

# **Using Secondary Data**

**Patrick S. Romano, MD, MPH**  
**Professor of Medicine and Pediatrics**

# Data Sources

- **Primary data**

- Data that you or your colleagues collect specifically for the purpose of answering your research question.

- **Secondary data**

- Existing data collected for another purpose that you employ to answer your research question.

# Advantages of Primary Data

1. **You collect** exactly the data elements that you need to answer your research question.
2. **You can test** an intervention, such as an experimental drug or an educational program, in the purest way (a double-blind randomized controlled trial).
3. **You control** the data collection process, so you can ensure data quality, minimize the number of missing values, and assess the reliability of your instruments.

## **Advantages of Primary Data (cont)**

- 4. You avoid the tendency** with secondary data to "dredge" for research questions instead of developing your research questions based on real knowledge deficits.
- 5. You select a sample** that is specifically designed to help answer your research question, so irrelevant exclusion criteria are not a concern.

# Advantages of Secondary Data

1. **Less expensive** to collect than primary data.
2. **It takes less time** to collect secondary data.
3. **If you are looking for a small effect size**, it may be impossible to collect primary data on a sufficient number of cases.
4. **Collecting primary data prospectively may be unethical** if a therapy is “standard of care.”
5. **The accessible population for primary data may be less representative of the target population** than that for secondary data.
6. **You may not need to worry about informed consent, human subjects restrictions, etc.**

# Uses of Secondary Data Sets

- **Disease surveillance, estimating incidence and prevalence**
- **Cross-sectional studies**
- **Retrospective cohort studies**
- **Natural experiments, interrupted time-series analyses**

# **Examples of Secondary Data Sources**

## **1. Vital statistics**

- Statistical Master Death File/Multiple Cause of Death File**
- Birth file**
- Fetal death file**
- Birth-infant death linked file**

# **Examples of Secondary Data Sources**

## **2. Disease registries**

- California Birth Defects Monitoring Program**
- California Cancer Registry (Surveillance, Epidemiology, End Results)**
- Fatal Accident Reporting System**
- California Pesticide Illness Surveillance Program**
- California HIV/AIDS Reporting System**
- Trauma registries (county or hospital)**

# **Examples of Secondary Data Sources (cont)**

## **3. National surveys:**

- National Health Care Survey (NHCS)**
  - National Ambulatory Medical Care Survey (NAMCS)**
  - National Hospital Ambulatory Medical Care Survey (NHAMCS)**
  - National Survey of Ambulatory Surgery**
  - National Nursing Home Survey (NNHS)**
  - National Home and Hospice Care Survey (NHHCS)**
  - National Hospital Discharge Survey (NHDS)**

## **Examples of Secondary Data Sources (cont)**

### **3. National (cross-sectional) surveys:**

- National Health Interview Survey (NHIS)**
- National Health and Nutrition Examination Survey (NHANES)**
- National Immunization Survey (NIS)**
- National Survey of Family Growth (NSFG)**
- Behavioral Risk Factor Surveillance System (BRFSS)**
- Youth Risk Behavior Surveillance System (YRBSS)**
- Medicare Current Beneficiary Survey (MCBS)**

## **Examples of Secondary Data Sources (cont)**

### **3. National (longitudinal) surveys:**

- Medical Expenditure Panel Survey (MEPS)**
- National Longitudinal Survey of Youth**
- National Longitudinal Study of Adolescent Health**
- Early Childhood Longitudinal Study**
- Panel Study of Income Dynamics**
- NHANES I Epidemiologic Follow-up Study**

## **Examples of Secondary Data Sources (cont)**

### **4. Statewide surveys:**

- **California Health Interview Survey (CHIS)**  
<http://www.chis.ucla.edu>)

### **5. Hospital data:**

- **California Hospital Discharge Data (OSHPD)**
- **Healthcare Cost and Utilization Project (AHRQ)**
  - **Nationwide Inpatient Sample (NIS)**
  - **State Inpatient Databases (SID)**
  - **State Ambulatory Surgery Databases (SASD)**
  - **State Emergency Department Databases (SEDD)**
  - **Kids' Inpatient Database (KID)**

## **Examples of Secondary Data Sources (cont)**

- 6. Financial or service utilization data (typically linked to an eligibility or enrollment file with demographic data)**
  - MediCal Paid Claims File**
  - Medicare Provider Analysis and Review (MEDPAR) and Standard Analytic Files (DME, HHA, hospice, SNF, inpt, outpt, MD)**
  - VA Patient Treatment File**
  - California Department of Developmental Services Masterfile**

## **Examples of Secondary Data Sources (cont)**

- 7. Previous or ongoing research studies (“ancillary studies”)**
  - Cardiovascular Health Study (CHS)
  - Women’s Health Initiative (WHI)
  - Your mentor
  
- 8. Other data bases**
  - Computerized health information systems (Regenstrief, Mayo, UCD?)
  - Industrial records

