Large Database Resources for Research in Epidemiology, Health Services & Health Policy: Data Strengths and Limitations

Ellen P. McCarthy, PhD, MPH
Associate Professor of Medicine
Division of General Medicine and Primary Care
Beth Israel Deaconess Medical Center
Assistant Dean for Development and Diversity
Office for Diversity and Community Partnership
Harvard Medical School
Surgical Outcomes Analysis & Research
SOAR
September 8, 2014

No conflicts of interest to disclose

- Acknowledge Dr. John Ayanian’s major contributions to the development of this course on Research with Large Database course (HPM299) at the Harvard School of Public Health.

Today’s Objectives

- Introduce a wide range of Federal data sources
- How to access the data sources
- Advantages and disadvantages / Strengths and Limitations

Why should you consider using secondary databases in your research?

- Develop a population perspective in your work
- May answer research questions quickly & efficiently
- “Hypothesis generating” results help to focus subsequent primary data collection efforts
- Balance your research portfolio & complement studies that require primary data collection
- Enhance your skills in study design & statistical analysis
- Learning curve is on one database, but the skills you develop are readily applicable to many others

Primary Use of Federal Data Sources

- Vital statistics or disease surveillance
  - US Census
  - National Death Index
  - State cancer registries
  - Monitoring of reportable diseases
- National surveys to inform public policy and public health
Examples include:
  - National Ambulatory Medical Care (NAMCS)
  - National Health and Nutrition Examination Survey (NHANES)
  - Medicare Current Beneficiary Survey (MCBS)

Primary Use of Federal Data Sources

- Reimbursement and management of medical services
  - Medicare & Medicaid claims for fee-for-service patients
  - Statewide hospital discharge abstracts
  - Some states have claims from all payers
- Data to evaluate quality of care & outcomes (often disease specific)
  - National Surgical Quality Improvement Project (NSQIP)
  - Cancer registry data (SEER Program)
  - United States Renal Data System (USRDS)
### Types of Large Data

- Administrative data
- Registry data
- Linked data
- National survey data

### Public Use vs. Restricted Use Databases

**Public Use**
- Typically available for public use at little or no cost
- Readily downloaded from the internet
- Generally well-documented; programs to import the data into common statistical software
- IRB determination is almost always Exempt

**Restricted Use**
- Typically requires an application process and purchase of data
- Usually data provided on DVD or other medium
- Requires signed data use agreement (DUA), which specifies confidentiality requirements & security safeguards (i.e., how data must be stored, protected, and presented)
- IRB review is usually Expedited

---

### Healthcare Cost & Utilization Project (HCUP)

- Collection of state administrative databases maintained by Agency for Healthcare Research & Quality (AHRQ):
  - State Inpatient Databases (SID)
  - National Inpatient Sample (NIS)
  - Kid's Inpatient Database (KID)
  - State Ambulatory Surgery & Services Databases (SASD)
  - State Emergency Department Databases (SEDID)
  - Nationwide Emergency Department Sample (NEDS)

**HCUP website:** [www.ahrq.gov/data/hcup/](http://www.ahrq.gov/data/hcup/)

**Online Training:** [www.hcup-us.ahrq.gov/overviewcourse.jsp](http://www.hcup-us.ahrq.gov/overviewcourse.jsp)

---

### State Inpatient Databases (SID)

- Discharge abstracts from 47 states:
  - 26 million discharges/year
  - 97% of all U.S. discharges
  - 28 states make data available through AHRQ
- Useful for studying single state, comparing two or a few states, or small-area variations
- Available for 1990-2011 (& 2012 for most states)
- Cost of $35/year (Arizona, Maryland) to $1,610/year (Utah)
National Inpatient Sample
NIS
• Largest all-payer inpatient database in US
  • ~8 million discharges/year
  • In 2012, ~7.2 million discharges from ~4,300 hospitals in 44 States
  • Covers ~95% of US population
• Useful for national estimates and trends, regional comparisons, large samples
  • Available for 1988-2012
  • Cost of $350/year ($50/year for students)

