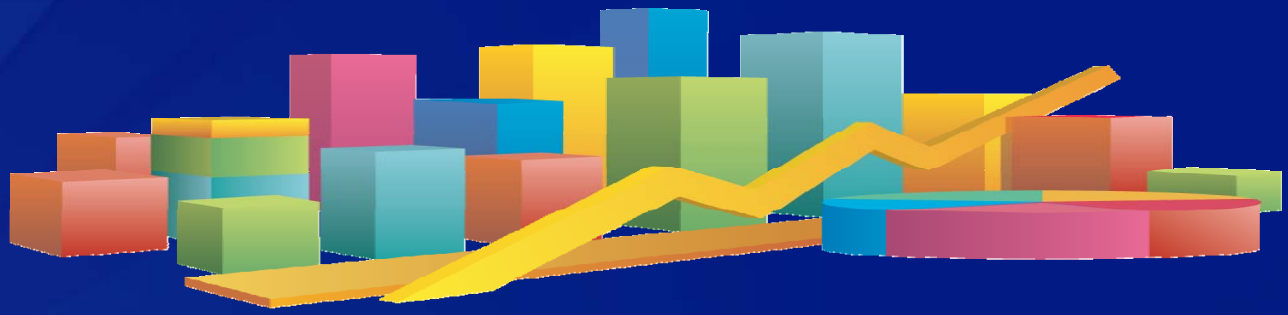
Why Can't You Just Give Me the Data?

Confidentiality, Access and Dissemination

Eve Powell-Griner, Ph.D.
Confidentiality Officer
National Center for Health Statistics



Overview

- Data distribution/use and protection requirements
- Issues in data granularity
- Main Steps In Producing Non-Disclosive Statistics

Data Distribution and Protection Requirements

- **Data are collected to be used**
  - Distribution in some form to particular parties is assumed from the outset – but in what form and to whom?
- **Distribution often limited by:**
  - Federal Law (Privacy Act; Public Health Service Act; HIPAA; etc.)
  - State law
  - Policies or ownership
  - Promises made to the person/institution providing the information (how will it be used, who will use it, how it will be protected)

NCHS: Two Competing Mandates

- **Wide dissemination of the data**
  - Public Health Service Act of 1956
- **Assurance of confidentiality**
  - Privacy Act of 1974
  - Public Health Service Act, section 308(d)
  - Confidential Information Protection and Statistical Efficiency Act (CIPSEA)
  - Policies/laws of data “owners”
    - States own vital statistics data
    - CMS owns administrative files (Medicare, Medicaid) linked to NCHS surveys

Confidentiality Requirements = Distribution Limits I

- **Requirements**
  - Use: only for purposes collected (statistical)
  - Data cannot be released in identifiable form
    - No person or institution providing information can be identified

Confidentiality Requirements = Distribution Limits II

- **Distribution limits**
  - Parties who may have access to detailed information
  - Limits on information that may be accessed
    - Direct identifiers
      - e.g., name, address, SSN, Medicare Number
    - Indirect identifiers-- the building blocks for creating a “unique” identity
      - Demographic (age, sex, race, marital status)
      - Socio-economic (occupation, education)
      - Other (date of event, rare health condition/cause of death, area characteristics, program participation, institution’s characteristics)

Confidentiality Requirements = Distribution Limits III

- **Preliminary protection considerations directing constraints imposed**
  - Are data unique to this file
  - Is collection process replicable
  - Are there external files with comparable data
    - Voter registrations lists, health record files, birth, death, marriage records, commercial data bases, newspaper accounts, personal knowledge, occupational licensing registries, property records, driver's license record, state and federal files, administrative data, information vendor files
  - Are there unique cases on key variables and combinations (e.g. outlier such as 20 yr. old with Ph.D.)
  - Have data been "enriched" –e.g., external data have been added

Confidentiality Requirements = Distribution Limits IV

- **Resulting constraints on publicly disseminated data**
  - No direct identifiers
  - Broader coding structures of key indirect identifiers
    - Restrict geography (e.g., show only region; state)
    - Recode to remove detail (e.g., top/bottom coding)
    - Case suppression (e.g., remove a record)
    - Variable suppression (e.g., remove a data item)

