Disclosure Avoidance Issues at NCHS

Jennifer Madans, Ph.D.
Associate Director for Science, NCHS

Joint NSF-Census-IRS Workshop on
Synthetic Data and Confidentiality Protection
July 31, 2009

Overview

- NCHS mission
- NCHS disclosure avoidance practices
- Future considerations for synthetic data at NCHS

NCHS Mission

- To provide statistical information that will guide actions and policies to improve the health of the United States population

- To release data in a timely manner and make available on as wide a basis as practicable while protecting confidentiality

NCHS Balancing Acts

- Meeting the needs of user community
- Safeguarding the confidentiality of survey participants
  - Changes in computing, technology, and availability of other public information will further restrict NCHS public-use data dissemination in the future
- Resource allocation
  - Data collection versus data dissemination
  - Investments must be proportional to impact

NCHS Disclosure Avoidance Practices

- Conduct re-identification risk assessments, e.g. DRB review
- Apply standard SDC methods for public-use files
  - Remove direct identifiers, rounding, top-coding, etc.
- Most data collected released as public-use files
  - NCHS has not yet developed and released synthetic public use data sets

NCHS Disclosure Avoidance Practices

- Restricted data made available through RDC
  - Detailed geographic information
  - Data obtained through linked administrative records
  - Genetic data
  - Other detailed socio-demographic information, e.g. extreme age, income, etc.
- Some experience with perturbing a small amount of information to release otherwise restricted data as public-use files

NCHS Perturbed Files

- Goal: to release public-use NCHS Linked Mortality Files
  - Selected key mortality variables for release
  - Assessed re-identification risk
  - For select records at risk, perturbed date or cause of death
  - Compared analytic utility of the public-use file to the restricted-use file
  - Released public-use file

What would make synthetic data sets more attractive to NCHS ?

- Having "buy-in" from skeptical user community
  - Develop low burden, non-'anecdotal' methods to demonstrate data integrity
- Demonstrated utility for resource investment
  - Creation and maintenance needs to be cost efficient
  - Technical assistance to users not resource intensive
  - Risk of disclosure must be lower than current methods
- Statistical agency collaboration and/or academic partnerships