

Data Sources for Epidemiological Studies

Xian Wu

Division of Biostatistics and Epidemiology
Department of Healthcare Policy and Research



10.10.17

Agenda for today

- **Scientific** and **operational** considerations involved in planning a epidemiological study
- Data sources in epidemiological studies
- scenarios in which different data sources are best used in epidemiological studies



Feasibility assessments

- **Critical first step to ensure scientific and operational integrity of a study**
- **Ideal study to address a given research question is often not wholly feasible**
- **Purpose**
 - **Characterize circumstances in which it is feasible to address research question**
 - **Identify trade-offs between scientific and operational considerations**

Scientific considerations

- **Outline ideal study to address a given research question**
 - **Define study objectives**
 - **Identify key data elements**
 - **Exposure of interest**
 - **Outcome of interest**
 - **Population**
 - **Statistical measure**
 - **Timeframe**
 - **Determine study design (descriptive studies v.s. analytic studies)**
 - **Subjects selected according to exposure (eg, cohort study)**
 - **Subjects selected according to outcome (eg, case-control study)**
 - **Subjects selected according to neither exposure nor outcome (eg, cross-sectional study)**
 - **Estimate sample size requirement**



Operational considerations

- **Identify potential data sources**
 - identify data source with sufficient number of patients who meet key inclusion and exclusion criteria (eg, diagnosed with indication or treated with drug of interest)
- **Requirements of review/approval by Institutional Review Boards (IRBs) and Clinical Study Evaluation Committee (CSEC)**
- **Time/funding**
 - Typical timelines for local regulatory/ethics approvals



Data sources used in epidemiological studies

Types of data sources

- **Primary data sources**

- Data directly collected from study participants for the purposes of the study

- **Secondary data sources**

- Data are collected from existing health care databases or medical records, where all of the events of interest have already occurred at the time of data are queried

- Collected for administrative/reimbursement purposes by insurance provider, as clinical data by general practitioner, or as part of universal healthcare coverage



Advantages of primary data sources

- Data collection is tailored to study objectives, eg:
 - Focus on measurement of confounders
 - Availability of lab data
 - Capture of less severe diagnoses
 - Indication for medication use more explicit
 - Capture of inpatient medications, over-the-counter medications, and medications taken on as-needed basis
 - Can obtain information on clinical assessments needed for valid measurement but not universally performed as standard of care

Disadvantages of primary data sources

- Expensive and time-intensive
- May be infeasible for studies requiring large sample sizes or long follow-up
- Many operational considerations, eg:
 - Subject informed consent
 - Identification, initiation, and management of study sites
 - Data monitoring

Types of secondary data sources

- **Unstructured data**

- Data do not already exist in a structured (ie, coded) database
- Information from individual patient medical records must be abstracted and converted into structured data for study purposes

- **Structured data**

- Data already exist in a structured (ie, coded) database
- eg, administrative claims database, registries, surveys.

- **Hybrid data**

- Data already existing in a structured (ie, coded) database are supplemented by unstructured data
 - Text fields (eg, physician notes) in the database or medical record information are reviewed, categorized/coded, and added to the structured database
 - Natural language processing: algorithm-based approach to identify relevant text from unstructured data contribute to coded fields

Data sources for different epidemiological studies

- **Clinical epidemiology/ Pharmacoepidemiology**
 - Administrative claims database
 - Clinical registries
- **Cancer epidemiology**
 - Surveillance, Epidemiology, and End Results Program (SEER)
 - SEER-Medicare linked database (Medicare beneficiaries with cancer)
 - National Cancer Database (NCDB)
- **Social epidemiology**
 - National Health and Nutrition Examination Survey (NHANES)
<https://www.cdc.gov/nchs/nhanes/index.htm>
 - Behavioral Risk Factor Surveillance System (BRFSS)
<https://www.cdc.gov/brfss/index.html>
 - NYC Community Health Survey (NYC CHS)
<https://www1.nyc.gov/site/doh/data/data-sets/community-health-survey-public-use-data.page>



Clinical epidemiology/ Pharmacoepidemiology

- **Administrative claims databases**
 - eg, government insurance programs, private insurance companies, provincial health plans
 - Generally in US and Canada
- **Electronic medical record-based databases, healthcare registries and record linkage systems**
 - eg, general practitioner-based data sources, population-based registries
 - few in US, many in Europe





HCUP User Support (HCUP-US)

The HCUP (pronounced "H-CUP") family of health care databases and related software tools and products is made possible by a Federal-State-Industry partnership sponsored by the Agency for Healthcare Research and Quality (AHRQ)

The Healthcare Cost and Utilization Project (HCUP, pronounced "H-Cup") is a family of health care databases and related software tools and products developed through a **Federal-State-Industry partnership** and sponsored by the Agency for Healthcare Research and Quality (AHRQ). HCUP databases bring together the data collection efforts of **State data organizations, hospital associations, private data organizations, and the Federal government** to create a national information resource of encounter-level health care data (HCUP Partners). HCUP includes **the largest collection of longitudinal hospital care data in the United States**, with **all-payer**, encounter-level information beginning in 1988.





HCUP User Support (HCUP-US)

The HCUP (pronounced "H-CUP") family of health care databases and related software tools and products is made possible by a Federal-State-Industry partnership sponsored by the Agency for Healthcare Research and Quality (AHRQ)

The National (Nationwide) Inpatient Sample (NIS) containing data on more than seven million hospital stays each year.

The Kids' Inpatient Database (KID) is a nationwide sample of pediatric inpatient discharges.

The Nationwide Emergency Department Sample (NEDS) is a database that yields national estimates of emergency department (ED) visits.

The Nationwide Readmissions Database (NRD) is a unique and powerful database designed to support various types of analyses of national readmission rates for all payers and the uninsured.

The State Inpatient Databases (SID) contain the universe of inpatient discharge abstracts from participating states.

The State Ambulatory Surgery and Services Databases (SASD) include data for ambulatory surgery and other outpatient services from hospital-owned facilities.

The State Emergency Department Databases (SEDD) contain data from hospital-affiliated emergency departments for visits that do not result in hospitalizations.



