



UNIVERSITY *of* MARYLAND
SCHOOL OF MEDICINE

Secondary Data Sources

Jennifer S Albrecht, PhD

Associate Professor

Department of Epidemiology and Public Health

University of Maryland School of Medicine

August 10, 2021

Outline

- Secondary data definition
- Advantages
- Disadvantages
- Research questions
- Sources

What is Secondary Data?

Data collected by someone other than the user

- Health, demographics, biometrics, income
 - Research purposes
 - Other Purposes
 - Administrative Claims
 - Record keeping
 - Allocation of resources

Useful to answer questions beyond the original intent

Advantages of Secondary Data

- Already collected – someone else has already dealt with:
 - logistics
 - cost
- Access to data that would be difficult to collect
 - Nationally representative
 - Years of data
 - Standardized collection
- Help is available
 - Documentation
 - Technical support



Secondary Data

- May qualify as exempt from Institutional Review Board (IRB) review
 - No identifying information (sometimes)
 - Data already collected
 - Still need to submit the protocol to the IRB for the exempt determination



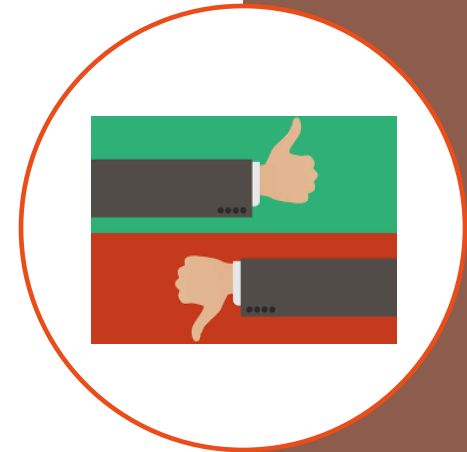
Limitations of Secondary Data

- Not *exactly* the right question
- Wrong population
- Complex survey designs
 - Must use weights to obtain accurate estimates
- Can't link to other data (if de-identified)
- Multiple teams may be working on the same questions with the same data



Benefits > Limitations?

- Learn new analysis techniques
- Challenge yourself to think about new aspects of the topic – what question can I answer with the data available?
- Understanding the limitations of your data helps you place your results within the larger literature
- Opportunity to collaborate with new investigators



Group Chat Question

Have you used
secondary data?
What was the source?



Developing a Research Question with Secondary Data

1. Start with a general topic or question
2. Conduct a literature review to identify gaps
3. Identify a few research questions
4. Identify data source
5. Refine research question



Identify Data Source

- **Best bet** – Inter-university Consortium for Political and Social Research (ICPSR)
 - searchable database at UMichigan

<https://www.icpsr.umich.edu/index.html>

- Find data source containing variables of interest
- Investigate variables and narrow down to a few sources

Identify Data Source

- Consult documentation
 - Questionnaires – text of questions
 - Codebook - distribution of responses and sample size
 - Check that
 - Appropriate variables exist
 - Adequate sample size available
 - Sufficient variation in distribution available
 - What's needed for access?



Before You Start

- Survey data are often weighted to bring them up to national-level estimates
- Critical to understand how to work with weighted data.
- Documentation contains important information regarding how to analyze the data

