Using Population-Based Data in Secondary Data Analysis

Presenters: Arthur Davidson, MD, MSPH; Allison Kempe, MD, MPH

Learning Objectives:

- Describe and categorize an array of population-based data resources for use in secondary analyses
- Describe the advantages and disadvantages of secondary datasets and how to access population-based data
- Describe examples of important publicly available datasets

Description of session:

A didactic presentation will describe population-based data resources used to answer research questions, as well as their advantages/disadvantages when used in secondary data research. Access and suitability issues of population-based datasets for research will be described with a short interactive activity (described below).

Interactive Activity:

Think of a research question of interest to you or within your current focus area (*examples below*):

Clinical questions:

- Among patients diagnosed with appendicitis, is hospital readmission higher for those receiving initial antibiotic therapy alone compared with patients receiving initial surgical intervention?
- Have rates of insulin plus oral medication among patients with Type 2 diabetes changed over time?

Policy or population-based questions:

- Do states with more restrictive personal exemption policies for childhood immunizations have higher rates of completion of recommended immunizations for 19-35 month old children?
- What trends in obesity prevalence have been observed in US youth and adults, by sex and age, during the past decade?

Considerations:

1. What key features of a population-based database are needed to answer this question? Think of the unit of analysis, frequency of data collection, granularity and/or longitudinality.
2. Assess public availability (yes/no); What barriers might exist to access these data? Who would be key partners?
3. What regulatory or legal issues should an investigator weigh to obtain and then analyze these types of data?

Analytic Plan Template – *Course Material (Annotated)*

Date:
Working Paper/Study Title/Grant application:

Lead Investigator(s):
Project Team Members:

> *Note: This is the right time and place to begin the discussion around authorship and position in the listed authors. First and last should probably be defined early.*

Specific Aim(s): *Specify which specific aim this manuscript addresses (if applicable)*

Resources:

> *Note: Important to define whether the resources are readily available and how to tactically and tactfully acquire this support. May require rethinking who are the team members.*

Who are the statistician and data analyst for this manuscript?

> *Note: Do you have the qualitative or quantitative skills to conduct and analyze the study?*

Do you need additional resources (e.g., additional programming, database management)?

> *Note: What types of data will you be using? Is there someone skilled in acquiring data from a secondary source? How easy will access to those secondary data sources be? What is the track record working with data from this source? What is the format for data received and how will that make use easier/harder?*

Research Objective(s): *State your research question(s) or goals*

> *Note: Is this a real research study or a quality improvement effort? How would one or the other influence the need for regulatory review? Different from specific aims, the research objectives are typically less structured and more a narrative of the intent of the work.*

Study Design:

> *Note: There may be an iterative relationship between the qualitative and the quantitative elements of your research design. These are not clearly accounted for in this analytic plan. How might this document change to achieve that? A few resources to review various study designs:*
>
> Study Designs in Epidemiology
> Secondary Data Analysis: A Method of which the Time Has …
> An Introduction to Secondary Data Analysis - National Center …

Time Frame (e.g., cross-sectional, longitudinal, retrospective cohort, cohort):

RCT or Observational:

Other comments:

PATIENT COHORT AND SUBJECTS: Describe the patients who are part of this manuscript. Specify **inclusion** and **exclusion** criteria and **study site(s)**.

> *Note: A principle in clinical trials is that the analysis must take into account the level at which randomization occurred. The number of observations in the analysis should match the number of 'units' that were randomized. In a simple parallel group design for a clinical trial, participants are individually randomized to one of two intervention groups, and a single measurement for each outcome from each participant is collected and analyzed.*
>
> *Note: Variations on this design:*
> - *groups of individuals were randomized together to the same intervention (i.e. cluster-randomized trials);*
> - *individuals undergo more than one intervention (e.g., in a cross-over trial, or simultaneous treatment of multiple sites on each individual); or*
> - *multiple observations for the same outcome (e.g., repeated measurements, recurring events, measurements on different body parts).*
>
> *Note: Are we discussing patients, providers, systems or a combination of all of these? What is the fundamental unit of analysis?*

Inclusion Criteria

Exclusion Criteria

Study Site

DATA SOURCES (identify existing data and additional data needed):

> *Note: All sources should be linked to the dependent and independent variables below*

HYPOTHESES: State key hypotheses explicitly. If there is no specific hypothesis (e.g., descriptive) simply state what is proposed.