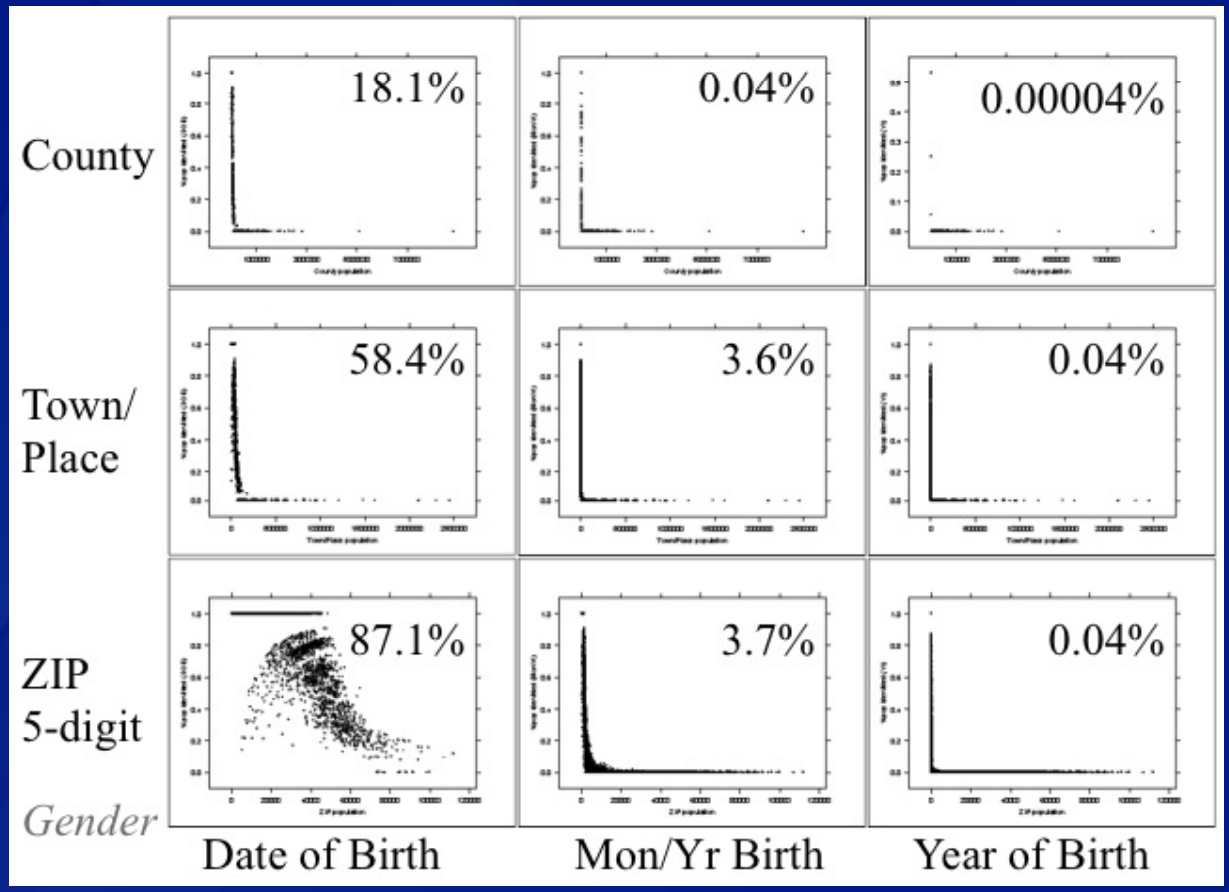
Issues in Granularity

- **The smaller the population, the more easily an individual or institution can be unique**
  - Rule of thumb: 100,000 inhabitants
  - HIPAA an exception (20,000 3-digit zip)
- **The greater the number of variables pertaining to an individual, the greater the likelihood combination of them increases identifiability**

Examples

- **Administrative Records**
  - Birth date, sex, zip code from the Cambridge voter registration list (about 55,000 voters) matched to information with the health data contained in anonymous hospital discharge records identified MA Governor (1997)
- **Vital Statistics**
  - Using age, race, state of death and cause of death, can identify decedent with rare, but public cause of death, using web search engines (e.g. white woman age 85 dying of homicide in a state with a small population) (2012)
- **Genome Project**
  - 42% of a sample of anonymous participants in a high-profile DNA study re-identified by using zip code, date of birth and gender from DNA study plus voter registration rolls. (2013)