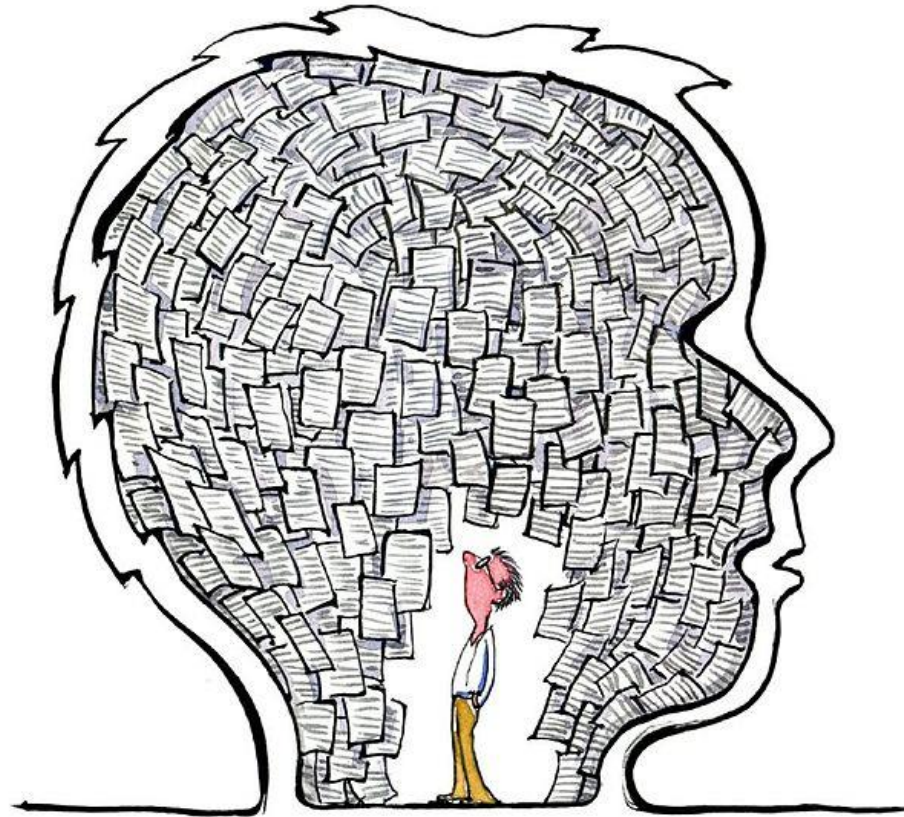
Pop Quiz: Examples of Secondary Data

Sofia is a student who is considering sources of secondary data for her dissertation proposal. Which of the following would be considered *secondary data* for Sofia's proposal?

- a) Patient prescription medication records from the VA
- b) Data her mentor collected 3 years ago
- c) A survey administered by the Centers for Disease Control and Prevention
- d) Data she collected last year with an aim that is different from her dissertation proposal

Sources of Secondary Data

United
States



International

The United States



Sources of Secondary Data: The Big Picture

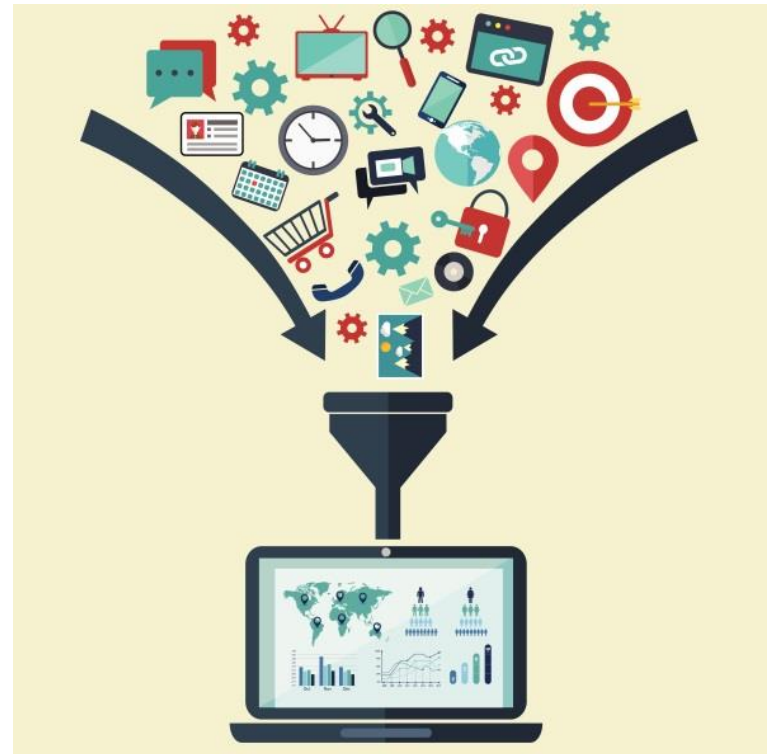
- **US Government**

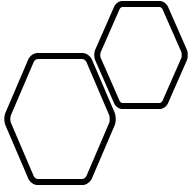
- US Bureau of the Census
- Bureau of Labor Statistics
- National Center for Health Statistics
- Substance Abuse and Mental Health Data Archive
- NIH – funded projects >\$500K must share data
- Medicaid/Medicare administrative claims
- FDA
- CDC - <https://wonder.cdc.gov/>