> H1:
> H2:
> H3:

VARIABLES FOR ANALYSIS *(link these to each research question or hypothesis to be tested):*

> *Note: If a variable is both in the dependent and independent variable list, it probably is worthwhile to parse them out into several tables after each hypothesis.*
>
> *Note: Adding which hypotheses are associated with these variable helps to assure all variables have a clear purpose related to SA and hypotheses*

Dependent Variables

Variable	Description	Type	Source	Hypothesis
Primary Outcomes				

Secondary Outcomes				

Note: Mark the hypothesis to which the variable is related both here in the dependent and below in the independent variable tale.

Independent Variables *(identify as main exposure variable, covariates, potential confounders)*
modify as needed)

Note: Which of the following fairly common variables are part of your study, which variables are totally unique to your study?

Example table:

Variable	Description	Type	Source	Hypothesis
Age	Created variable using DOB and enrollment date (in years)	continuous	administrative	
Gender	Male=1; female=0	dichotomous	administrative	
Race/ethnicity	Recode from administrative data, collapse into Latino=1, African American=2, Non-Hispanic white=3, Others?	categorical		
comorbidities	Apply case-finding algorithms – use individually or count	From ICD-9 codes	Sum of comorbidities from eligible visits	
Insurance	Private Federal -Medicaid, Medicare State – CIPC/uninsured	From last visit		
Intervention group	1=intervention, 0=controls			
time	Time since enrollment (in months, days?)			
BMI	BMI = (Weight in Pounds / (Height in inches x Height in inches)) x 703		Calculated	
Add as many				

as necessary to describe your study variables….				

Note: Might want to describe the case finding process in detail here to assure clearly documented

DATA ANALYSIS
Some general analytic approaches (expand and modify as needed).

> *Note: How does your study fit with items a-e? Are there other topics that you should consider for your study or more broadly for all studies in your domain?*

1. *Setting and subjects.* Usually start by describing the sample and addressing issues of external and internal validity
    a. Generate frequency distributions and summary statistics (e.g., means, standard deviations, median, rates) on outcome variables, sociodemographic and clinical variables, and other relevant variables of interest. For continuous outcomes, examine distributions to determine whether normality assumptions hold or if transformations or other approaches may be needed.
    b. Are the patients in this clinic similar to target population?
        i. Usually start by computing descriptive statistics for sample – frequencies, means (sd)
    c. Are people who refuse similar to participants?
        i. If possible, compare participants to non-participants: t-tests, chi-square tests, or just compute 95% CI on means and proportions for participants
    d. Are dropouts similar to completers (longitudinal designs)?
        i. Compare dropouts to completers and assess for differences in baseline covariates and outcomes using chi-square tests, t-tests, Kendall's tau
        ii. Also determine whether there is differential dropout by study group. For longitudinal designs this will help determine whether the data are 1) MCAR = Missing completely at random, 2) MAR= missing at random: ignorable,3) MNAR= missing not at random: missingness related to something you may not observe The first two are ignorable but analytic requirements differ; the last is non-ignorable. Variables related to missingness need to be included in the analysis.
    e. If an RCT, compare treatment groups on key baseline variables using chi-square tests and t-tests
        i. This will help determine which covariates are potential confounders and need to be included in the analysis.
2. Bivariate analyses (parametric/nonparametric, correlations vs. categorical statistics)
3. Multivariate analyses
    a. Choice of model and rationale (e.g., logistic regression, linear regression, survival analysis, factor analysis)
    b. Strategy for c*ovariate identification and selection.* Screen by domains (e.g., sociodemographic or clinical) and retain all independent variables that are associated

with the outcome at ≈ $p<.20$ for inclusion in initial multivariate models. Final models will include covariates that are associated with missingness (if longitudinal), treatment group, or the outcome (at ≈ $p<.15$ in multivariate models, depending on sample size).
c. Assessment of appropriateness/fit of model
d. Strategies to validate model (split sample, separate sample, etc.)

Analyses to address study questions/hypotheses. *Some text here will help with writing later on. This would be a good place to mention specific analyses (e.g., multivariate linear regression, etc) and highlight pros and cons or issues that need to be addressed.* The primary outcome for this analysis is XXX.

