Database Theory and Applications for
Biomedical Research and Practice

BMIN 502 / EPID 635
Week 9

Introduction to secondary data sources

Institute for
Biomedical Informatics

Sources of data for clinical research

- Primary
  - » Surveys
  - » Outbreak investigations
  - » Ongoing research studies
  - » Registries
- Secondary
  - » Medical records
  - » Surveys
  - » Registries
  - » The US Census
  - » Insurance claims

Using existing data

- Three approaches
  - » Secondary data analysis
  - » Ancillary studies
  - » Systematic reviews
- Advantages
  - » (Relatively) fast data collection
  - » (Relatively) cheap
- Disadvantages
  - » The main one: Possibly poor quality
    - – No control over *how* data were collected
    - – No control over *what* data were collected
    - – No control over *when* data were collected

Types of datasets for secondary analysis

- Individual datasets
  - » Prior research studies
  - » Publicly available datasets
  - » Registries
- Aggregate datasets
  - » Datasets on groups of observation units
  - » Useful for ecologic studies

Using secondary data:

Discovering a research question to fit the data

- Identify an appropriate database
- Learn all you can about the data
  - » Primary documentation: Data dictionaries, use documents, descriptions by source agency
  - » Secondary documentation: papers, communication with prior users/user groups
- Identify variables of interest for your research question
- Review the literature and consult with experts
- Get the data into the appropriate platform
- Perform descriptive analyses and formulate hypotheses
- Analyze!

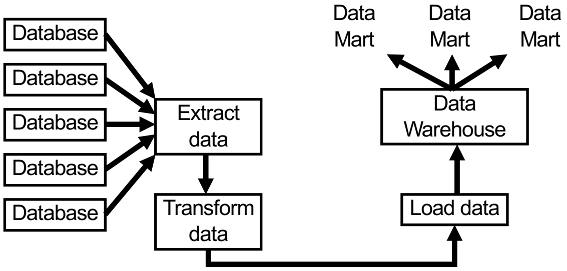
How are secondary data packaged?

- Aggregated reports
- Flat text files
- Spreadsheets
- Datasets
- Databases
- Data warehouses

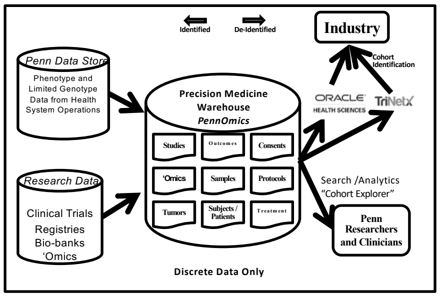
The data warehouse

- A centralized resource for long-term data storage
- Supports the activities of entire organizations
- Takes input from distributed databases on scheduled batch basis
- Provides access to large-scale, temporal data
- Provides a platform for decision support

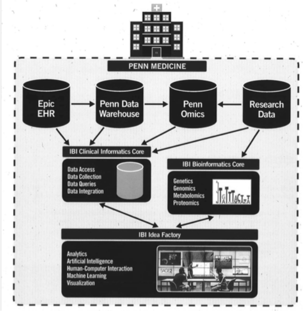
How does a warehouse work?



A local example: The Penn Research Data Warehouse



An integrated informatics ecosystem



PENN MEDICINE

Epic EHR

Penn Data Warehouse

Penn Omics

Research Data

IBI Clinical Informatics Core

Data Access
Data Collection
Data Queries
Data Integration

IBI Bioinformatics Core

Genetics
Genomics
Metabolomics
Proteomics

IBI Idea Factory

Analytics
Artificial Intelligence
Human-Computer Interaction
Machine Learning
Visualization

Some secondary data resources worth considering

Bureau of the Census
National Center for Health Statistics
National Cancer Institute
Agency for Healthcare Research and Quality

The Census: First things first…

- What is a census?
  - » Enumeration, not a sample
- Characteristics of the US Census
  - » Performed every 10 years
  - » Census (everyone)
    - – Short form
  - » Sample
    - – Long form

American Community Survey

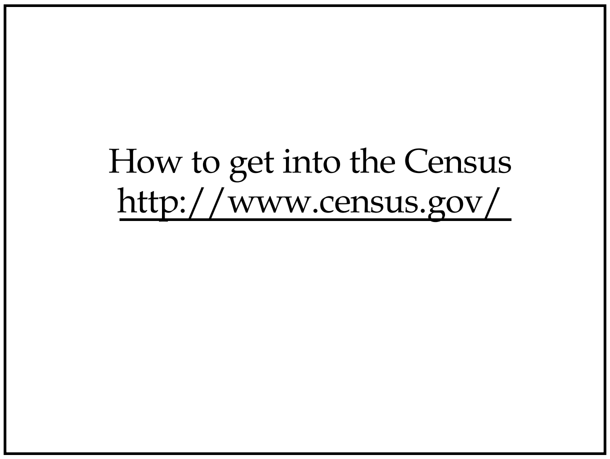
- Collects data on ~3M American households yearly
- Sample drawn from every US county
- Supports “critical government functions”
- Will eliminate the decennial long form

Population Estimates Program

- Population estimates between censuses
- Four waves each year:
  - » Winter - The United States and States
  - » Spring - Counties
  - » Summer - Cities and Towns
  - » Fall - Metropolitan and Micropolitan Statistical Areas
- Estimates refer to population as of July 1 of previous year