# **The Power of Linking Secondary Data Sets**

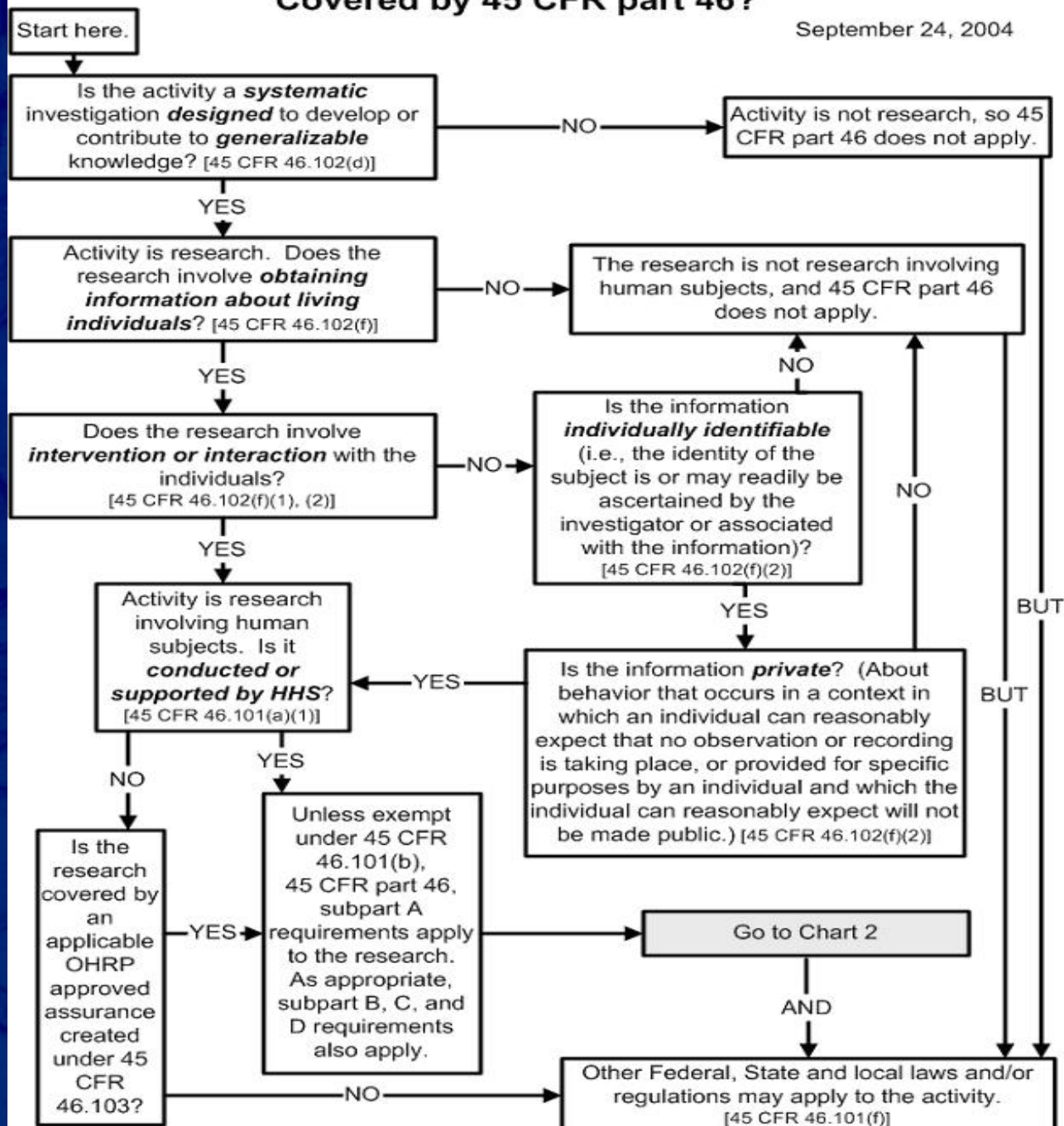
- **Linked California Birth Cohort File with the Patient Discharge Data Set for all deliveries.**
- **Linked California Statistical Master Death File with the Patient Discharge Data Set**
- **Linked California Cancer Registry with the CMS/Medicare enrollment file, the California Medicaid enrollment file, and the Patient Discharge Data Set**
- **Link your own data with the National Death Index or other (nonconfidential) data sets**