    H1.
    H2.
    H3.

Include empty, mocked up tables, if possible
        Note: Imagine the final product or manuscript. What would be your 4 or 5 tops tables or figures? Create shell tables to visualize how your analytic approach will create a summary that communicates your intended output and that will address your primary research question.

Next steps, meetings, assignment of responsibilities, etc:
        Note: Create a work breakdown structure that allows you to define all the steps, dependencies, priorities and timelines.
        Note: Use the plan to drive accountability and forward momentum

An abbreviated description of some secondary datasets

Examples of some publicly available datasets from the National Center for Health Statistics (www.cdc.gov/nchs/)

- **National Health Interview Survey (NHIS)**
  - Data on a broad range of health topics are collected through personal household interviews
  - Multipurpose survey: health, illness, services
  - Complex sample design with clustering and stratification
  - Sample representative of non-institutionalized US civilian population living at addressed dwellings
  - Administered annually since 1957
  - Approximately 100k respondents/year from 1986 – 2021
  - Typical measures:
    - demographics,
    - employment,
    - health status,
    - activity limitations,
    - healthcare utilization
- **National Health and Nutrition Examination Survey (NHANES)**
  - Program of studies designed to assess the health and nutritional status of adults and children in the US
  - Unique in that it combines interviews and physical examinations
  - Complex sample design with clustering and stratification
  - Sample representative of non-institutionalized US civilian population living at addressed dwellings
  - Administered 1971-'80, 1988-'94, 1999-current
  - Approximately 30K respondents/year earlier period, now about 5K/year
  - Typical Measurements:
    - physiological,
    - diet & nutrition,
    - blood and urinary labs,
    - alcohol and tobacco use
- **National Ambulatory Medical Care Survey (NAMCS)**
  - National survey designed to meet the need for objective, reliable information about the provision and use of ambulatory medical care services in the US

Using Population-Based Data in Secondary Data Analysis

Presenters: Arthur Davidson, MD, MSPH; Allison Kempe, MD, MPH

[Notes]

COPRH Con
Colorado Pragmatic
Research in Health
Conference

ACCORDS
ADULT AND CHILD CONSORTIUM FOR HEALTH OUTCOMES
RESEARCH AND DELIVERY SCIENCE
UNIVERSITY OF COLORADO | CHILDREN'S HOSPITAL COLORADO

Colorado Clinical and Translational
Sciences Institute (CCTSI)
UNIVERSITY OF COLORADO **DENVER | ANSCHUTZ MEDICAL CAMPUS**

- Findings are based on a sample of visits to non-federal employed office-based physicians who are primarily engaged in direct patient care
- age, sex, race, ethnicity, and visit characteristics such as patient's reason for visit, physician's diagnosis, services ordered or provided, and treatments, including medication therapy
- data about the physician and their practice characteristics are collected during a survey induction interview (~2K/year followed)
- 1973 – present (public use)
- **National Immunization Survey (NIS)**
  - List-assisted random-digit-dialing telephone survey followed by a mailed survey to children's immunization providers
  - Began data collection to monitor childhood immunization coverage
  - Target population for the NIS is children between the ages of 19 and 35 months living in the United States at the time of the interview
  - Added teen-focused surveys 13-17 years in 2006
  - COVID surveys in adults (>18 years)
  - 1994 – present (children surveyed annually)
- **National Hospital Discharge Survey**
  - National probability survey designed to meet the need for information on characteristics of inpatients discharged from non-Federal short-stay hospitals in the United States
  - Used to examine important topics of interest in public health and for a variety of activities by governmental, scientific, academic, and commercial institutions
  - Contains over 266,000 records from a sample of hospital discharge records
  - Conducted annually since 1965 to 2010
  - More recently, converted to National Hospital Care Survey integrates inpatient data formerly collected by the NHDS with the emergency department (ED), outpatient department (OPD), and ambulatory surgery center (ASC) data collected by the National Hospital Ambulatory Medical Care Survey (NHAMCS).
- **National Survey of Ambulatory Surgery**
  - The only national study of ambulatory surgical care in hospital-based and freestanding ambulatory surgery centers
  - Conducted from 1994-1996, 2006
  - Now efforts to integrate with the National Hospital Ambulatory Medical Care Survey (NHAMCS)
- **National Nursing Home Survey/ National Study of Long-Term Care Providers**
  - A continuing series of national sample surveys of nursing homes characteristics, their residents, services, and staff
  - All nursing homes included in this survey had at least three beds and were either certified (by Medicare or Medicaid) or had a state license to operate as a nursing home
  - Conducted 1973-'74, 1977, 1985, 1995, 1997, 1999, 2004
  - 1500 nursing homes were selected in 2004
  - NSLTCP includes inpatient rehabilitation facilities and long-term care hospitals,adult day services centers, assisted living and similar residential care communities, home health agencies, hospices, and nursing homes

