



Creating preliminary data: Large dataset research

EAST Research Short Course
Wednesday, January 15th, 2020
Orlando, FL

Heena P Santry, MD MS FACS
Associate Professor of Surgery
Director, Center for Surgical Health Assessment, Research & Policy
Ohio State Wexner Medical Center

SHARP



Nothing to disclose



THE OHIO STATE UNIVERSITY

WEXNER MEDICAL CENTER

SHARP



Myths about large dataset research

- It's quick
- It's easy
- I have a med student who knows stats



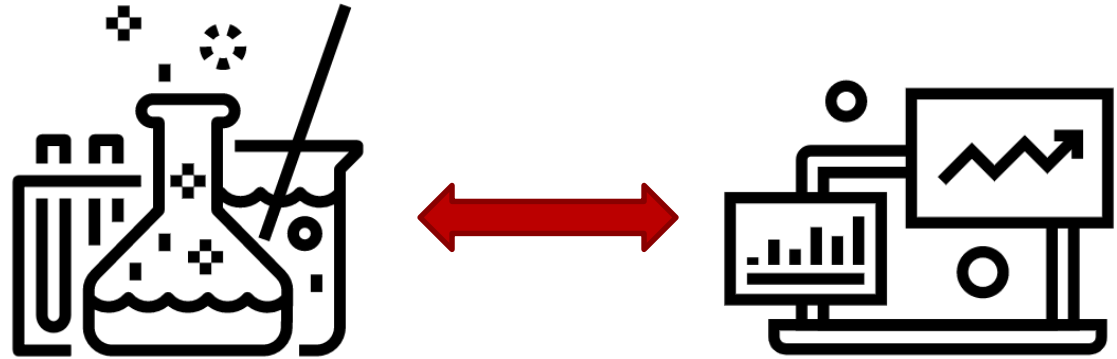
Realities of dataset research

- Critical thinking
- Time
- Statistical expertise



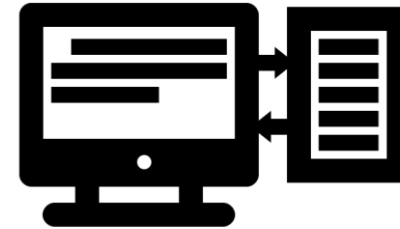
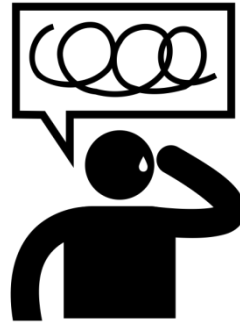
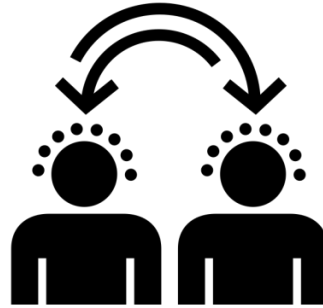
Large dataset research is SCIENCE

- Robust research design
- Hypothesis driven
- Novel



Education

- MSHS
- MPH
- Certificate
- Epidemiology
- Biostatistics



Resources - Reading

SGIM
Society of General Internal Medicine
Creating Value for Patients

SGIM About SGIM Communities Meetings Publications Resource Library Career Center Membership

SGIM Communities Research Dataset Compendium

COMMUNITIES

ACLIGIM

Advocacy

Clinical Practice

Education

Research

Dataset Compendium

Public Datasets: Description

Public Datasets: Topic Grid

Proprietary Datasets

Funding Corner

Research eConsults

Other SGIM Committees

https://www.sgim.org/jms

Dataset Compendium Overview

The SGIM Research Dataset Compendium is designed to assist investigators conducting research on existing datasets, with a particular emphasis on health services research, clinical epidemiology, and research on medical education. The detailed information provided by the SGIM compendium distinguishes it from other web-based compendia, which typically provide lists of datasets but give little information about their strengths and weaknesses and the insights of experienced users about making best use of the data.

About

This site is a project of the SGIM Research Committee since 2008. Information in this site was initially compiled under the direction of Michael Steinman (chief developer), with assistance from John Ayanian, Ken Covinsky, Christina Wee, Stacey Jackson, Bruce Landon, Mitchell Wong, Amy Woodward, and others. Alex Smith took over leadership of the compendium in 2010 and expanded the resource in 2012 to include proprietary datasets.