Sources of Secondary Data: The Big Picture

- **Insurance Industry**
 - Commercial claims
- **State Government**
 - Inpatient encounters
 - Deaths
- **Hospitals**
 - Electronic medical record
 - Central data repository
- **Other**
 - Mentors
 - Colleagues





Free Data



National Center for Health Statistics

- <https://www.cdc.gov/nchs/index.htm>
- Conduct major data collection programs to
 - Identify and address health issues
 - Guide public health and health policy decisions
- Research and methodology
 - Survey design
 - Use of technology

National Center for Health Statistics

Population Surveys

- National Health Interview Survey (NHIS)
- National Health Nutrition and Examination Survey (NHANES)
- National Survey of Family Growth (NSFG)

Vital Records

- National Vital Statistics System (NVSS)
- National Death Index (NDI)

National Center for Health Statistics

Data Linkages

- Linked by NCHS to
 - National death index
 - Medicare/Medicaid
 - Social security
 - Department of housing

<https://www.cdc.gov/nchs/data/datalinkage/LinkageTable.pdf>

Group Question



Does anyone have a research question that could be answered using the data sources we've looked at?

EXAMPLE

Quality of Hospice Care for Patients with Dementia

- The aim of this study was to quantify differences in quality-of-care measures between hospice patients with and without dementia

Quality of Hospice Care for Patients with Dementia

- *Design*: Cross-sectional analysis of data.
- *Setting*: 2007 National Home and Hospice Care Survey.
- *Participants*: 4,711 discharges from hospice care.

Quality of Hospice Care for Patients with Dementia

- 450/4,711 (9.5%) with diagnosis of dementia.
- Individuals with dementia were more likely to:
 - receive tube feeding (OR 2.6; 95% CI 1.4, 4.5)
- and less likely to:
 - have a report of pain (OR = 0.6; 95% CI 0.3, 0.9)
- Consider as potential quality-of-care measures among hospice patients with dementia.

Sources of Secondary Data

- Interuniversity Consortium for Political and Social Research
 - <https://www.icpsr.umich.edu/icpsrweb/index.jsp>
- National Epidemiologic Survey on Alcohol and Related Conditions
 - <https://www.niaaa.nih.gov/research/nesarc-iii>
- Health and Retirement Study
 - <http://hrsonline.isr.umich.edu/>

Sources of Secondary Data

- National Longitudinal Study of Adolescent to Adult Health
 - <http://www.cpc.unc.edu/projects/addhealth>
- Substance Abuse and Mental Health Data Archive
 - <http://www.datafiles.samhsa.gov/>
- Behavioral Risk Factor Surveillance System
 - <https://www.cdc.gov/brfss/index.html>



There are three rabbits here. Really.

Sources of Secondary Data

- Medical Expenditure Panel Survey
 - <https://meps.ahrq.gov/mepsweb/>
- American Community Survey
 - <https://www.census.gov/programs-surveys/acs/about.html>
- American Time Use Survey
 - <https://www.bls.gov/tus/home.htm>

Sources of Secondary Data

- Women's Health Initiative
 - <https://www.nhlbi.nih.gov/whi/>
- U.S. National Library of Medicine
 - <https://www.nlm.nih.gov/hsrinfo/datasites.html>

Data that isn't free



Administrative Claims

- Medical and pharmacy claims
- Beneficiaries characteristics and enrollment info
- Longitudinal follow-up (years!!) with complete capture of health info
 - *Except bundled services*

Administrative Claims

Closed claims data for 130+ million patients available since 2006 to derive robust and meaningful insight