Category	Database	Pricing	2014	2013	2012	2011	2010	2009	2008
Nationwide	NIS	All Others	☐ \$500	☐ \$350	☐ \$350	☐ \$350	☐ \$350	☐ \$350	☐ \$350
Nationwide	NIS	Students	☐ \$100	☐ \$100	☐ \$50	☐ \$50	☐ \$50	☐ \$50	☐ \$50
Nationwide	KID	All Others	NA	NA	☐ \$350	NA	NA	☐ \$350	NA
Nationwide	KID	Students	NA	NA	☐ \$50	NA	NA	☐ \$50	NA
Nationwide	NEDS	All Others	☐ \$750	☐ \$500	☐ \$500	☐ \$500	☐ \$500	☐ \$500	☐ \$500
Nationwide	NEDS	Students	☐ \$150	☐ \$150	☐ \$75	☐ \$75	☐ \$75	☐ \$75	☐ \$75
Nationwide	NRD	All Others	☐ \$750	☐ \$500	☐ \$500	☐ \$500	☐ \$500	NA	NA
Nationwide	NRD	Students	☐ \$150	☐ \$150	☐ \$150	☐ \$150	☐ \$150	NA	NA

New Jersey	SID	All Applicants	NA	☐ \$175	☐ \$160	☐ \$160	☐ \$160	☐ \$160	☐ \$160
New Jersey	SASD	All Applicants	NA	☐ \$175	☐ \$160	☐ \$160	☐ \$160	☐ \$160	☐ \$160
New Jersey	SEDD	All Applicants	NA	☐ \$175	☐ \$160	☐ \$160	☐ \$160	☐ \$160	☐ \$160
New Mexico	SID	All Applicants	NA	☐ \$500 (R)	☐ \$485 (R)	☐ \$485 (R)	☐ \$485 (R)	☐ \$485 (R)	☐ \$485 (R)
New York	SID	AHRQ Grantee	NA	☐ \$400 (R)	☐ \$385 (R)	☐ \$385 (R)	☐ \$385 (R)	☐ \$385 (R)	☐ \$385 (R)
New York	SID	All Others	NA	☐ \$750 (R)	☐ \$735 (R)	☐ \$735 (R)	☐ \$735 (R)	☐ \$735 (R)	☐ \$735 (R)
New York	SID	Not-For-Profit Affiliation	NA	☐ \$400 (R)	☐ \$385 (R)	☐ \$385 (R)	☐ \$385 (R)	☐ \$385 (R)	☐ \$385 (R)
New York	SID	Students	NA	☐ \$200 (R)	☐ \$185 (R)	☐ \$185 (R)	☐ \$185 (R)	☐ \$185 (R)	☐ \$185 (R)
New York	SASD	AHRQ Grantee	NA	☐ \$300 (R)	☐ \$285 (R)	☐ \$285 (R)	☐ \$285 (R)	☐ \$285 (R)	☐ \$285 (R)
New York	SASD	All Others	NA	☐ \$550 (R)	☐ \$535 (R)	☐ \$535 (R)	☐ \$535 (R)	☐ \$535 (R)	☐ \$535 (R)
New York	SASD	Not-For-Profit Affiliation	NA	☐ \$300 (R)	☐ \$285 (R)	☐ \$285 (R)	☐ \$285 (R)	☐ \$285 (R)	☐ \$285 (R)
New York	SASD	Students	NA	☐ \$150 (R)	☐ \$135 (R)	☐ \$135 (R)	☐ \$135 (R)	☐ \$135 (R)	☐ \$135 (R)
New York	SEDD	AHRQ Grantee	NA	☐ \$550 (R)	☐ \$535 (R)	☐ \$535 (R)	☐ \$535 (R)	☐ \$535 (R)	☐ \$535 (R)
New York	SEDD	All Others	NA	☐ \$1,050 (R)	☐ \$1,035 (R)	☐ \$1,035 (R)	☐ \$1,035 (R)	☐ \$1,035 (R)	☐ \$1,035 (R)
New York	SEDD	Not-For-Profit Affiliation	NA	☐ \$550 (R)	☐ \$535 (R)	☐ \$535 (R)	☐ \$535 (R)	☐ \$535 (R)	☐ \$535 (R)
New York	SEDD	Students	NA	☐ \$300 (R)	☐ \$285 (R)	☐ \$285 (R)	☐ \$285 (R)	☐ \$285 (R)	☐ \$285 (R)



Medicare

Medicaid/CHIP

Medicare-Medicaid
Coordination

Private
Insurance

Innovation
Center

Regulations &
Guidance

Research, Statistics,
Data & Systems

Outreach &
Education

[Home](#) > [Research, Statistics, Data and Systems](#) > [Files for Order - General Information](#) > [Files for Order - General Information](#)

Files for Order - General Information

[Identifiable Data Files](#)

[Limited Data Set \(LDS\) Files](#)

[Non-Identifiable Data Files](#)

Files for Order - General Information

The Centers for Medicare & Medicaid Services (CMS) makes certain data files available for order. These data files include information collected through the operation of the Medicare and Medicaid programs. There are three types of data made available.

[Identifiable Data Files](#)

Identifiable data files (IDFs) are available to certain stakeholders as allowed by federal laws and regulations as well as CMS policy. IDFs contain protected health information (PHI) and/or personally identifiable information (PII) and CMS is committed to ensuring this information is protected.

[Limited Data Sets](#)

Limited Data Sets (LDS) are available to certain stakeholders as allowed by federal laws and regulations as well as CMS policy. LDS contain beneficiary level health information but exclude specific direct identifiers as outlined in the Health Insurance Portability and Accountability Act of 1996 (HIPAA). LDS are considered identifiable even though they do not contain Personally Identifiable Information (PII).

[Non-Identifiable Data Files](#)

Non-Identifiable Data Files are available to stakeholders for order. Non-Identifiable Data Files do not contain any protected health information (PHI) or personally identifiable information (PII).

Please use the navigation bar on the left to view information about the three categories of data available, including: Identifiable Data Files; Limited Data Sets; and, Non-Identifiable Data Files.