Special thanks to Sneha Patel, Julie Machulsky, and Francine Jetton for background research, website development, technical assistance, and support, and to John Ayanian and Ellen McCarthy for resource development. In addition, we thank the many people who volunteered their service as dataset experts.

How to use this site:

This site is divided into four main sections. Users are encouraged to browse the different sections of the site rather than focus only on one area.

- [User's guide to working with secondary data:](#) Tips for working with secondary data.
- [Public datasets:](#) Descriptions and expert evaluations of high-value datasets
- [Proprietary datasets:](#) Collaborate with senior SGIM researchers using their proprietary datasets
- [Other dataset compendia, repositories, and resources:](#) Brief descriptions and links to other dataset compendia, data

This Issue Views **3,276** | Citations **8** | Altmetric **63** | Comments **1**

Editorial

June 2018

Tips for Analyzing Large Data Sets From the *JAMA Surgery* Statistical Editors

Amy H. Kaji, MD, PhD^{1,4}; Alfred W. Rademaker, PhD^{2,4}; Terry Hyslop, PhD^{3,4}

» [Author Affiliations](#)

JAMA Surg. 2018;153(6):508-509. doi:10.1001/jamasurg.2018.0647

Download PDF

Full Text

Cite This

Permissions

Comment

Browse and subscribe to work podcasts!

PDF Help

SHARP

Resources - Training



Your source for CMS data support

FIND CMS DATA FILES REQUEST CMS DATA FILES SEARCH DATA VARIABLES LEARN ABOUT CMS DATA

Introduction to the Use of Medicare Data for Research

Medicare data are useful for understanding the cost and use of healthcare, but they are also very complex. Understanding the origin, structure and contents of the data is essential to designing a successful research project. This workshop delivers an overview of Medicare benefit and payment design, describes available research files, and provides insight into how these features affect inference. Course faculty share what they have learned from a combined 40+ years of using Medicare data for research.

REGISTER NOW

Workshop Date: August 6, 2019 to August 7, 2019

ResDAC Faculty:

- Beth Virginia, PhD, MPH
- Helen Parsons, PhD, MPH
- Nathan Shapiro, PhD, MS
- Stephanie Jarossek, PhD, BSN

Agenda: [CMS 101 Agenda - Samele.pdf](#)

Location: Carlson School of Management, 3M Auditorium
(Room 1.150, Minneapolis, MN)



U.S. Department of Health & Human Services

About Us Careers Contact Us Español FAQ E-mail

AHRQ's 9-24-2019 HCUP Data Users Workshop – Registration Now Open

Agency for Healthcare Research and Quality (AHRQ) open this Bulletin at 08/05/2019 03:31 PM EDT

REGISTRATION OPEN!

AHRQ ONE-DAY HANDS-ON HCUP DATA USERS' WORKSHOP

TUESDAY, SEPTEMBER 24, 2019, 9:00 A.M. – 4:00 P.M. ET

Workshop Topic: An In-Depth Exploration of HCUP Resources to Study Hospital Utilization for Opioid, Alcohol, and Other Substances

Location: Rockville, MD

Summary: Register today for AHRQ's full-day hands-on Workshop on Tuesday, September 24, 2019 from 9:00 A.M. – 4:00 P.M. ET! Registration information is also available on the [HCUP Workshops & Webinars page of the HCUP User Support \(HCUP-US\) website](#) at: www.hcup-us.ahrq.gov/hcup_web_workshop.jsp.

This intermediate-level data users' workshop is designed for health services researchers and analysts who want to learn more about using HCUP databases and products. Examples of HCUP resources will be demonstrated through data analyses on opioid, alcohol, and substance use topics. Instructors will show how to use HCUP Fast Stats and HCUPPhet to access substance use statistics as well as how to evaluate those conditions using the HCUP nationwide databases – the National Inpatient Sample (NIS), Nationwide Readmissions Database (NRD), Kids' Inpatient Database (KID), and Nationwide Emergency Department Sample (NEDS). Comparisons will be made to demonstrate the information that can be obtained from each database and help users better understand which databases are best suited to help answer their research questions. Attendees will learn how to utilize HCUP User Support (HCUP-US) documentation to their advantage when developing analyses. Instructional and reference materials will be distributed and discussed. Participants should bring their own computers to follow the examples. Given the content and pace of this course, participants would benefit from having prior experience with HCUP or with large administrative databases. If not familiar with HCUP databases, please attend the HCUP Overview