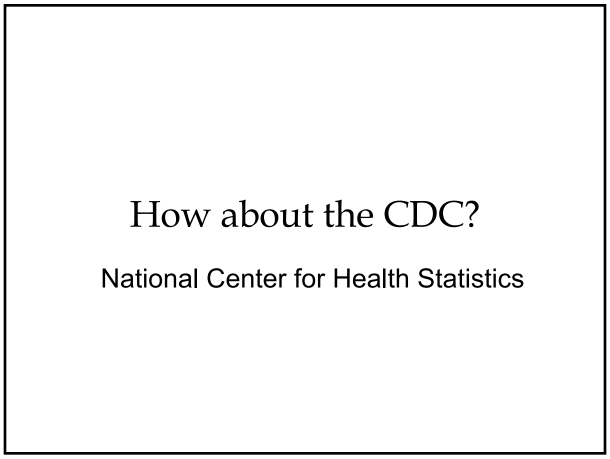
Other surveys at census.gov

- Survey of Income and Program Participation (SIPP)
  - » Evaluation of government programs
- Current Population Survey (CPS)
  - » 50K households/month
  - » Development of government programs
- American Housing Survey (AHS)
  - » Biennial sample of ~55K households (repetitive)
  - » Data on housing characteristics and household flow
- And numerous others for other government agencies…



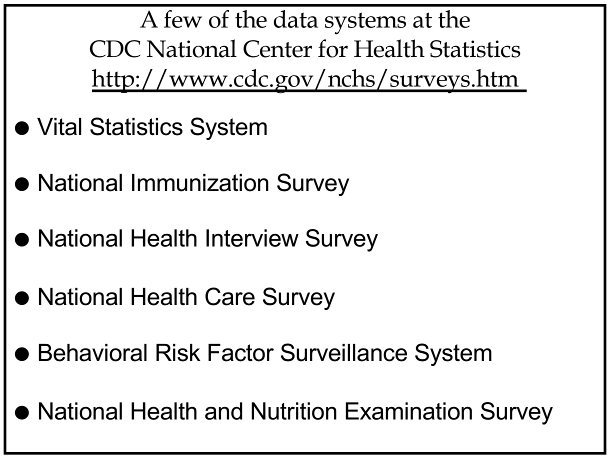
How to get into the Census

http://www.census.gov/



How about the CDC?

National Center for Health Statistics



A few of the data systems at the
CDC National Center for Health Statistics
http://www.cdc.gov/nchs/surveys.htm

- Vital Statistics System
- National Immunization Survey
- National Health Interview Survey
- National Health Care Survey
- Behavioral Risk Factor Surveillance System
- National Health and Nutrition Examination Survey

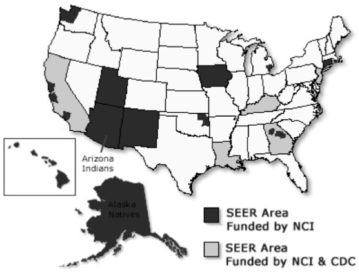
National Cancer Institute

Surveillance, Epidemiology, and End Results (SEER) Program

- Authoritative source of information on cancer incidence and survival in the United States
- Started in 1973
- Only source of population-based historical as well as current information on patient survival and stage of disease
- Data from 20 cancer registries covering about 26 percent of the US population

SEER registry coverage

http://seer.cancer.gov/registries/



SEER data

- Patient demographics
- Primary tumor site
- Tumor morphology and stage at diagnosis
- First course of treatment
- Follow-up for vital status
- Full data documentation is at: http://seer.cancer.gov/data/documentation.html

SEER public use data

http://seer.cancer.gov/resources/

- Available for 1973-2010
- Requires signed data use agreement
  - » Renewable annually
- Two forms
  - » Online (SEER*Stat)
  - » Downloadable or CD with/without SEER*Stat

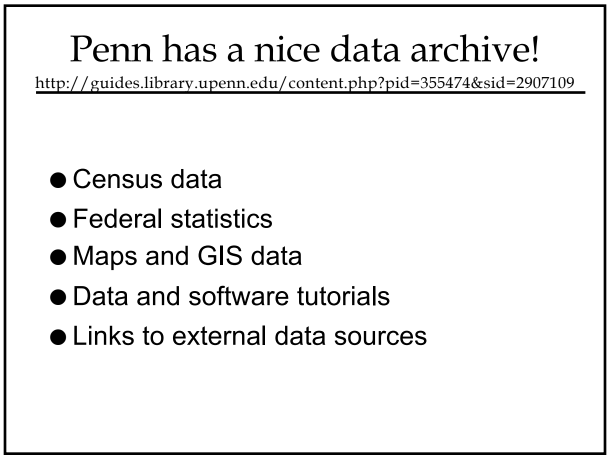
Agency for Healthcare Research and Quality

Healthcare Cost and Utilization Project (HCUP)

HCUP Databases

http://www.ahrq.gov/research/data/hcup/index.html

- **The Nationwide Inpatient Sample (NIS)**
  - » Inpatient data from a national sample of over 1,000 hospitals
- **Kids' Inpatient Database (KID)**
  - » Nationwide sample of pediatric inpatient discharges
- **State Inpatient Databases (SID)**
  - » Universe of inpatient discharge abstracts from participating states
- **The State Ambulatory Surgery Databases (SASD)**
  - » Data from ambulatory care encounters from hospital-affiliated and sometimes freestanding ambulatory surgery sites
- **State Emergency Department Databases (SEDD)**
  - » Data from hospital-affiliated emergency departments for visits that do not result in hospitalizations.



Penn has a nice data archive!

http://guides.library.upenn.edu/content.php?pid=355474&sid=2907109

- Census data
- Federal statistics
- Maps and GIS data
- Data and software tutorials
- Links to external data sources