Related Links

[CMS Data Payment Form on Pay.gov](#)

[CMS Data Navigator](#)

Page last Modified: 01/10/2017 9:15 AM

[Help with File Formats and Plug-Ins](#)

LDS PRICING and REQUEST ORDER FORM

File List - Select the files and years you would like by specifying 5% or 100% in appropriate cells.	Running Total all Files: \$0																	Price per Year	
	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	COST			5%	100%
Denominator (Annual) File 2006 - 2016	N/A												N/A	N/A				\$250	\$1,000
To order the QUARTERLY Denominator (MBSF) file, see SAF Quarterly tab ▶ QTR	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	-	-		N/A					
Master Beneficiary Summary (Annual) File Begins w/2016	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A		N/A	N/A				\$250	\$1,000
To order the QUARTERLY Denominator (MBSF) file, see SAF Quarterly tab ▶ QTR	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	-	-		N/A					
Carrier Standard Analytic File - Annual	N/A												N/A	N/A				\$1,700	N/A
To order the QUARTERLY Carrier file, go to the SAF Quarterly tab ▶ QTR	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A				N/A					
Durable Medical Equipment Standard Analytic File - Annual	N/A												N/A	N/A				\$800	N/A
To order the QUARTERLY DME file, go to the SAF Quarterly tab ▶ QTR	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A				N/A					
Home Health Standard Analytic File - Annual	N/A												N/A	N/A				\$300	\$2,000
To order the QUARTERLY HHA file, go to the SAF Quarterly tab ▶ QTR	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A				N/A					
Hospice Standard Analytic File - Annual	N/A												N/A	N/A				\$300	\$1,000
To order the QUARTERLY Hospice file, go to the SAF Quarterly tab ▶ QTR	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A				N/A					
Inpatient Standard Analytic File - Annual	N/A												N/A	N/A				\$400	\$3,000
To order the QUARTERLY Inpatient file, go to the SAF Quarterly tab ▶ QTR	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A				N/A					
Outpatient Standard Analytic File - Annual	N/A												N/A	N/A				\$1,000	\$7,000
To order the QUARTERLY Outpatient file, go to the SAF Quarterly tab ▶ QTR	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A				N/A					
Skilled Nursing Facility Standard Analytic File - Annual	N/A												N/A	N/A				\$300	\$1,000
To order the QUARTERLY SNF file, go to the SAF Quarterly tab ▶ QTR	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A				N/A					
Provider Master Crosswalk - (must submit DUA/FormB) *see note below	N/A	N/A	N/A	N/A	N/A	N/A	N/A		N/A	N/A	N/A	N/A	N/A	N/A				N/A	\$0
(OPPS) Supplemental File - *see note below	N/A	N/A	N/A	N/A	N/A							N/A	N/A	N/A				N/A	\$0
Inpatient Psychiatric Prospective Payment System (IPF PPS)	N/A	N/A	N/A	N/A					N/A	N/A	N/A	N/A						N/A	\$3,000



Other examples of administrative databases

Examples of Administrative Healthcare Databases in US and Canada

Database	Characteristics	Eligible Population
US		
Group Health Cooperative, Washington	HMO	460,000
Kaiser Permanente, Northern California	HMO	2.8 million
Kaiser Permanente, NW Division	HMO	430,000
Harvard Pilgrim Health Care, New England	HMO	1.5 million
Tennessee Medicaid Database	Health insurance for recipients of social welfare	1.4 million
New Jersey Medicaid Database	Health insurance for recipients of social welfare	700,000
Veterans Affairs Database	US veterans	6.1 million
Pharmetrics	26 HMOs	60 million
Healthcore	Recipients of health insurance plans	34 million
United Healthcare	Recipients of health insurance plans	25 million
Canada		
Saskatchewan Health Database, Saskatchewan, Canada	Provincial health plan	1 million
RAMQ Database, Quebec, Canada	Provincial health plan for elderly	750,000
Ontario Health Insurance, Canada	Provincial health plan for elderly	1.4 million



Examples of EMR databases and registries

Examples of European Medical Record Databases, Healthcare Registries and Insurance Plans

Database	Country	Characteristics	Eligible Population
General Practitioner Databases			
GPRD	England	GP database	5 million
THIN	England	GP database	2.7 million
IPCI	Netherlands	GP database	1 million
PHARMO Record Linkage System	Netherlands	GP database	2 million
Tayside MEMO	Scotland	GP database	400,000
HSD-Thales	Italy	GP database	800,000
Healthcare Registries			
Denmark	Denmark	Healthcare registries	Maximum 5 million
Sweden	Sweden	Healthcare registries	Maximum 10 million
Other			
Bremen Institute of Prevention	Germany	Statutory health insurance recipients	13 million

1: General Practice Research Database 2: The Health Information Network 3: Integrated Primary Care Information
4: Medicines Monitoring Unit 5: Health Services Database