Cost

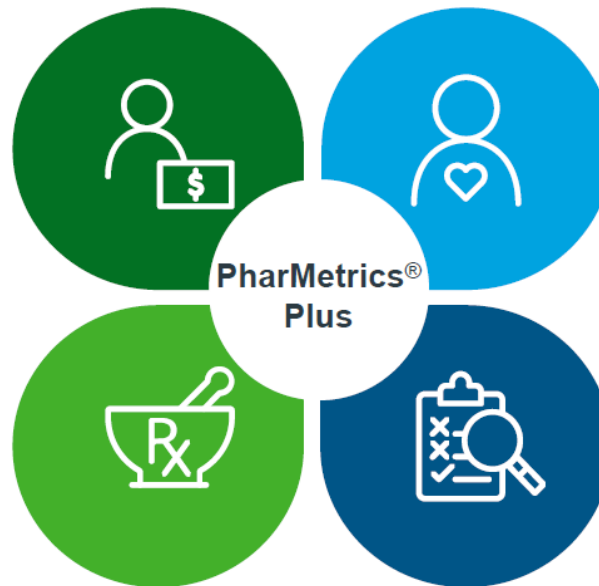
True cost data for various markets

- Allowed/Paid Amounts
- Coinsurance
- Copayment
- Deductible
- Charges

Medication Use

Every pharmacy claim billed through insurance

- Retail/Mail Order/Specialty Pharmacy Use
- Infusions/in-office treatment(s) administered
- Date of Fill
- On/Off Formulary Status
- DAW code & Dispensing Fee



Patient Information

Geo-granular data available for patients

- Patient year of birth & gender
- Census region, state, and zip3
- Payer Type and enrollment dates

Diagnosis & Hospitalization

Every healthcare facility interaction billed through insurance

- Diagnosis codes: ICD-9/ICD-10
- Procedure Codes: CPT, HCPCS
- Primary Care and Specialty Visits
- Provider specialties
- Inpatient Stays (Dates, Discharge status, confinement number, admitting diagnoses)

This information is for IQVIA but similar for all. Adjudicated claims submitted for reimbursement.

Medicare Data

- 98% of Americans ≥ 65 are covered
- Access through ResDAC
 - <https://www.resdac.org/>
- Parts A (Hospital), B (Physician) and D (Prescription Medications)
- Medicare Current Beneficiary Survey
 - <https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/MCBS/index.html?redirect=/mcbs>
- Pharmaceutical Research Computing
(prc@rx.umaryland.edu)

OptumLabs Data Warehouse

- Access to over 15 years of United Healthcare Claims
- UMB is no longer a partner 😞
- Must partner with another University like Hopkins or Mayo
 - Bonus: Opportunity to collaborate!