Kids’ Inpatient Database
KID
• Largest database on hospital care of newborns, children & adolescents (<age 21)
  • Drawn from SID data
  • 2 to 3 million discharges per year
  • ~4,100 hospitals in 44 states in 2012
    • (22 states in 1997, 36 in 2003)
  • AHA data on hospital characteristics since 2003
  • Disease severity measures since 2003
  • Cost of $350/year ($50/year for students)

HCUP Studies

NIS & KID
Data Elements
• Age, sex, race/ethnicity, payer, zip code income
• Hospital ID number, encrypted physician/surgeon ID
• Admission & discharge dates, total and ICU LOS, days from admission to procedures
• Admission source & type and discharge disposition
• Diagnosis & procedure codes (1-15), DRG
• Total charges
• Birth weight (KID only) and neonatal/maternal flag

NIS & KID
Data Use and Availability
• No disclosure of individual patients or cells <10 subjects
• No individual-level linkages to other datasets
• No sharing of dataset with other researchers or groups
• No proprietary or commercial use of data
• Data stored securely with access limited to authorized users
• HCUP Data Use Agreement web-based training required: www.hcup-us.ahrq.gov/tech_assist/dua.jsp
State Ambulatory Surgery & Services Databases SASD

- Same-day surgery abstracts from 18 states available through AHRQ in 2011

Useful for:
- Analyses of same-day surgeries
- Trends in outpatient surgery utilization, access, charges, insurance type, and outcomes
- Shift in surgical procedures to outpatient facilities
- Diffusion of medical technology

- Available for 1997-2011 (2012 available for some states)
- $35 - $1,610 per state per year

HCUP Resources AHRQ Quality Indicators

- Prevention quality indicators (PQIs): 16 ambulatory care-sensitive conditions
- Inpatient quality indicators (IQIs):
  - 15 medical & surgical mortality measures
  - 11 utilization measures
  - 6 volume measures
- Patient safety indicators (PSIs): 20 inpatient complications
- Pediatric quality indicators (PDIs): 18 indicators
- Downloadable for free from AHRQ website: http://qualityindicators.ahrq.gov/

Medicare Data

- Available only for fee-for-service beneficiaries
- Medicare claims: diagnoses, procedures, charges
  - Part A: inpatient, home health care, hospice services
  - Part B: physician services, outpatient services
  - Part C: Medicare Advantage – claims generally not available
  - Part D: prescription drugs (added in 2006)

- Enrollment files: demographics, eligibility, HMO enrollment, vital status/date of death
- Possible linkages: provider of services files, Medicare Physician Identification and Eligibility Records (MPIER) (e.g., physician specialty), AMA, Area Resource File

Medicare Advantage

- National Average of 30% in 2014

- Share of Medicare Beneficiaries Enrolled in Medicare Advantage Plans, by State, 2014

  51% in MN
  3% in WY
  <1% in AK
Racial Disparities in Abdominal Aortic Aneurysm Repair Among Male Medicare Beneficiaries

Outcomes for Mitral Valve Surgery Among Medicare Fee-for-Service Beneficiaries, 1999 to 2009

John A. Dikkes, MD, Yan Wu, PhD, Maria M. Doss, PhD, MPH
Tsuyoshi Seki, MD, PhD, Carlos J. Toto Ros, MD, PhD
Sandeep K. Khosla, MD, Jamal M. Idrees, MD, MPH

Outcomes for Whites and Blacks at Hospitals That Disproportionately Care for Black Medicare Beneficiaries

Murray Eaton and Todd D. Joffe

All Payer Claims Data

Administrative Data

Limitations

- Always less clinical data than you would like
- Lack data on functional status and disease severity
- Difficult to distinguish complications from comorbidity conditions
- Must understand basic inclusion & exclusion criteria (e.g., lack data on managed care patients) as well as coverage/reimbursement of key factors under study
- Potential for recording bias
- Chronic conditions may be undercoded in severely ill patients
- Procedures that are not well-reimbursed may be undercoded in favor of those with better reimbursement
- Longitudinal analyses can be limited by:
  - Changing eligibility (e.g., Medicaid, switch to managed care insurance)
  - Inconsistent identifiers across settings of care