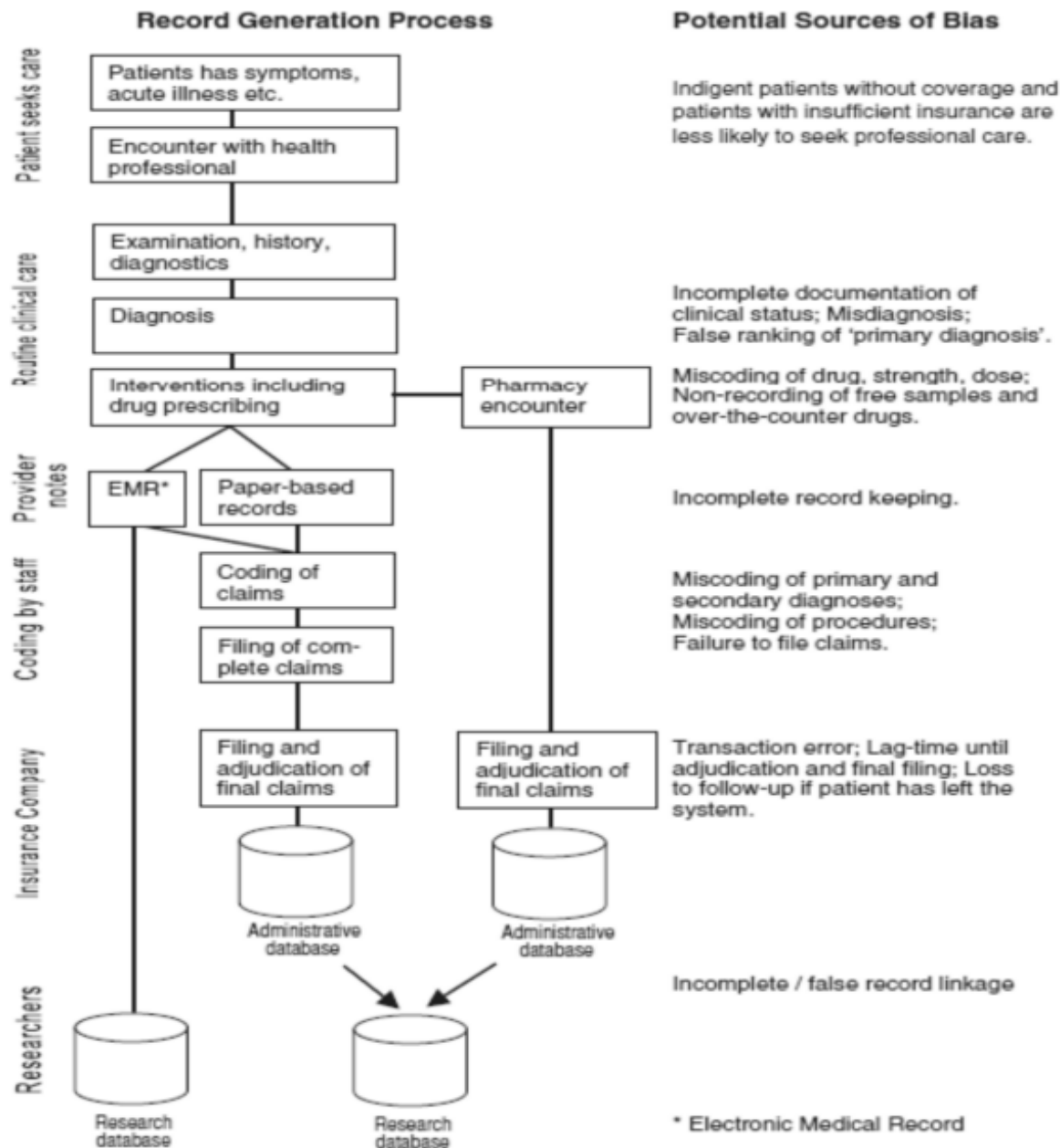


Examples of coding systems

Example of coding systems		
Overview of Coding Schemes Useful in Secondary Database Research		
Coding Scheme	Content	Comments
International Classification of Disease (ICD)	Diseases and procedures	ICD-9-CM is used for coding diagnoses and procedures, ICD-10 is used for causes of death, ICD-10-CM is under development; overseen by the World Health Organization, maintained in the United States by the National Center for Health Statistics
Current Procedural Terminology (CPT)	Products, services, and some drugs	Maintained by the American Medical Association, the 4th edition is most current; includes services performed by providers as well as drugs administered during provision of care
Healthcare Common Procedure Coding System (HCPSC)	Products and services	Maintained by the Centers for Medicare and Medicaid Services; covers products and services not in the CPT
National Drug Code (NDC)	Drugs	Maintained by the U.S. Food and Drug Administration
American Hospital Formulary Service (AHFS)	Drugs	Published and maintained by the American Society of Health-System Pharmacists
Anatomical Therapeutic Chemical Classification (ATC)	Drugs	Maintained by the World Health Organization

ICD-9-CM = International Classification of Disease, Ninth Revision, Clinical Modification; ICD-10-CM=ICD-CM, Tenth Revision.





ICD Code Problems

- **Errors in coding can occur:**
 - Improper documentation in the medical record**
 - Lack of documentation by provider**
 - Medical record coder
 - inexperience**
 - miscoding****
 - Unbundling (assignment of codes for each part of a diagnosis- instead of the overall diagnosis)**
 - Upcoding (assignment of codes for higher reimbursement over codes for less reimbursement)**



Example of record linkage

Patient Characteristics						
Patient ID	Patient Acct	Admit Date	LOS	Age	Gender	Discharge
12AB9809	803234	26JUN2007	7	62	Male	Alive
26BW6722	805335	27JUL2007	23	88	Female	Dead
12AB9809	813553	01AUG2007	3	62	Male	Dead
76CP0351	829781	02AUG2007	5	48	Female	Alive

Laboratory Data					
Patient ID	Patient Acct	Test	Value	Report	Test Date
12AB9809	803234	Potassium	5.5	High	26JUN2007
12AB9809	803234	Sodium	136	Normal	26JUN2007
12AB9809	803234	Creatinine	1.9	High	26JUN2007
12AB9809	803234	Glucose	108	Normal	26JUN2007
26BW6722	805335	Creatinine	0.9	Normal	28JUL2007
26BW6722	805335	Glucose	120	High	28JUL2007
12AB9809	813553	Creatinine	1.2	Normal	01AUG2007
12AB9809	813553	Creatinine	1.4	High	02AUG2007

Diagnosis Data		
Patient ID	Patient Acct	Primary Diagnosis
12AB9809	803234	Pneumonia
12BW6722	805335	Myocardial infarction
12AB9809	813553	Urinary tract infection
76CP0351	829781	Pneumonia


Pharmacy Data							
Patient ID	Patient Acct	Drug	Mg Dose	Route	Frequency	Start	Stop
12AB9809	803234	Quinapril	20	PO	BID	26JUN2007	02JUL2007
12AB9809	803234	Ceftriaxone	1000	IV	Q24H	28JUN2007	29JUN2007
12AB9809	803234	Ibuprofen	600	PO	Q6HP	26JUN2007	02JUL2007
26BW6722	805335	Metformin	500	PO	BID	28JUL2007	18AUG2007
26BW6722	805335	Simvastatin	20	PO	Q24H	28JUL2007	18AUG2007
12AB9809	813553	Quinapril	20	PO	BID	01AUG2007	03AUG2007
12AB9809	813553	Morphine	5	IV	Q2HP	01AUG2007	03AUG2007
12AB9809	813553	Cefepime	1000	IV	12H	01AUG2007	03AUG2007
76CP0351	829781	Azithromycin	500	PO	Q24H	02AUG2007	06AUG2007
76CP0351	829781	Acetaminophen	650	PO	Q6HP	02AUG2007	06AUG2007

Combined Dataset								
Patient ID	Patient Acct	Primary Diagnosis	LOS	Age	Gender	Discharge	Drug	Value
12AB9809	813553	Urinary tract infection	3	62	Male	Dead	Cefepime	1.2
76CP0351	829781	Pneumonia	5	48	Female	Alive	Azithromycin	

Original Investigation

Traumatic Spinal Cord Injury in the United States, 1993-2012

Nitin B. Jain, MD, MSPH; Gregory D. Ayers, MS; Emily N. Peterson, MS; Mitchel B. Harris, MD; Leslie Morse, DO; Kevin C. O'Connor, MD; Eric Garshick, MD, MOH

 Supplemental content at jama.com

IMPORTANCE Acute traumatic spinal cord injury results in disability and use of health care resources, yet data on contemporary national trends of traumatic spinal cord injury incidence and etiology are limited.