CDC Centers for Disease Control and Prevention
CDC 247: Saving Lives, Protecting People™

Search

Genomics & Precision Health

Genomics & Precision Health



Genomics & Precision Health

About Us

Hot Topics of the Day

Weekly Update

PHGKB

Reports and Publications

Genomics and Precision Health Blog

Events and Multimedia



Genomics, Big Data and Data Science in Public Health

August 9, 2019, 9:30 am – 12:00 pm EDT
CDC Chamblee Campus, Building 107, Room 1A

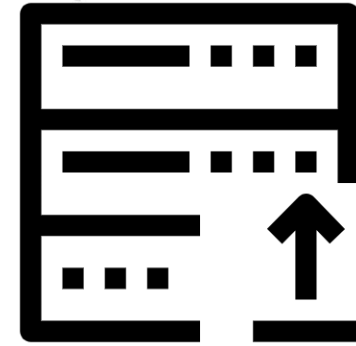
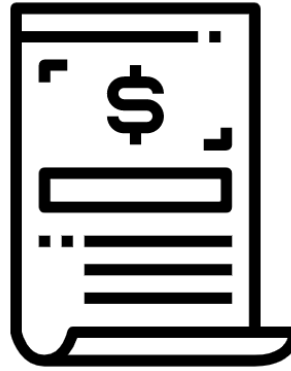
Big Data encompasses the ever increasing amounts of health-related information from disparate sources that can provide more precision by place, time, and persons than previously available. Although genomic and other molecular technologies helped launch Big Data, the field now offers emerging

SHARP

THE OHIO STATE UNIVERSITY
WEXNER MEDICAL CENTER

Types of data

- Administrative
- Registries
- Quality programs
- Survey



Administrative data

Nationwide HCUP Databases

HCUP's Nationwide databases can be used to identify, track, and analyze national trends in healthcare utilization, access, charges, quality, and outcomes.

National (Nationwide) Inpatient Sample (NIS)

- [NIS Database Documentation](#)

Kids' Inpatient Database (KID)

- [KID Database Documentation](#)

Nationwide Ambulatory Surgery Sample (NASS)

- [NASS Database Documentation](#)

Nationwide Emergency Department Sample (NEDS)

- [NEDS Database Documentation](#)

Nationwide Readmissions Database (NRD)

- [NRD Database Documentation](#)

State-Specific HCUP Databases

HCUP's State-specific databases can be used to investigate State-specific and multi-State trends in healthcare utilization, access, charges, quality, and outcomes.

State Inpatient Databases (SID)

- [SID Database Documentation](#)

State Ambulatory Surgery and Services Databases (SASD)

- [SASD Database Documentation](#)

State Emergency Department Databases (SEDD)

- [SEDD Database Documentation](#)

Find, Request and Use CMS Data

GETTING STARTED

New to CMS data

How to begin

- Who is in the data? →
- What is in the data? →
- What type of data is right for me? →

SUBMITTING A REQUEST

Find the documents you need & submit a request

How to request identifiable data

- Timeline and process →
- CMS data fee information →
- Get the documents you need →

LEARN ABOUT CMS DATA


Get answers about CMS data

How to understand & use the data

- CMS data training →
- Articles about the data →
- Medicaid data quality resources →

Registry data

[American College of Surgeons](#) > [Quality Programs](#) > [Annual Call for Data: National Trauma Data Bank \(NTDB\)](#) > About NTDB



About NTDB

The National Trauma Data Bank® (NTDB®) is the largest aggregation of U.S. trauma registry data ever assembled.

Annual Call for Data (NTDB)

- About NTDB
- Annual Call for Data Instructions
- National Trauma Data Standard (NTDS)
- TQP Participant Hub
- Past NTDB Data Points
- TQP Participant Use File
- NTDB Reports and Publications
- Contact NTDB

NTDB Participation

NTDB participants not only contribute to the growing knowledge base for trauma research, but they also gain access to a variety of reports that can be used to benefit their hospital.

Research Data Set

The NTDB creates and distributes [research data sets](#) that can be used by researchers. To gain access to NTDB data, researchers must submit requests through our online application process.