# Issues in Using Secondary Data

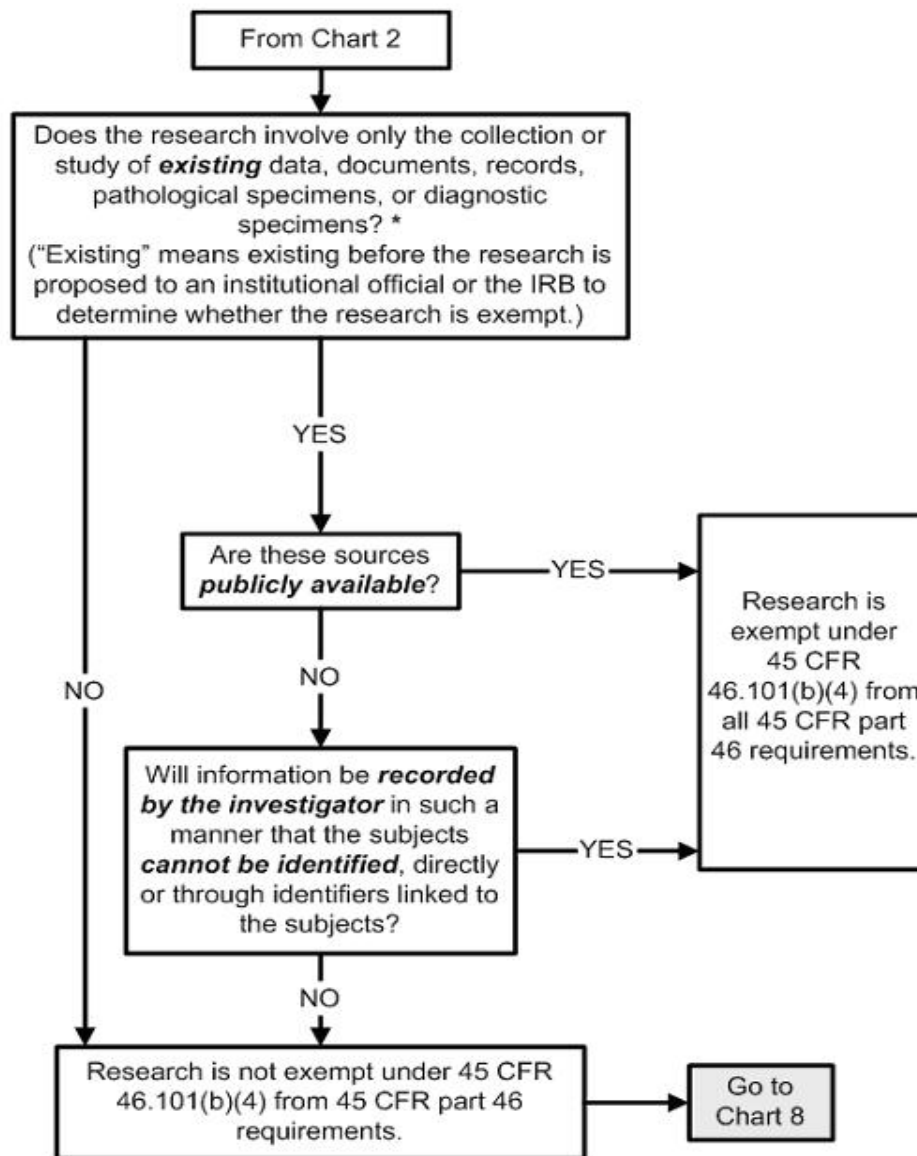
- Research using secondary data may not qualify as human subjects research under 45 CFR 46.102(f).
- Even if it qualifies, it may be exempt from IRB review under 45 CFR 46.101(b).
- But it may still fall under the HIPAA (Health Insurance Portability and Accountability Act) Privacy Rule, so “Privacy Board” (IRB) review and a data use agreement may be needed to get the “limited data set” variables you want.
  - All elements of dates (except year)... directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89...
  - All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code...
  - <http://privacyruleandresearch.nih.gov/>

# Chart 1: Is an Activity Research Involving Human Subjects Covered by 45 CFR part 46?

September 24, 2004



## Chart 5: Does Exemption 45 CFR 46.101(b)(4) (for Existing Data Documents and Specimens) Apply?



\* Note: See OHRP guidance on research use of stored data or tissues and on stem cells at <http://www.hhs.gov/ohrp/policy/index.html#tissues> and #stem, and on coded data or specimens at #coded for further information on those topics.

September 24, 2004

# Issues in Using Secondary Data

- **Missing data are often a problem.**
  - Are data “missing completely at random” (i.e., the probability of an observation being missing is uniform, and does not depend on observed or unobserved measurements)?
  - If so, then analyses of nonmissing data give valid inferences.
  - Are data “missing at random” (i.e., the mechanism for missingness can be expressed solely based on the *observed data*, and does not depend on any unobserved data)?
  - If so, then missing data can be imputed using nonmissing data on the same case (or condition the analysis on the observed data that affect missingness).

# Approaches to Missing Data

- **Completers analysis**
  - Inefficient (drops observations with missing data)
  - Confusing (different N for different analyses)
  - Potentially biased (missing not at random)
- **Simple mean imputation**
  - Dilutes effects, reduces variances
- **Regression mean imputation**
  - Less biased, but variance of imputations is still too small
- **Create new category for missing observations**
  - May not reduce bias, because missing category may be very heterogeneous
- **Weighted analyses (preferred)**
  - Weight each observation by the inverse of the probability that it was nonmissing
- **Hot deck imputation (preferred)**
  - A *random draw* (or series of random draws) is made from some suitable distribution

# Issues in Analyzing Secondary Data

- **Study documentation carefully.**
- **Population surveys usually have a complex sample design, in which some persons in the population have a higher probability of being sampled than others.**
- **Users must apply “sample weights” and often special analytic procedures to account for the data structure and estimate the “design effect.”**
- **Software: SUDAAN, STATA, WesvarPC. Now SAS, SPSS too...**