OBJECTIVE To assess trends in acute traumatic spinal cord injury incidence, etiology, mortality, and associated surgical procedures in the United States from 1993 to 2012.

DESIGN, SETTING, AND PARTICIPANTS Analysis of survey data from the US Nationwide Inpatient Sample databases for 1993-2012, including a total of 63 109 patients with acute traumatic spinal cord injury.

MAIN OUTCOMES AND MEASURES Age- and sex-stratified incidence of acute traumatic spinal cord injury; trends in etiology and in-hospital mortality of acute traumatic spinal cord injury.

RESULTS In 1993, the estimated incidence of acute spinal cord injury was 53 cases (95% CI, 52-54 cases) per 1 million persons based on 2659 actual cases. In 2012, the estimated incidence was 54 cases (95% CI, 53-55 cases) per 1 million population based on 3393 cases (average annual percentage change, 0.2%; 95% CI, -0.5% to 0.9%). Incidence rates among the younger male population declined from 1993 to 2012: for age 16 to 24 years, from 144 cases/million (2405 cases) to 87 cases/million (1770 cases) (average annual percentage change, -2.5%; 95% CI, -3.3% to -1.8%); for age 25 to 44 years, from 96 cases/million (3959 cases) to 71 cases/million (2930 cases), (average annual percentage change, -1.2%; 95% CI, -2.1% to -0.3%). A high rate of increase was observed in men aged 65 to 74 years (from 84 cases/million in 1993 [695 cases] to 131 cases/million [1465 cases]; average annual percentage change, 2.7%; 95% CI, 2.0%-3.5%). The percentage of spinal cord injury associated with falls increased significantly from 28% (95% CI, 26%-30%) in 1997-2000 to 66% (95% CI, 64%-68%) in 2010-2012 in those aged 65 years or older ($P < .001$). Although overall in-hospital mortality increased from 6.6% (95% CI, 6.1%-7.0%) in 1993-1996 to 7.5% (95% CI, 7.0%-8.0%) in 2010-2012 ($P < .001$), mortality decreased significantly from 24.2% (95% CI, 19.7%-28.7%) in 1993-1996 to 20.1% (95% CI, 17.0%-23.2%) in 2010-2012 ($P = .003$) among persons aged 85 years or older.



SHORT REPORT

Open Access

Emergency Department Visits for Heat Stroke in the United States, 2009 and 2010

Xian Wu¹, Joanne E Brady^{1,2}, Henry Rosenberg^{3,4} and Guohua Li^{1,2,5*}

Abstract

Background: The effect of extreme heat on health has become a growing public health concern due to climate change. We aimed to examine the epidemiological patterns of hospital-based emergency department (ED) visits for heat stroke in the United States.

Findings: We analyzed data from the 2009 and 2010 Nationwide Emergency Department Sample, the largest ED data system sponsored by the Agency for Healthcare Research and Quality. ED visits for heat stroke were identified by screening the recorded diagnoses using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) code 992.0. Annual incidence rates of ED visits for heat stroke were computed according to demographic characteristics and geographic regions. In 2009 and 2010, there were an estimated 8,251 ED visits for heat stroke in the United States, yielding an annual incidence rate of 1.34 visits per 100,000 population (95% Confidence Interval [CI] = 1.23-1.45). Significantly higher incidence rates were found in males (1.99 per 100,000; 95% CI = 1.82-2.16), adults aged ≥ 80 years (4.45 per 100,000; 95% CI = 3.73-5.18), and residents living in the southern region (1.61 per 100,000; 95% CI = 1.43-1.79). The majority (63.1%) of ED visits for heat stroke occurred during the summer months of June, July and August. Over one-half (54.6%) of the ED visits for heat stroke required hospitalization and 3.5% of the patients died in the ED or hospital.

Conclusions: Heat stroke results in approximately 4,100 ED visits each year in the United States, with the majority occurring in the summer months and requiring admission to the hospital. Men, the elderly, and people living in the south region are at heightened risk.

Keywords: Emergency medical service; Epidemiology; Global warming; Heat stroke; Public health



Incident Stroke and Mortality Associated with New-onset Atrial Fibrillation in Patients Hospitalized with Severe Sepsis

Allan J. Walkey, MD, MSc,

The Pulmonary Center, Division of Pulmonary and Critical Care Medicine, Department of Medicine, Boston University School of Medicine, Boston, MA, USA.

Renda Soylemez Wiener, MD, MPH,

The Pulmonary Center, Division of Pulmonary and Critical Care Medicine, Boston University School of Medicine, Boston, MA USA. Center for Health Quality, Outcomes, & Economic Research, Edith Nourse Rogers Memorial VA Hospital, Bedford, MA, USA. The Dartmouth Institute for Health Policy and Clinical Practice, Dartmouth Medical School, Hanover, NH, USA

Joanna M. Ghobrial, MD,

Department of Medicine, Division of Cardiology, University of Washington School of Medicine, Seattle, WA USA.

Lesley H. Curtis, PhD, and

Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC, USA.
Department of Medicine, Duke University School of Medicine, Durham, NC, USA.

Emelia J. Benjamin, MD, ScM

National Heart Lung and Blood Institute's and Boston University's Framingham Heart Study, Framingham, MA, USA. Cardiology and Preventive Medicine, Whitaker Cardiovascular Institute, Department of Medicine, Boston University School of Medicine, Boston, MA, USA. Epidemiology Department, Boston University School of Public Health, Boston, MA, USA.

Abstract

Context—New-onset fibrillation (AF) has been reported in 6–20% of patients with severe sepsis. Whereas chronic AF is a known risk factor for stroke and death, the clinical significance of new-onset AF in the setting of severe sepsis is uncertain.

Objective—To determine the in-hospital stroke and in-hospital mortality risks associated with new-onset AF in patients with severe sepsis.

Design—Retrospective population-based cohort of California State Inpatient Database administrative claims data from 1/1/2007 through 12/31/2007.

Setting—Non-Federal acute care hospitals.



