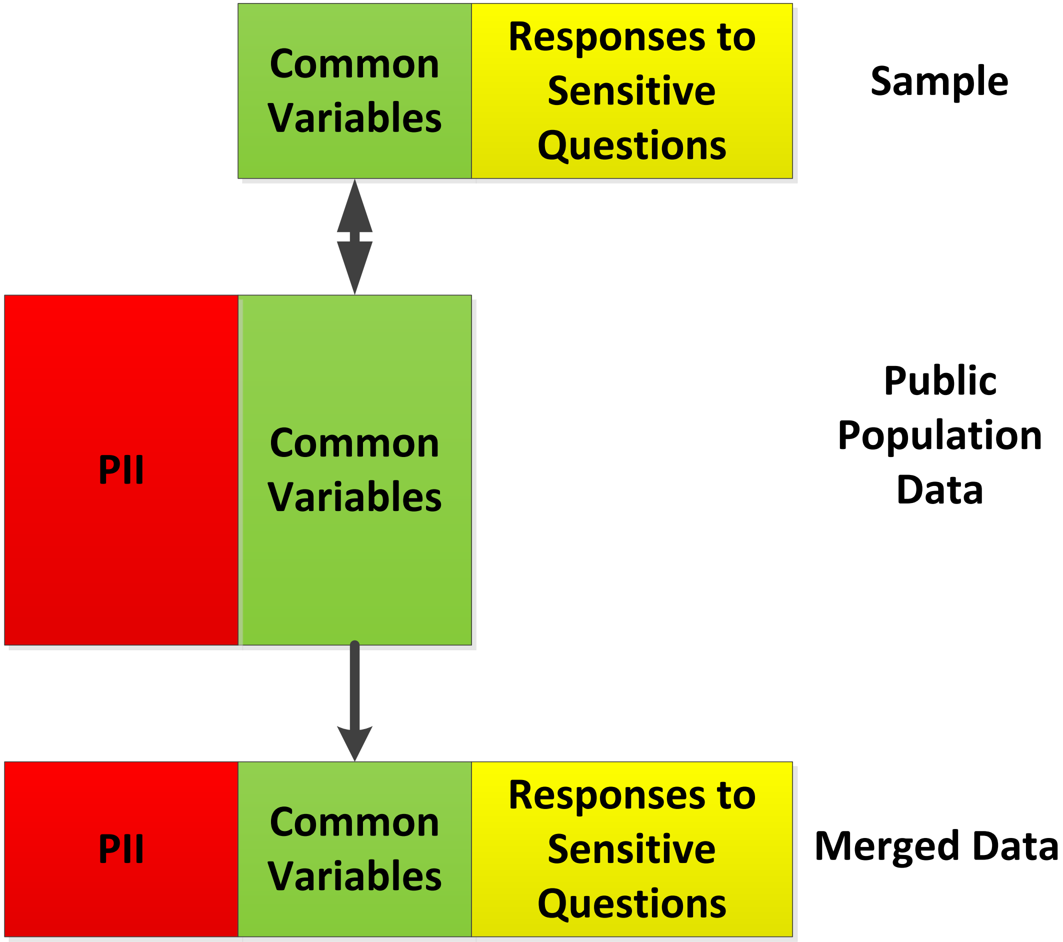
- Combinations of indirect identifiers
  - Sample design and weighting variables
  - Demographics
    - Examples: Age, race, sex, education attainment, employment, income
  - Geography (residence, workplace)
    - Examples: Region, state, country, sub-county
  - Contextual variables
    - Unemployment rate in small area
  - Outliers (continuous variables, spatial)
  - Open ended questions (look for names, locations, occupations)

Disclosure Risks Within Dataset (3)

- Combinations of indirect identifiers (continued)
  - Recommended practice
    - Estimate the risk
      - Exhaustive tabulations identify sample uniques or sparse combinations of variables
        - k-anonymity (Sweeney, 2002)
      - Special Unique Detector Algorithm (SUDA) (Elliot, 2002)
      - Re-identification risk
        - Log-linear models (Skinner and Shlomo, 2008)
        - Mu-Argus – sampling weights incorporated
    - Risk measures can also be used…
      - To re-assess their current confidentiality rules
      - To set risk thresholds for their studies