- 2012 – current NSLTCP

Data Set/Source	Description	Variables
Behavioral Risk Factor Surveillance System (BRFSS) **NCHS**	CDC telephone survey designed to collect state-specific general population data on behaviors that are related to the leading causes of premature death. The basic philosophy is to collect data on actual behaviors, rather than on attitudes or knowledge, to support risk reduction and disease prevention activities. Data collection began in 14 states in 1984 and all states have been participating in this survey since 1994, which allows states to compare risk factor prevalence with other states and monitor the effects of interventions over time, as well as permits the assessment of geographic patterns of risk factor prevalence.	Risk Factors • Acute drinking • Cholesterol awareness • Chronic drinking • Cigarette use • Drinking and driving • Exercise • Hypertension • Hypertension awareness • Overweight • Safety belt usage • Smokeless tobacco Disability • Physical inactivity/activity Race/ethnicity

Colorado Pragmatic Research in Health (COPRH) Conference: **Breakout Session:** *Using population-based data for secondary analysis*

May 25, 2021   **Supplemental Handout**

Data Set/Source	Description	Variables
Healthcare Cost and Utilization Project (HCUP) **AHRQ**	AHRQ-sponsored family of administrative, longitudinal databases, web-based products, and software tools developed as part of a Federal- State-Industry partnership to build a standardized, multi-state health data system. HCUP is based on data collected by individual states and provided to AHRQ by the states. These data are used for research on hospital utilization, access, charges, quality and outcomes. They are used to describe patterns of care for uncommon as well as common diseases, analyze hospital procedures, including those that are performed infrequently, and study the care of population sub-groups such as minorities, children, women, and the uninsured. Researchers and policymakers use HCUP data to identify, track, analyze ~~and compare hospital~~ statistics at the national, regional and state levels. **HCUPnet** gives you easy access to national ~~statistics and trends and~~ selected state statistics about hospital stays. **HCUPnet** generates statistics using the 1997 data from the Nationwide Inpatient Sample (NIS) and from the State Inpatient Databases (SID) for those states that have agreed to participate.	*NIS* <u>Discharge-Level</u> • Linkage elements • Physician identifiers • Data source identifiers • Area identifiers • Patient demographics • Clinical information • Days and dates • Admission/discharge status • Payment information Hospital-Level • Linkage elements • Sampling stratum characteristics • Weights *SID* <u>Discharge-Level</u> • Linkage elements • Physician identifiers • Data source identifiers • Area identifiers • Patient demographics • Clinical information • Days and dates • Admissions/discharge status • Payment information <u>State-specific (varies across states)</u> • Physician specialty • Readmission indicator • Diagnoses present at admission • Type of admission • Birth weight • Detailed charges • Expected payer • Encrypted patient zipcode • Encrypted patient code <u>Hospital-Level</u>: • Linkage elements No survey but rather compilation of administrative longitudinal databases and user-friendly software

Data Set/Source	Description	Variables
Medical Expenditure Panel Survey (MEPS) **AHRQ**	Four surveys in one: • Household component • Insurance component • Medical provider component • Nursing home component AHRQ survey designed to continually provide policymakers, health care administrators, businesses, and others with timely and comprehensive information about health care use, health care costs in the United States, and to improve the accuracy of their economic projections. MEPS collects data on the specific health services that Americans use, how frequently they use them, the cost of these services and how they are paid for, as well as the data on the cost, scope, and breadth of private health insurance held by and available to the U.S. population. MEPS can link data on health services spending and health insurance to the demographic, employment, economic, health status, and other characteristics of survey respondents. NMCES 1977-87 (periodic); MEPS began in 1996 (continuing longitudinal).	Household Component • Health care use • Expenditures • Sources of payment • Insurance status • Functional limitations and disabilities • Restricted activity days • Access to care • Acute and chronic conditions Insurance Component • Health insurance plans • Premiums, deductibles, co-pays Medical Provider Component • Expenditure data from hospitals, home health providers, pharmacies, office-based physicians providing care to household component respondents Survey instruments: www.meps.ahrq.gov/survey.htm