ICD-9-CM Validation


Validation of the severe sepsis *ICD-9-CM* 995.92 at our institution (see eSupplement) demonstrated moderate sensitivity [52%, (95% Confidence Interval (CI), 39–65%)] and high specificity [98%, (95% CI, 92–100%)], similar to validation findings in other hospitals.²³ Previous studies have demonstrated 95% sensitivity and 99% specificity for AF *ICD-9-CM* 427.3× claims.¹³ We validated present on admission modifiers (see eSupplement) for severe sepsis and AF claims; agreement between severe sepsis present on admission status and blinded chart review was 91% (kappa 0.77) and an agreement between AF present on admission status and blinded chart review was 90% (kappa 0.74), similar to previous findings.²⁴ Prior validation of ischemic stroke *ICD-9-CM* codes has shown variable accuracy,^{14,25–27} however, a strategy using *ICD-9-CM* 433, 434, 436 at any diagnostic position previously demonstrated 86% sensitivity and 95% specificity (kappa=0.82).¹⁴ *ICD-9-CM* 434.11 used to indicate embolic stroke has been demonstrated to be accurate in 73% of patients; patients coded with 434.11 who did not have a clear embolic stroke on chart review were characterized as having either ischemic stroke of athero-thrombotic or uncertain etiology.²⁶



Original Investigation

Assessment of the CHA₂DS₂-VASC Score in Predicting Ischemic Stroke, Thromboembolism, and Death in Patients With Heart Failure With and Without Atrial Fibrillation

Line Melgaard, MSc; Anders Gorst-Rasmussen, MSc, PhD; Deidre A. Lane, PhD;
Lars Hvilsted Rasmussen, MD, PhD; Torben Bjerregaard Larsen, MD, PhD; Gregory Y. H. Lip, MD

 Supplemental content at
jama.com

IMPORTANCE The CHA₂DS₂-VASC score (congestive heart failure, hypertension, age ≥ 75 years [doubled], diabetes, stroke/transient ischemic attack/thromboembolism [doubled], vascular disease [prior myocardial infarction, peripheral artery disease, or aortic plaque], age 65-75 years, sex category [female]) is used clinically for stroke risk stratification in atrial fibrillation (AF). Its usefulness in a population of patients with heart failure (HF) is unclear.

OBJECTIVE To investigate whether CHA₂DS₂-VASC predicts ischemic stroke, thromboembolism, and death in a cohort of patients with HF with and without AF.

DESIGN, SETTING, AND POPULATION Nationwide prospective cohort study using Danish registries, including 42 987 patients (21.9% with concomitant AF) not receiving anticoagulation who were diagnosed as having incident HF during 2000-2012. End of follow-up was December 31, 2012.

EXPOSURES Levels of the CHA₂DS₂-VASC score (based on 10 possible points, with higher scores indicating higher risk), stratified by concomitant AF at baseline. Analyses took into account the competing risk of death.

MAIN OUTCOMES AND MEASURES Ischemic stroke, thromboembolism, and death within 1 year after HF diagnosis.

RESULTS In patients without AF, the risks of ischemic stroke, thromboembolism, and death were 3.1% (n = 977), 9.9% (n = 3187), and 21.8% (n = 6956), respectively; risks were greater with increasing CHA₂DS₂-VASC scores as follows, for scores of 1 through 6, respectively: (1) ischemic stroke with concomitant AF: 4.5%, 3.7%, 3.2%, 4.3%, 5.6%, and 8.4%; without concomitant AF: 1.5%, 1.5%, 2.0%, 3.0%, 3.7%, and 7% and (2) all-cause death with concomitant AF: 19.8%, 19.5%, 26.1%, 35.1%, 37.7%, and 45.5%; without concomitant AF: 7.6%, 8.3%, 17.8%, 25.6%, 27.9%, and 35.0%. At high CHA₂DS₂-VASC scores (≥ 4), the absolute risk of thromboembolism was high regardless of presence of AF (for a score of 4, 9.7% vs 8.2% for patients without and with concomitant AF, respectively; overall $P < .001$ for interaction). C statistics and negative predictive values indicate that the CHA₂DS₂-VASC score performed modestly in this HF population with and without AF (for ischemic stroke, 1-year C statistics, 0.67 [95% CI, 0.65-0.68] and 0.64 [95% CI, 0.61-0.67], respectively; 1-year negative predictive values, 92% [95% CI, 91%-93%] and 91% [95% CI, 88%-95%], respectively).

Author Affiliations: Aalborg
Thrombosis Research Unit,
Department of Clinical Medicine,
Faculty of Health, Aalborg University,
Aalborg, Denmark (Melgaard,
Gorst-Rasmussen, Rasmussen)



Methods

Registry Data Sources

We used 3 nationwide registries in this study: (1) the Danish National Patient Register,¹⁴ which has registered all hospital admissions along with diagnoses since 1977 and has coded all diagnoses according to the *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10)* since 1994; (2) the Danish National Prescription Registry,¹⁵ which contains data on all prescriptions dispensed from Danish pharmacies since 1994, coded according to the Anatomical Therapeutic Chemical (ATC) Classification System; and (3) the Danish Civil Registration System, which holds information on date of birth, migration, vital status, date

of death, and sex of all persons living in Denmark.¹⁶ Data were linked via a unique personal identification number used in all Danish national registries. All 3 registries were used up to December 31, 2012 (end of follow-up). These registries have previously been well validated,^{14,15,17} and the diagnoses of HF, AF, and ischemic stroke have been found to be valid.¹⁷⁻¹⁹

No ethical approval is required for anonymous register studies in Denmark. The study was approved by the Danish Data Protection Agency.

Study Population

The study population was identified as patients aged 50 years or older discharged with a primary diagnosis of incident HF (*ICD-10* codes I50, I42.0, I11.0, I13.0, and I13.2) in the period January 1, 2000, to December 31, 2012. Patients with AF were identified by a hospital diagnosis of AF or atrial flutter (*ICD-10* code I48) between 1994 and baseline. We excluded patients treated with a vitamin K antagonist (ATC codes B01AA03 and B01AA04) within 6 months prior to the HF diagnosis. Moreover, patients with a diagnosis of cancer (*ICD-10* codes C00-C97) within 5 years before HF diagnosis or with a prior diagnosis of chronic obstructive pulmonary disease (COPD [*ICD-10* code J44]) were excluded.