Data Privacy Lab: Birthdate, Gender, Zip Code As Re-Identifying Keys

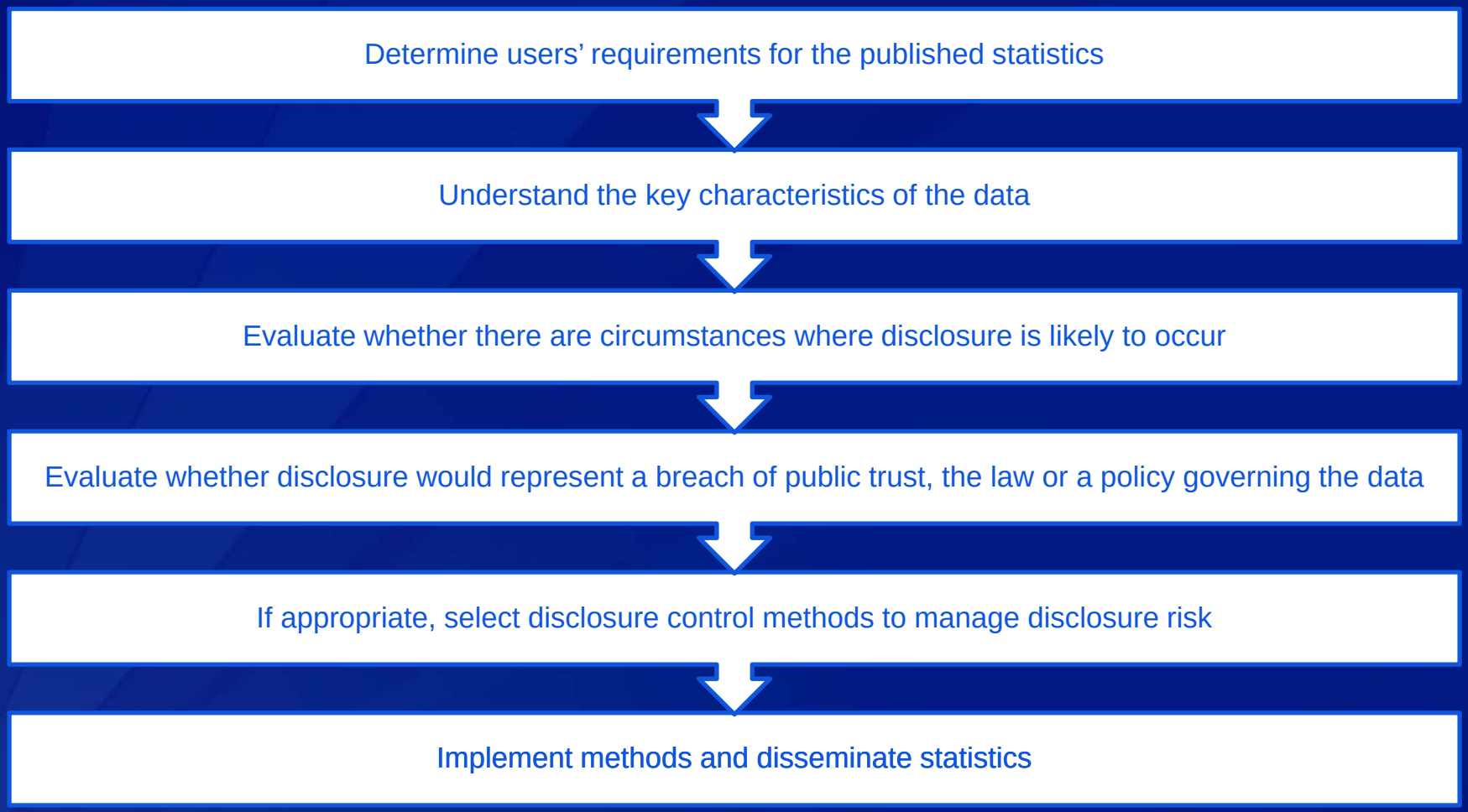


L. Sweeney, Harvard Department of Government: http://aboutmyinfo.net/about.html

Possible Strategies for Wider Dissemination

- More detail for selected key variables and less for others
- Use rates rather than counts
- Use controlled rounding
- For mapping, use ranges rather than counts or rates
- Reduce number of tables that can be linked
- Use spatial and/or temporal aggregation
- Use suppression (complementary stronger than primary)

Main Steps For Ensuring Access to Non-Disclosive Statistics



Thank You

For more information contact:
Eve Powell-Griner, Ph.D.
Email: EPowell-Griner@cdc.gov
Phone: 301-458-4601

National Center for Health Statistics
Centers for Disease Control and Prevention