Colorado Pragmatic Research in Health (COPRH) Conference: **Breakout Session:** *Using population-based data for secondary analysis*

May 25, 2021   **Supplemental Handout**

Data Set/Source	Description	Variables
National Ambulatory Medical Care Survey (NAMCS) **NCHS**	NTIS national probability sample survey of patient visits to the offices of non-federally employed office- based physicians who are primarily engaged in office-based, direct patient care, but not in the specialties of anesthesiology, pathology, or radiology. The survey includes information on patient, physician and visit characteristics. The survey measures health care utilization across a variety of providers. The NAMCS was conducted annually from 1973-81, again in 1985, and resumed as an annual survey in 1989.	**Patient characteristics:** • Age • Sex • Race/ethnicity • Whether the patient currently smokes cigarettes **Physician characteristics:** • Physician specialty • Professional identity • Geographic location **Visit characteristics:** • Patient's reason(s) for visit • Injury-related visits • Cause of injury • Physician's diagnoses • Expected source(s) of payment • Ambulatory surgical procedures performed • Diagnostic/screening services • Therapeutic/preventive services • Medication/injections ordered supplied, or administered • Providers seen • Referral status • Prior visit status • Disposition • Duration

Data Set/Source	Description	Variables
National Hospital Ambulatory Medical Care Survey (NHAMCS) **NCHS**	NCHS (CDC) survey designed to collect data on the utilization and provision of ambulatory care services in hospital emergency and outpatient departments. Findings are based on a national sample of visits to the emergency departments and outpatient departments of non-institutional general and short-stay hospitals, exclusive of Federal, military, and Veterans Administration hospitals, located in the 50 States and the District Columbia. The survey uses a four-stage probability design with samples of geographically defined areas, hospitals within these areas, clinics with hospitals, and patient visits within clinics.	• Patient demographics • Expected source of payment • Reason for visit • Cause of injury • Physician diagnoses • Diagnostic/screening services • Procedures • Medications • Providers seen • Disposition • Information on selected hospital characteristics

Data Set/Source	Description	Variables
National Health and Nutrition Examination Survey (NHANES I, II, III) **NCHS**	NCHS (CDC) survey designed to collect information about the health and diet of people in the United States. This survey combines a home interview with health tests that are done in a Mobile Examination Center (MEC virtual tour). NHANES I (NHEFS – epidemiologic follow-up study): Investigates the relationship between clinical, nutritional, and behavioral factors assessed in the first National Health and Nutrition Examination Survey and subsequent morbidity, mortality, and hospital utilization, as well as changes in risk factors, functional limitations and institution- alization. NHANES III: • estimates the national prevalence of selected diseases and risk factors • estimates national population reference distributions of selected health parameters • documents and investigates reasons for secular trends in selected diseases and risk factors • contributes to an understanding of disease etiology • investigates the natural history of selected diseases	**NHANES III--** Target diseases/conditions • Cardiovascular disease • COPD • Diabetes • Kidney disease • Gallbladder disease • Osteoporosis • Arthritis • Infectious diseases • Substance abuse • Dental health • Allergy • Cancer • Mental health • Hearing • Nutrition • Monitoring • Risk factors • Physical activity • Reproductive health • Tobacco • Child health • Health of older Americans • Occupational health • Environmental health • Longitudinal follow-up • Storage of biologic specimens

Colorado Pragmatic Research in Health (COPRH) Conference: **Breakout Session:** *Using population-based data for secondary analysis*