Strengths

- Population-based with very large numbers
- Many can provide longitudinal, person-level histories of clinical and health care use over time and across settings of care
- Well-developed, publicly available algorithms for comorbidity conditions and risk adjustment
- Conduct robust analyses (e.g., subgroup analyses, assessment of confounding and interaction, multi-level modeling)

Potential Pitfalls

- MUST take caution with strength in numbers
- Excellent statistical power, but beware of spurious results
- Specify ahead of time what “clinically meaningful” means to your study!
- Start with a hypothesis-driven research question and pre-specified analysis plan
- Avoid fishing (data mining)
- Take caution with multiple testing
- Unmeasured confounders nearly always exist
- Ask questions that are interesting regardless of whether the results are positive or negative!
  - Just because you can answer it, doesn’t mean it is interesting!
  - “Bad” papers often take longer and are more time consuming than “good” papers
Types of Large Data

- Administrative data
- Registry data
- Linked data
- National survey data

National Surgical Quality Improvement Project (NSQIP)

http://site.acsnsqip.org/
American College of Surgeons
- Implemented to measure and improve quality of surgical care across surgical specialties in private hospitals nationally validated, risk-adjusted outcomes
- Identify opportunities to improve quality and outcomes in general and vascular surgery (e.g., reduce complications of care, reduce disparities in care, improve mortality)
- List of participating hospitals

NSQIP Sample ~ 1.6 million patients from 1991-present

Includes all “major cases” done under general, spinal or epidural
- General and vascular surgeries first 40 consecutive cases over 8-day period (46 cycles per year)
- ~1,600 cases from high-volume hospitals & minimum of 900 cases from low volume hospitals
- Subspecialties: gynecologic, neurologic, cardiac, thoracic, orthopaedic, otolaryngologic, plastic, urologic

Targeted Procedures (regardless of anesthesia)
- Carotid endarterectomy
- Inguinal herniorrhaphy
- Parathyroidectomy
- Thyroidectomy
- Breast lumpectomy
- Endovascular AAA repair

NSQIP Data Elements

Clinical data abstracted by Surgical Clinical Reviewer (SCR)
- Preoperative: demographics, 30 clinical variables, 12 laboratory variables
- Intraoperative: 15 clinical variables
- Postoperative: 10 laboratory variables, 21 categories of postoperative morbidity, 30-day postoperative morbidity and mortality, length of stay in hospital
- ~95% capture of 30-day outcomes
- Various methods (e.g., medical record, follow-up office visits, phone calls, patient letters)
- Hospitals with follow-up rates <80% excluded

NSQIP Limitations

Limitations due to privacy safeguards and resource constraints
- Dates of surgery are not released, other dates are recoded to durations (e.g., length of stay, lab tests)
- No provider or facility identifiers, no geographic information
- Patients age 90 and older grouped together
- Variables tend to be more generic in nature due to wide range of surgeries may lack data important to your surgery of interest
- Most patients do not receive all possible preop lab tests (careful of missing data)
- Does not capture preventive measures
NSQIP Access and Availability

- 2012 Participant Use File (PUF)
- 543,885 cases submitted from 374 sites
- Contains patient level, aggregate data does not include identifying information on hospitals, healthcare providers, or patients
- PUF excludes hospitals with <80% follow-up on 30-day measures and >5% inter-rater reliability
- Official appointment at an enrolled site, signed Data Use Agreement (DUA)
  
  http://www.acsnsqip.org/acsNsqipData/jsp/pub/useAgreement.jsp

National Trauma Data Bank (NTDB)
American College of Surgeons

- Largest Trauma Registry in United States
- Nationally representative sample of trauma centers in the US requires survey procedures
- Admission and discharge status, patient demographics, payment method, injury and diagnosis (mechanism, e-code, ICD-9 or AIS code), procedure codes, injury severity scores (ISS, GCS), and outcome variables (e.g., LOS, ICU days)
- In 2012, ~833,000 valid trauma diagnosis from 805 hospitals
- Requires approval from ACS Committee on Trauma, signed Data Use Agreement (DUA), costs about $300 per year
  
  http://www.facs.org/trauma/ntdb/ntdbapp.html

Surveillance, Epidemiology, and End Results ~ SEER Program

- Epidemiologic surveillance system
- Developed by the National Center Institute in 1973
- Population-based tumor registries
- Covers ~26% of the U.S. population
- Captures ~99% of cancers diagnoses
- Priority on racial and ethnic diversity across sites
  