Other Sources of Claims Data

- IQVIA
- Marketscan
- Medicaid

Link to Other Data Sources

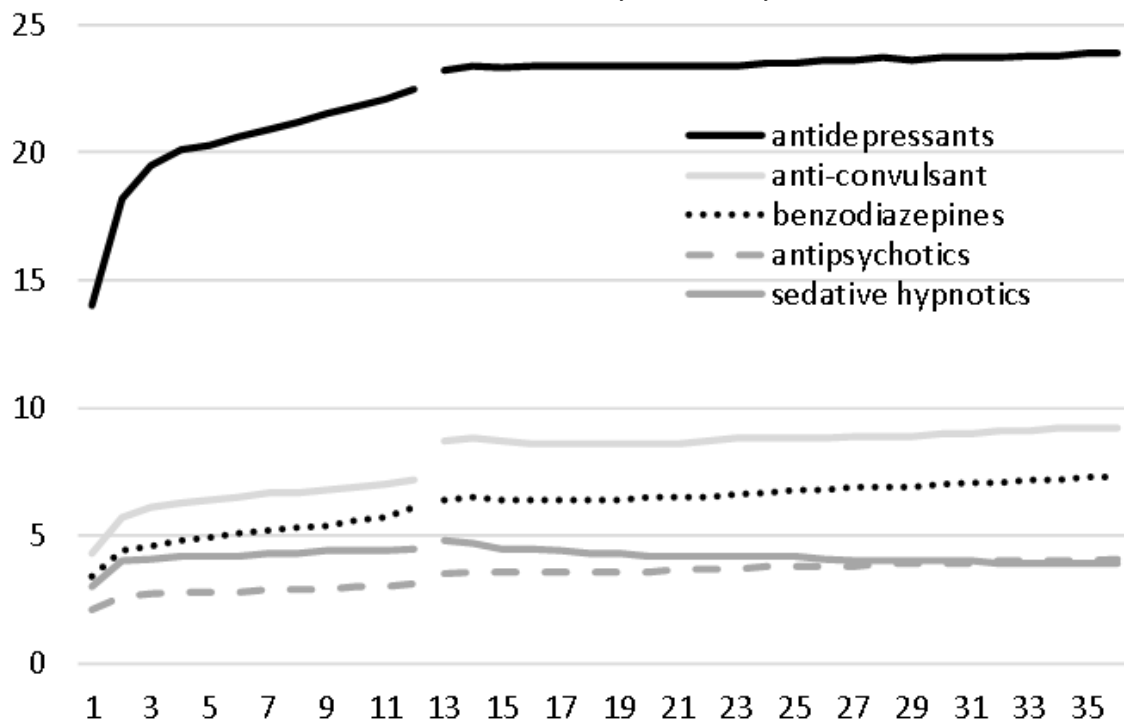
- **Big Data** - the electronic medical record
 - Test results
 - Lab values
 - Provider notes
- Ecological variables
 - Access to healthcare
 - Poverty
 - Education



Group Question

- Has anyone used administrative claims data for research?
- Does anyone have a question they think could be answered using administrative claims data?

Average Monthly Prevalence of Psychotropic Medication use, n=207,354



Example
Patterns of
Psychotropic
Medication Use among
Individuals with
Traumatic Brain Injury

International Data



Sources of Secondary Data

- Demographic and Health Surveys Program
 - <http://dhsprogram.com/>
- Gateway to Global Aging Data
 - <https://g2aging.org/>
- World Bank Microdata
 - <http://microdata.worldbank.org/index.php/catalog/central/about>

Sources of Secondary Data

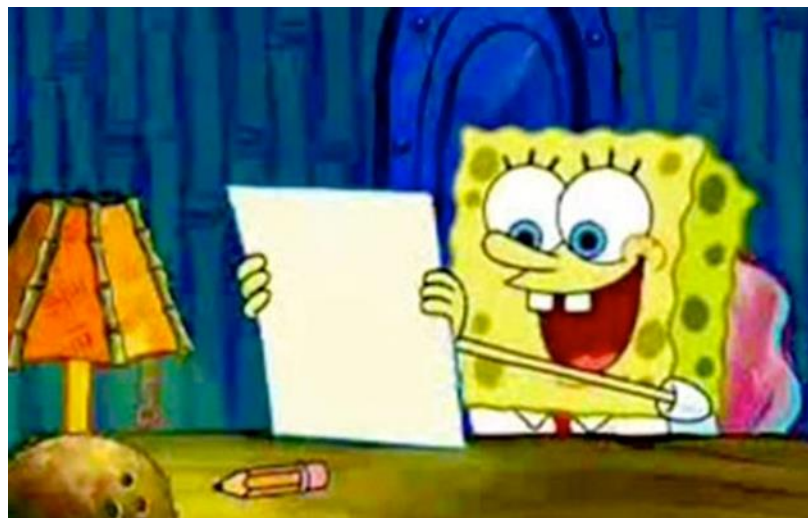
- Eurostat (statistics only)
 - <http://ec.europa.eu/eurostat/web/main/home>
- Global Health Information & Resources (UMich)
 - <http://guides.lib.umich.edu/c.php?g=282776&p=1884201>
- UN Data
 - <http://data.un.org/Default.aspx>

Take home messages



- Secondary data provides easily-accessible opportunities for analysis
- Develop strategies to address limitations

Remember.....



- Secondary data means that someone else is expert at using the data.
- Familiarize yourself as much as you can with the data
- Approach help desk, colleagues, or authors of papers of interest with intelligent questions.