May 25, 2021   **Supplemental Handout**

Data Set/Source	Description	Variables
National Home Care and Hospice Survey (NHCHS) **NCHS**	NCHS (CDC) is a continuing series of surveys of home and hospice care agencies in the United States (probability sample). Information was collected about agencies that provide home and hospice care and about their current patients and discharges. The survey includes all types of agencies that provide home health and hospice care without regard to whether they are Medicare certified or whether they are licensed. Home health agencies and hospices are usually defined in terms of the type of care they provide. Home health care is provided to individuals and families in their place of residence for the purpose of promoting, maintaining, or restoring health or for maximizing the level of independence while minimizing the effects of disability and illness, including terminal illness. Hospice care is defined as a program of palliative and supportive care services providing physical, psychological, social, and spiritual care for dying persons, their families, and other loved ones. Hospitce services are available in both home and inpatient settings. Data are collected through personal interviews with administrators and staff.	Agency File • Agency identifier code • Number of current patients • Type of ownership • Affiliation • Certification status • Staff hours • Services available Current Patient File • Patient demographics • Current living arrangements • Referral source • Diagnoses at admission and at time of survey • Surgical and diagnostic procedures related to admission • Type of care received • Primary care giver/relationship • Aids used • Vision and hearing status • Activities of daily living • Instrumental activities of daily living • Services provided • Service providers • Number of visits • Amount billed for care/dates • Sources of payment Discharge Patient File – same as current patient file plus: • Living arrangements at discharge • Dxs at admission and discharge • Reason for discharge

Data Set	Description	Variables
National Health Interview Survey (NHIS)	NCHS (CDC) is the collection and analysis of morbidity data on health and disability, current major health issues, and conditions for the civilian non-institutionalized U.S. population by various socioeconomic and demographic characteristics. Continuous data collection since 1957, survey redesigned in 1997 to collect data on all household members through household interviews by US Census Bureau interviewers. Cross- sectional, complex multi-stage area probability sample design, and linkage to the National Death Index. E-coding for injuries, including medical and therapeutic misadventures, began in 1993. Since 1985, over sampling of black and Hispanic persons has been done in various data years.	• Housing characteristics • Family structure and living arrangements • Relationships/social contacts • Health care utilization • Health conditions/impairments • Functional status, assistance with basic activities • Occupation and retirement • Health opinions • Health insurance • Access to health care/transportation • Behaviors (tobacco, physical activity, alcohol) • Use of assistive devices/medical implants • Immunizations • AIDS

Colorado Pragmatic Research in Health (COPRH) Conference: **Breakout Session:** *Using population-based data for secondary analysis*

May 25, 2021   **Supplemental Handout**

Data Set/Source	Description	Variables
National Survey of Family Growth (NSFG) **NCHS**	NCHS (CDC) is a multipurpose survey based on personal interviews with a national sample of women aged 15 to 44 years in the civilian non-institutionalized population of the United States. Its main function is to collect data about factors affecting pregnancy and women's health in the United States. Studies were conducted in 1973, 1976, 1982, 1988, 1990, and in 1995.	• Number of children women have had/number they expect in the future • Intended and unintended births • Sexual intercourse • Marriage and cohabitation • Contraceptive use • Infertility, impaired fecundity, and sterilization operations • Breastfeeding, maternity leave, and child care • Adoption, stepchildren, and foster children • Health insurance coverage • Family planning and other medical services • Smoking by women 15-44 • HIV testing • Pelvic inflammatory disease • Sex education

Data Set/Source	Description	Variables	
American Community Survey **Census**	American Community Survey (ACS) is an ongoing survey that provides vital information on a yearly basis about our nation and its people. Information from the survey generates data that help determine how more than $675 billion in federal and state funds are distributed each year. Through the ACS, we know more about jobs and occupations, educational attainment, veterans, whether people own or rent their homes, and other topics. Public officials, planners, and entrepreneurs use this information to assess the past and plan the future. ACS data help communities plan for hospitals and schools, support school lunch programs, improve emergency services, build bridges, and inform businesses looking to add jobs and expand to new markets, and more.	• Age • Ancestry • Citizenship Status • Commuting (Journey to Work) and Place of Work • Disability Status • Educational Attainment and School Enrollment • Employment Status • Fertility • Grandparents as Caregivers • Health Insurance Coverage • Hispanic or Latino Origin • Income and Earnings • Industry, Occupation, and Class of Worker	/ • Language Spoken at Home • Marital History, Marital Status • Migration/Residence 1 Year Ago • Period of Military Service • Place of Birth • Poverty Status • Race • Relationship to Householder • Sex • Undergraduate Field of Degree • VA Service-Connected Disability • Status • Veteran Status • Work Status Last Year • Year of Entry