NCHS Disclosure Concerns
In Our Modern World

Donna Miller, MS
Confidentiality Officer
National Center for Health Statistics
National Center for Health Statistics (NCHS)

- Collects data from birth and death records, medical records, interview surveys, direct physical examinations, laboratory testing
- Release data: public use files, on-line systems
- Survey, vital statistics data are covered under confidentiality assurances
Balancing Confidentiality and Data Access to Protect NCHS Data Products

- **NCHS Disclosure Review Board**
  - Formally reviews, approves proposed NCHS public use files
  - Files deemed too sensitive to publically release are made available in the Research Data Centers.
NCHS Disclosure Concerns

- **External Data**
  - Foundations for Evidence-Based Policymaking Act of 2018, Title II, Open Government Data Act
  - Proliferation of commercial data collections
    - Commercial Data Brokers
    - Social Media

- **Advances in Technology**
  - Hardware
  - Software
Foundations for Evidence-Based Policymaking Act of 2018, Title II, Open Government Data Act

- Default state of new and modernized Federal information resources shall be open and machine readable
- More government data available to the public

[Data.gov website screenshot with search bar and message: 232,195 datasets found]
How Do We Prevent Disclosure of NCHS Respondents and Establishments: Open Government Data Act?

- More open data = more linking opportunities
- Future research is needed to determine the types of data Federal agencies will publicly release.
  - What impact will this have on the disclosure risks of NCHS survey respondents?

**NCHS PUF**
- Age
- Sex
- Race/Ethnicity
- Education
- Marital status
- Region
- ADLs
- Type of Insurance
- Veteran Status
  - Active duty time
  - frames

**Standardize Like Variables**

**VA Admin. Data**
- Age
- Sex
- Race/Ethnicity
- Education
- Marital status
- Region
- ADLs
- Type of Insurance
- Veteran Status
  - Active duty time
  - frames

**Non-Matches**

**Potential Matches**

**True Matches**

**Non Matches**

**True Matches**
Proliferation of Commercial Data Collections

- **Commercial Data Brokers**
  - Companies collecting consumers’ personal information, via online and offline sources, and resell or share that information with others
    - Types of Consumer Data
      - Transactional data sales & customer lists
      - Loyalty card data (retailers)
      - Non-profit organizations’ member or donor lists
    - Little legal oversight

- **Social Media**
How Do We Prevent Disclosure of NCHS Respondents and Establishments: Proliferation of Commercial Data Collections

- Future research is needed to determine if/how commercial data collections could be used to reidentify NCHS survey respondents
  - Commercial Data Brokers
    - Contents of commercial files not fully known
    - Can Brokers augment NCHS survey data with data they collect?
  - Social Media
    - Cambridge Analytica copy cats
      - Numerous “How To” articles describing how to extract data on internet
    - Share current locations via photographs
Advances in Technology

- **Hardware**
  - Unlimited, cheap storage space
  - More computing power

- **Software**
  - Machine learning, artificial intelligence
    - Process of learning starts with observational data
    - Look for patterns in the data
    - Build models to generalize
    - Mimic human intelligence, cognitive abilities
  - Web scraping
    - Various methods to collected data from internet
How Do We Prevent Disclosure of NCHS Respondents and Establishments: Advancing Technology?

- Future research is needed on how to safely make data publicly available while at the same time reducing the risk of disclosure.

- **CDC WONDER**, ad-hoc query system for the analysis of public health data
  - 2016 Obstetrics and Gynecology article violated WONDER suppression rules, reported cell sizes <10 at subnational level, described how they obtained small cell sizes

- **NCHS public use data**
  - 2018 journal article used machine learning techniques to “re-identify” respondents
Final Thoughts

- La Tanya Sweeney and others identified threats for reidentification.
- Deidentification methods advanced to address the threats.
- Now we need to determine what new threats pose the greatest disclosure risks:
  - Adjust our deidentification methods to protect the confidentiality of our survey respondents.
- Resources are needed:
  - Funds
  - Subject Matter Experts
  - Training current staff.
Thank You

For more information contact:
Donna Miller, MS
Email: DMMiller1@cdc.gov
Phone: 301-458-4135

National Center for Health Statistics
Centers for Disease Control and Prevention