# De-identifying research data

Erick Baumgartner  Vrinda Kapoor

30 June, 2021

# Overview

- Research team should always work with and analyze de-identified data, except when
  - planning follow-up survey
  - monitoring in-coming data
- Publicly released data should be de-identified
- To de-identify data:
  - Remove or drop identifiers (PII)
  - De-identify necessary PII variables by masking, encoding, and anonymizing

# Personally Identifiable Information (PII)

- Variables that can, either on their own or in combination with other variables, be used to identify a single surveyed individual with reasonable certainty
- Any information that can be used to link survey data with respondents

# Personally Identifiable Information (PII)

Direct identifiers

- Name of survey respondent, household members
- GPS coordinates
- Biometrics
- Record identifier (SSN, IP address, bank account, medical record number)
- Pictures of houses or individuals

# Personally Identifiable Information (PII)

Indirect identifiers: Combination of:

- Age
- Gender
- Caste/Ethnicity
- Grades, salary, job position
- Physical attributes, disability, medical condition
- Any outliers (number of children) + location

63% of the US population is uniquely identified by gender + date of birth + zip code!

# Why remove PII?

- You are handling someone's confidential (often sensitive) information
- Respect for respondents
- Informed consent is an agreement which states that we keep data confidential
- Required by ethics review boards (IRBs)
- Often legally required (HIPAA, GDPR, etc.)

# Who can access PII?

- Add your name to the project IRB
- Possess human subjects certification
  - CITI
  - NIH

# How to remove PII

- Drop all PII variables not necessary for analysis
- Keep it in another encrypted folder, which can be accessed if you need it for monitoring, back-checks, etc.

# How to remove PII

- De-identify all PII necessary for analysis by
  - Encoding by dropping the value label
  - Masking by limiting disclosure of continuous PII variables needed for analysis
  - Categorization: making categories of cont. variables
  - Rounding, top-coding
  - Adding noise by adding a variable with zero mean and positive variance
- Anonymize secondary and administrative dataset by creating new ID

# How to remove PII

- Document all the changes you make to de-identify the PII
- If creating new ID, keep the crosswalk encrypted and safe

# How to remove PII

- Manually look for identifiers
- Use *pii_scan* by JPAL; output is all flagged variables
  - https://github.com/J-PAL/stata_PII_scan
- Flag questions when designing the instrument

# When to remove PII?

- As early as possible!
- Saves time and effort later
- Opportunity to ask while collecting data: Is this really needed?

# When to remove PII?

- Data must be de-identified before publishing on microdata catalogue or made public
- Be conservative
- If PII is needed for analysis, then give restricted access for replication

# Disclosure risk

- Risks that data could be re-identfied
- Think about trade-off between disclosure risk and information loss
- How difficult is it to re-identify data?
- What harm could re-identification cause?
  - Illegal activity
  - Data from conflict zones
  - Political activity, vote-buying during elections
  - Voting behavior
  - Medical records, blood samples
  - Financial records

# Disclosure risk

- Can use sdcMicro, package in R
- Used for generating anonymized microdata
- Practical guide by IHSN: https://sdcpractice.readthedocs.io/en/latest/

Thank you!