Comorbidities at baseline were identified using the Danish National Patient Register and the Danish National Prescription Registry. Ascertainment of baseline medication status was based on medication purchase in a 45-day window before or after the date of HF diagnosis. *ICD-10* codes and ATC codes were used to define comorbidities and medical therapies (eTable 1 in the Supplement).



Surgical outcomes

- **American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP)**

From the patient's medical chart, not insurance claims: In a study comparing ACS NSQIP data to administrative and claims data collected by the University Health System Consortium (UHC) program,² ACS NSQIP identified 61 percent more complications than UHC, including 97 percent more surgical site infections.

Risk-adjusted: ACS NSQIP lets you compare apples to apples. Your data is risk-adjusted, based on models in use for more than 20 years. Caring for a chronically ill 75-year-old is very different from treating a healthy 21-year-old, and quality measures should take these differences into account.

Case-mix-adjusted: ACS NSQIP allows a hospital that takes on more complex surgical cases to meaningfully calibrate its results against one that performs more straightforward procedures. ACS NSQIP accounts for the complexity of operations performed, allowing for more accurate national benchmarking.

Based on 30-day patient outcomes: Studies show half or more of all complications occur after the patient leaves the hospital, often leading to costly readmissions. ACS NSQIP tracks patients for 30 days after their operation, providing a more complete picture of their care. either.



Original Investigation

Underlying Reasons Associated With Hospital Readmission Following Surgery in the United States

Ryan P. Merkow, MD, MS; Mila H. Ju, MD, MS; Jeanette W. Chung, PhD; Bruce L. Hall, MD, PhD, MBA; Mark E. Cohen, PhD; Mark V. Williams, MD; Thomas C. Tsai, MD, MPH; Clifford Y. Ko, MD, MS, MSHS; Karl Y. Bilimoria, MD, MS

IMPORTANCE Financial penalties for readmission have been expanded beyond medical conditions to include surgical procedures. Hospitals are working to reduce readmissions; however, little is known about the reasons for surgical readmission.

OBJECTIVE To characterize the reasons, timing, and factors associated with unplanned postoperative readmissions.

DESIGN, SETTING, AND PARTICIPANTS Patients undergoing surgery at one of 346 continuously enrolled US hospitals participating in the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) between January 1, 2012, and December 31, 2012, had clinically abstracted information examined. Readmission rates and reasons (ascertained by clinical data abstractors at each hospital) were assessed for all surgical procedures and for 6 representative operations: bariatric procedures, colectomy or proctectomy, hysterectomy, total hip or knee arthroplasty, ventral hernia repair, and lower extremity vascular bypass.

MAIN OUTCOMES AND MEASURES Unplanned 30-day readmission and reason for readmission.

← Editorial page 467

+ JAMA Report Video and Author Video Interview at jama.com

+ Supplemental content at jama.com

+ CME Quiz at jamanetworkcme.com and CME Questions page 518

Cancer Epidemiology

- **Surveillance, Epidemiology, and End Results Program (SEER)**

- **SEER-Medicare linked database (Medicare beneficiaries with cancer)**

- **National Cancer Database (NCDB)**

The nationally recognized National Cancer Database (NCDB)—jointly sponsored by the American College of Surgeons and the American Cancer Society—is a clinical oncology database sourced from hospital registry data that are collected in more than 1,500 Commission on Cancer (CoC)-accredited facilities. NCDB data are used to analyze and track patients with malignant neoplastic diseases, their treatments, and outcomes. Data represent more than 70 percent of newly diagnosed cancer cases nationwide and more than 34 million historical records.

Time to Surgery and Breast Cancer Survival in the United States

Richard J. Bleicher, MD; Karen Ruth, MS; Elin R. Sigurdson, MD, PhD; J. Robert Beck, MD; Eric Ross, PhD; Yu-Ning Wong, MD, MPH; Sameer A. Patel, MD; Marcia Boraas, MD; Eric I. Chang, MD; Neal S. Topham, MD; Brian L. Egleston, PhD

IMPORTANCE Time to surgery (TTS) is of concern to patients and clinicians, but controversy surrounds its effect on breast cancer survival. There remains little national data evaluating the association.

OBJECTIVE To investigate the relationship between the time from diagnosis to breast cancer surgery and survival, using separate analyses of 2 of the largest cancer databases in the United States.

DESIGN, SETTING, AND PARTICIPANTS Two independent population-based studies were conducted of prospectively collected national data from the Surveillance, Epidemiology, and End Results (SEER)-Medicare-linked database and the National Cancer Database (NCDB). The SEER-Medicare cohort included Medicare patients older than 65 years, and the NCDB cohort included patients cared for at Commission on Cancer-accredited facilities throughout the United States. Each analysis assessed overall survival as a function of time between diagnosis and surgery by evaluating 5 intervals (≤ 30 , 31-60, 61-90, 91-120, and 121-180 days) and disease-specific survival at 60-day intervals. All patients were diagnosed with noninflammatory, nonmetastatic, invasive breast cancer and underwent surgery as initial treatment.

MAIN OUTCOMES AND MEASURES Overall and disease-specific survival as a function of time between diagnosis and surgery, after adjusting for patient, demographic, and tumor-related factors.