  Released annually in April
  
  Data use agreement required
  
  Available on DVD or by downloading via FTP

SEER Tumor Registries

- Connecticut
- Iowa
- New Mexico
- Utah
- Hawaii
- Detroit, Michigan
- Atlanta, Georgia
- Seattle-Puget Sound, Washington
- San Francisco-Oakland, California

Established in 1992 (primarily Hispanics)
- San Jose-Monterey, California
- Los Angeles County, California

Added in 2001
- California (remainder), Kentucky, Louisiana, New Jersey

Special Populations
- Rural Georgia (10 counties, predominantly African Am) (added in 1978)
- Native American populations in Arizona (added in 1980), Cherokee Nation in Oklahoma (added in 1997)

SEER Tumor Registries

http://seer.cancer.gov/registries/
Generalizability of SEER compared to US population

- SEER contracts with nonprofit organizations in coverage areas
- Identify incident cancers, except some skin cancers
- Abstract records for every resident cancer patient
  - hospitals in and out of coverage area
  - private labs, radiotherapy units, nursing homes
  - death certificates where cancer is leading cause of death

SEER Program Case Ascertainment

- Population-based & nearly complete case ascertainment
- Captures second primary tumors
- Complete and valid data on treatments with surgery and radiation therapy
- Enriched with diverse and immigrant populations
- Excellent for trend analyses in incidence and mortality

SEER Program Types of Data

- Demographics: Age at diagnosis, race/ethnicity, marital status, birthplace, county of residence
- Cancer site: Primary or secondary site, numbers and sequence of primaries, laterality, diagnosis date, diagnostic confirmation
- Tumor characteristics: Stage at diagnosis, tumor size, extent of disease, numbers of positive nodes, histology and grade
- Initial treatment: Site-specific surgery, radiation, surgery-RT sequence
- Mortality: Follow-up time, status, cause of death

SEER Program Strengths

- Population-based & nearly complete case ascertainment
- Captures second primary tumors
- Complete and valid data on treatments with surgery and radiation therapy
- Enriched with diverse and immigrant populations
- Excellent for trend analyses in incidence and mortality

SEER Program Limitations

- Limited to information around time of diagnosis with the exception of date and cause of death
- Does not capture metastases following diagnosis
- Recurrence
- Comorbid illness
- Socioeconomic status or income
- Health insurance or access to care
- Does not release information on chemotherapy use
Types of Large Data

- Administrative data
- Registry data
- Linked data
- National survey data

SEER-Medicare Database

- SEER cases through 2009 linked with Medicare data through 2010
- Medicare enrollment information
- Medicare claims for inpatient, outpatient, provider, home care and hospice services
- Medicare prescription drug data (starting 2007)
- Census indicators of income and education of census tract/zip code of residence
- 5% sample of non-cancer beneficiaries

Additional Sources of “Linked” Data

- Medicare Current Beneficiary Survey
  http://www.resdac.org/MCBS/data_available.asp
- Health and Retirement Survey
  http://hrsonline.isr.umich.edu/
- United States Renal Data System (USRDS) Database
  http://www.usrds.org/products.asp

Types of Large Data
Federal Survey Data ~ Advantages

- Someone else has done the hard work → recent data with large sample sizes
- Rigorous methods: study design, sampling, instrument development & validation
- Oversample important populations (e.g., racial/ethnic minorities, persons with disabilities)
- Nationally-representative data when analyzed properly
- Many surveys feature core questions that are consistent from year to year → excellent for studies of:
  - Trends (e.g., changes over time, impact of changes in policy or guidelines)
  - Rare diseases or outcomes
  - Vulnerable populations

Federal Survey Data ~ Disadvantages

- Not specifically designed to answer your question of interest
- May need to be creative to define variables
- May lack important information (e.g., confounders, disease severity)
- Limited to no ability to assess non-respondents
- Limitations to research involving non-English, non-Spanish speakers
- Limitations to aging research
- Many surveys limited to community-dwellers
- Upper age limit is usually set at 85+ years

Questions