Objectives of the Session

- Define secondary data
- Identify and discuss key conceptual and practical issues involved in secondary data analysis
- Identify the types and sources of secondary data
- Recognize the strength and limitations of secondary data
- Use online tools for locating and learning about publicly available datasets relevant to your research
- Apply study approach to their research
## Summary of Study Approaches

<table>
<thead>
<tr>
<th>Study Approach</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review/meta-analysis</td>
<td>Synthesize existing knowledge</td>
</tr>
<tr>
<td>Correlational (ecological)</td>
<td>Compare average levels of exposure and disease in several populations</td>
</tr>
<tr>
<td>Case series</td>
<td>Describe a group of individuals with a disease</td>
</tr>
<tr>
<td>Cross-sectional survey</td>
<td>Describe exposure and/or disease status in a population</td>
</tr>
<tr>
<td>Case-control study</td>
<td>Compare exposure histories in people with disease (cases) and people without diseases (controls)</td>
</tr>
<tr>
<td>Cohort study</td>
<td>Compare rates of new (incident) disease in people with different exposure histories or follow a population forward in time to look for incident diseases</td>
</tr>
<tr>
<td>Experimental study</td>
<td>Compare outcomes in participants assigned to an intervention or control group</td>
</tr>
<tr>
<td>Qualitative study</td>
<td>Seek to understand how individuals and communities perceive and make sense of the world and their experiences</td>
</tr>
</tbody>
</table>
Definition of Secondary Data

• Primary or Secondary data?
  – Depends on a relationship between who collect the data and who analyze it
  – Data collected by researchers for specific purpose is primary data
  – Data collected by someone else for other purpose is secondary data

Key Conceptual Issues

• Someone else’s data

• Treat data and research plan with same rigor as would for a primary data collection study

• Research questions should be conceptually driven, interesting a priori

• Know data as well as if you had collected it yourself
  – Who is in the cohort?
  – Strengths and limitations of data collection procedures, instruments
Secondary Data – Use

• Hypothesis generation/testing
• Publications
• Assess the health of a community or population;
• Search for possible causes of disease, injury and disability;
• Plan programs to meet community needs;
• Measure progress in prevention and control efforts
• Pilot data for grant (e.g., R01) proposals

Selecting a Database

• Compatibility with research question(s)
• Availability and expense
• Sample: representativeness, power
• Measures of interest present and valid
• Missing data and missing information
• Local expertise
Locating appropriate secondary data

• Procedures
  – Define your research question
  – Specify the study population
  – Specify other variables you would like to include in the analysis
  – Specify what kind of data is most appropriate for your research question (e.g., national survey, hospital record)
  – Create a list of data sets that include information related to your research question and examine them
  – Once you choose data set, examine the variables carefully (missing data, data management, codes, ...)
Types of Secondary data

- Survey (NHIS, NHANES, HRS, BRFSS)
- Administrative (Medicare claims)
- Discharge (HCUP SID and NIS)
- Medical chart / EMR
- Disease registries (SEER)
- Aggregate (US Census)
- Combinations and linkages

Data Sources

- National Center for Health Statistics (NCHS) http://www.cdc.gov/nchs/surveys.htm
- Center for Medicare and Medicaid Services (CMS) http://www.cms.hhs.gov/home/rsds.asp
Data Sources (cont.)

Examples of downloadable data from NCHS:
- National Health and Nutrition Examination Survey (NHANES)
- National Ambulatory Medical Care Survey (NAMCS)
- National Hospital Ambulatory Medical Care Survey (NHAMCS)
- National Hospital Discharge Survey (NHDS)
- National Home and Hospice Care Survey (NHHCS)
- National Nursing Home Survey (NNHS)
- National Survey of Ambulatory Surgery (NSAS)
- National Employer Health Insurance Survey (NEHIS)
- National Vital Statistics System (NVSS)
- National Health Interview Survey (NHIS)

Evaluation of Data Sources

- Purpose of the study
- Sponsor/collector of the data
- Mode of data collection
- Sampling procedures
- Consistency of data with other sources
- Documentation
- Number of observations
- Number of variables
- Coding scheme
- Summary statistics
Secondary Data - Examples

**Example 1:**

- The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations.

  [http://www.cdc.gov/nchs/nhanes.htm](http://www.cdc.gov/nchs/nhanes.htm)

---

NHANES

- **Public data:**
  - Demographics
    - Age, Gender, Income
  - Dietary
    - Tuna eaten during past 30 days
  - Examination
    - Blood pressure, Weight
  - Laboratory
    - Cholesterol, Albumin
  - Questionnaire
    - Physical activity, Health Status

- **Restricted data:**
  - Youth Alcohol use and youth sexual behavior
  - Genetic
Secondary Data (cont.)

• Example 2: http://healthpolicy.ucla.edu/chis/Pages/
  – California Health Interview Survey (CHIS): data collected every two years by the UCLA Center for Health Policy Research in collaboration with the California Department of Public Health and the Department of Health Care Services.
    • The California Health Interview Survey (CHIS) is the state health survey and a critical source of data on Californians as well as on the state’s various racial and ethnic groups.
    • Policymakers, researchers, health experts, members of the media and others depend on CHIS for comprehensive data on the health of Californians.
    • A researcher used and analyzed the data to answer his/her own research question (i.e., Analysis of secondary data)

Secondary Data -- Examples

• Example 3: http://www.cdc.gov/brfss/
  – Behavioral Risk Factor Surveillance System (BRFSS) data collected annually by the Center for Disease Control and Prevention (CDC) and State Health Department
    • A researcher used and analyzed the data to answer his own research question (i.e., Analysis of secondary data)

• Others:
  – OSHPD: http://www.oshpd.ca.gov/HID/Data_Request_Center/
  – SEER: http://seer.cancer.gov/data/
Secondary Data (Cont.)

• **Considerations related to causal inference**

  – Causality and causal inference are complex issues.

  – **Association does not prove causation**

    • A is associated with B; does not mean A cause B.
      – E.g., cross-sectional data

Types of Survey Sample Design

• Simple Random Sampling
  – Each member of the population has an equal and known chance of being selected

• Stratified sample
  – The population is first divided into non-overlapping subpopulations: *strata* such as gender, race or SES.
  – Sample from each strata.

• Systematic Random Sampling
  – the selection of every $k_{th}$ element from a sampling frame with the sampling interval $k (=N/n)$.

• Cluster sample
  – Elements are selected in groups or clusters
  – PSU: Primary Sampling Unit. This is the first unit that is sampled in the design. For example, school districts from CA may be sampled and then schools within districts may be sampled.

  • Complex Survey design: increase efficiency, reduce cost.
Sampling Stages in NHANES

Stage 1
- Counties

Stage 2
- Segments

Stage 3
- Households

Stage 4
- Individuals

Secondary Data – Strengths

- Large samples
- Fast and inexpensive
- Population estimates
- Can test trends over time
- Avoid data collection problems
- Economic (no need for resources for data collection)
- Data collection process is informed by expertise and professionalism that may not available to small research
- Provide bases for comparison
Secondary Data – Limitations

• Non-experimental
• Data availability
  – Information you would like to have may not have been collected or collected in different ways than what you need (e.g., age, race, ...)
  – Data collected but not available to secondary researchers (e.g., phone)
• Level of observation
• Constructs measured by fewer items (no scales)
• Oftentimes require special statistical techniques
• Quality of documentation
• Data quality
• Outdated data
• Most are cross-sectional

Applications

http://accelerate.ucsf.edu/research/celdac