← Editorial page 302

+ Author Audio Interview at jamaoncology.com

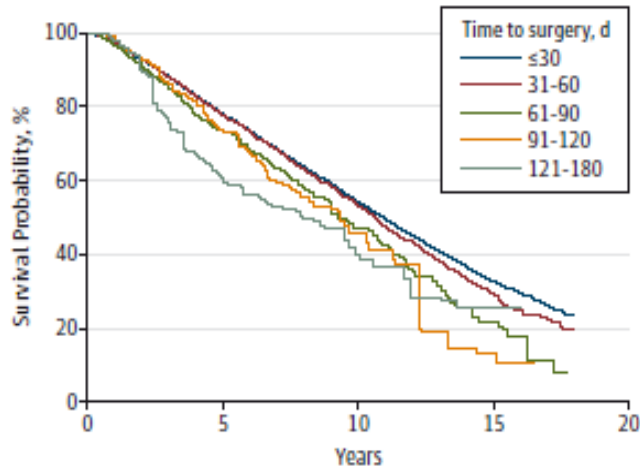
← Related article page 322

+ Supplemental content at jamaoncology.com



Figure 1. Adjusted Overall Survival

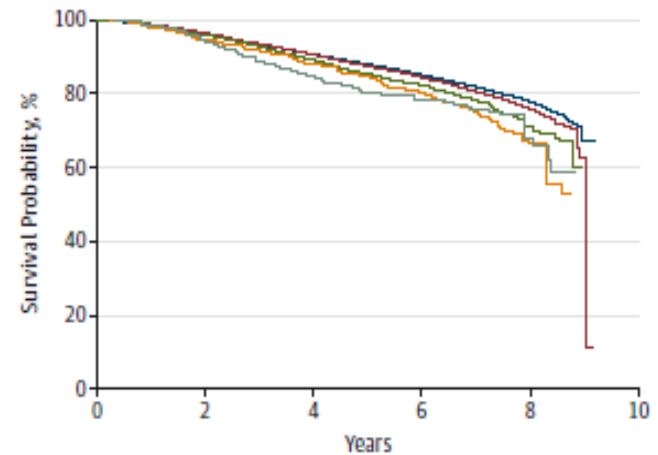
A SEER-Medicare



No. at risk

≤30 d	73491	38075	10870	2386
31-60 d	17345	6370	1132	212
61-90 d	2586	760	110	12
91-120 d	686	235	24	4
121-180 d	436	121	16	3

B NCDB



No. at risk

≤30 d	80505	73422	66532	43354	5811
31-60 d	28832	26272	23643	14721	1783
61-90 d	4697	4163	3667	2170	247
91-120 d	1152	991	854	497	40
121-180 d	604	513	413	239	27

Adjusted overall survival for Surveillance, Epidemiology, and End Results (SEER)-Medicare Database patients (A) and National Cancer Database (NCDB) patients (B) for preoperative delay intervals of ≤30, 31-60, 61-90, 91-120, and 121-180 days. The hazard ratio for each increasing delay in SEER-Medicare interval was 1.09 (95% CI, 1.06-1.13; $P < .001$). The hazard ratio for each increasing delay interval in NCDB was 1.10 (95% CI, 1.07-1.13; $P < .001$).

Social Epidemiology

- **National Health and Nutrition Examination Survey (NHANES)**
<https://www.cdc.gov/nchs/nhanes/index.htm>
 - a complex, stratified, multistage probability sampling design
 - nationally representative data on dietary intake, health conditions, and objectively measured body weight/height
- **Behavioral Risk Factor Surveillance System (BRFSS)**
 - Random-digit telephone survey conducted by state health departments on independent probability samples of state residents aged 18 years or more.
 - It is the world's largest ongoing telephone health system survey, containing data from more than 350,000 adults annually.
- <https://www.cdc.gov/brfss/index.html>
- **National Youth Risk Behavior Survey (YRBS)**
- **NYC Community Health Survey (NYC CHS)**
<https://www1.nyc.gov/site/doh/data/data-sets/community-health-survey-public-use-data.page>



Original Investigation

Prevalence of and Trends in Diabetes Among Adults in the United States, 1988-2012

Andy Menke, PhD; Sarah Casagrande, PhD; Linda Geiss, MA; Catherine C. Cowle, PhD

IMPORTANCE Previous studies have shown increasing prevalence of diabetes in the United States. New US data are available to estimate prevalence of and trends in diabetes.

OBJECTIVE To estimate the recent prevalence and update US trends in total diabetes, diagnosed diabetes, and undiagnosed diabetes using **National Health and Nutrition Examination Survey (NHANES) data**.

DESIGN, SETTING, AND PARTICIPANTS **Cross-sectional surveys** conducted between 1988-1994 and 1999-2012 of nationally representative samples of the civilian, noninstitutionalized US population; 2781 adults from 2011-2012 were used to estimate recent prevalence and an additional 23 634 adults from 1988-2010 were used to estimate trends.

MAIN OUTCOMES AND MEASURES The prevalence of diabetes was defined using a previous diagnosis of diabetes or, if diabetes was not previously diagnosed, by (1) a hemoglobin A_{1c} level of 6.5% or greater or a fasting plasma glucose (FPG) level of 126 mg/dL or greater (hemoglobin A_{1c} or FPG definition) or (2) additionally including 2-hour plasma glucose (2-hour PG) level of 200 mg/dL or greater (hemoglobin A_{1c}, FPG, or 2-hour PG definition). Prediabetes was defined as a hemoglobin A_{1c} level of 5.7% to 6.4%, an FPG level of 100 mg/dL to 125 mg/dL, or a 2-hour PG level of 140 mg/dL to 199 mg/dL.

- ← Editorial page 1005
- + Author Video Interview and JAMA Report Video at jama.com
- ← Related article page 1052
- + Supplemental content at jama.com

Other common data sources

- **US Census data**

- American Community Survey (ACS)

- <https://www.census.gov/programs-surveys/acs>

- https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=PEP_2016_PEPANNRES&src=pt

- **National Center for Health Statistics**

- <https://www.cdc.gov/nchs/index.htm>

- <https://wonder.cdc.gov/>



Advantages of secondary data sources

- Study can be executed rapidly and inexpensively
- Can be used for studies with large sample size or long follow-up requirements
- Operational issues significantly reduced
 - eg, subject informed consent and site management not needed (generally)
- Pharmacy information (dispensings) may more accurate than self-report and medical record, especially for those who are too ill or who have died
- Data linkage with other databases to obtain additional information (ie, death, cancer, etc.)

Disadvantages of secondary data sources

- Diagnoses may not be valid, particularly when data have been generated for reimbursement purposes
 - eg, recording of rule-out diagnoses
- Data on important confounders, such as disease severity, behavior data, etc., and lab results generally unavailable
- Data on over-the-counter and inpatient drug use generally lacking
- In databases with high patient turnover, information will be significantly truncated
 - population-based databases (vs. insurance claims) tend to have more stable population

Choosing between primary and secondary data collection

- Need to rank data sources for capturing required data elements, eg:
 - Sufficient number of patients who meet key inclusion and exclusion criteria
 - Recording of lab data required for valid measurement of outcome
 - Routine conduct of clinical assessments required for valid measurement of confounding diagnoses
- May consider hybrid approach
 - eg, supplementing aggregated secondary data sources with medical record review

Data source considerations

- Choice guided by
 - Research question
 - Validity of measurement of required data elements
 - Capability of addressing sources of bias
 - Sample size requirements

Thank you



**Weill
Cornell
Medicine**