Disclosure Risks External to Dataset

- Publically available data
  - Various on-line lists



Disclosure Risks External to Dataset (2)

- Publically available data (continued)
  - Recommended practice
    - Risk measures (prior slides)
    - Record linkage
      - Exact matching and Statistical matching
      - Summary in Winkler (1993)
        - » https://www.census.gov/srd/papers/pdf/rr93-8.pdf
      - Diniz da Silva, et al. (2010), evaluation of…
        - » Link Plus (CDC), RELAIS (ISTAT), FEBRL (Australian National University and the New South Wales Dept of Health), Others
      - CDC's FRIL

Disclosure Control Treatments

- Data coarsening
  - Recodes
    - Categories – Combine categories
    - Continuous variables -- specified categories
      - Top-codes
  - Variable suppression
    - Open-ended items
    - Items with 2 categories where one is sparse
- After coarsening, rerun risk assessment
- If risks remain, consider further SDC treatments
  - E.g., American Community Survey Public Use File
  - Random perturbation
  - Subsampling

Disclosure Risks in CAI Studies

- Knowledge of sample inclusion
  - Knowing whether or not a person is part of the survey has a large increasing effect on risk
    - Parent, Caretaker
    - Prison guard
  - Possible protections
    - Try to keep out of the room during interview
    - Dis-associate parent data from youth data for PUF
    - Subsample
    - Assign different booklets containing different subject matter to inmates

Disclosure Risks in CAI Studies (2)

- People within hearing range of interview
  - Answer phone, hand off phone, listen
    - Survey about risky behavior, crime
  - Possible protections
    - Be vague initially until discussing with the respondent
    - Use multiple choice responses
    - Use touch-tone dial responses
    - Audio Computer Assisted Self-Interview (ACASI)

Disclosure Risks in CAI Studies (3)

- Organizers of focus group
  - Knowledge of sensitive subject matter
  - Watch people walk into building or room
  - May be associated with them in some way (colleagues)
  - Possible protection
    - Registration desk manned by someone without knowledge of subject matter
    - Limit access to data -- data for internal use only

Disclosure Risks in CAI Studies (4)

- In-person interviews
  - Interviewer leaves help-line card
  - Spouse or roommate can pick it up and investigate
  - Advanced letters
  - Possible safeguard
    - Only hand over card if deemed useful (some distress)
    - Be somewhat vague in advanced letters
- Visible devices
  - Worn by respondents
  - May convey information about survey inclusion
  - Possible protection
    - Coarsen data
    - Limit access

Disclosure Risks in CAI Studies (5)

- Consent form
  - ‘Research staff that work with your sample will never **know** your name or any other personal information.’
  - Technically, as originally stated, it may not be true that the researcher may never know a person’s name, since combinations of variables provided could identify a unique person that is known to the researcher.
  - ‘Research staff that work with your sample will never **be given** your name and any other personal information.’

Disclosure Risks in CAI Studies (6)

- Call centers, help desks, interviewer notes
  - How are notes taken?
  - What PII are on the notes?
  - What happens to the notes and how is the information conveyed?
  - Possible safeguards
    - Electronic information only
    - Transfer in secure manner
    - Cover in data security plans
- Paradata, interviewer IDs
  - Be conscientious of location information
    - GPS coordinates
    - Location embedded in IDs

Summary

- Balance between disclosure risk and data utility
  - Heightened with sensitive data
- Safeguards need to be addressed throughout the spectrum of the CAI process
  - Planning
    - IRB approval
    - Data security plan
  - Implementation
    - Following protocols
    - Various scenarios, and combinations of variables

Summary (2)

- Key safeguards
  - Measure risks of combinations of factual pieces of information in the data
  - Try to protect against other knowing about the person being in the sample
- Data dissemination
  - Some basic important questions:
    - Who owns the data?
    - Where does the data go in the end?
    - What does the file have on it?
    - How are results published?
    - How many are in the population and what is the sample size?
  - Apply SDC treatments