NTDB Annual Reports

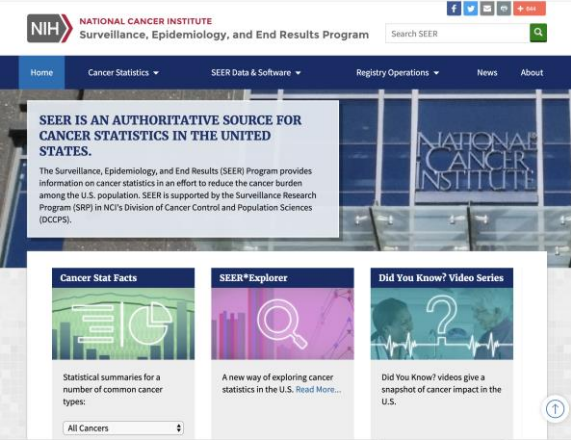
NTDB Annual Adult and Pediatric Reports contain descriptive information about trauma patients, including demographics, injury information, and outcomes.

NTDS Data Dictionary Download

The National Trauma Data Standard (NTDS) Data Dictionary is now available for download. Download the latest version.

NTDS Data Dictionary Revision Site

The NTDS Data Dictionary is a living, changing document. In order to continue improving the quality of the data,



SHARP

Quality program data



ACS NSQIP Participant Use Data File

The Participant Use Data File (PUF) is a Health Insurance Portability and Accountability Act (HIPAA)-compliant data file containing cases submitted to the American College of Surgeons National Surgical Quality Improvement Program® (ACS NSQIP®). The PUF contains patient-level, aggregate data and does not identify hospitals, health care providers, or patients. The intended purpose of this file is to provide researchers at participating sites with a data resource they can use to investigate and advance the quality of care delivered to the surgical patient through the analysis of cases captured by ACS NSQIP. The PUF is provided at no additional cost to employees (surgeons, Surgical Clinical Reviewers, researchers, etc.) of ACS NSQIP-participating hospitals.

The 2018 PUF contains 1,020,511 cases submitted from 722 NSQIP-participating sites. Twelve other separate NSQIP PUFs, containing a rich database of more than 6.6 million cases, are also available. Only cases included in corresponding Semiannual Report (SAR) risk-adjustment calculations are in the PUF datasets. For background information on the case inclusion/exclusion criteria and the data collection and submission processes, please visit the ACS NSQIP Program Specific page. Variable formats, variable definitions (which can change from year to year), and other supporting information are available in the PUF User Guides. The data files are made available in a delimited text, SAS, and SPSS file type.

The 2018 PUF User Guide is available.

Previously Available PUFs

- 2017—1,028,713 cases submitted by 708 hospitals (2017 PUF User Guide)
- 2016—1,000,393 cases submitted by 680 hospitals (2016 PUF User Guide)
- 2015—885,502 cases submitted by 603 hospitals (2015 PUF User Guide)
- 2014—750,387 cases submitted by 517 hospitals (2014 PUF User Guide)

Participant Use Data File

Participant Use Request Form

SHARP

Survey data

CDC Centers for Disease Control and Prevention
 CDC 24/7: Saving Lives, Protecting People™

National Center for Health Statistics

CDC > NCHS > National Health and Nutrition Examination Survey

National Health and Nutrition Examination Survey

NHANES Questionnaires, Datasets, and Related Documentation

Survey Methods
Plan & Operations, Sample Design, Estimation & Weighting Procedures, Analytic Guidelines, etc.

Search Variables
Simple keyword search for Continuous NHANES (1999 and on) variables

Continuous NHANES

NHANES 2019-2020	NHANES 2017-2018	NHANES 2015-2016	NHANES 2013-2014
------------------	------------------	------------------	------------------

HRS Data Products

Listings of available HRS data products, with access instructions and policies.

Public Data

Public Survey Data
A listing of publicly available biennial, off-year, and cross-year data products.

RAND HRS Data
HRS data products produced by the RAND Center for the Study of Aging.

Gateway Harmonized Data
HRS data products produced by the USC Program on Global Aging, Health, and Policy.

Contributed and Replication Data
Data products (unsupported by the HRS) provided by researchers sharing their work.

Register and Access Public Data
Log in to download public data products.

Restricted/Sensitive Data

Cognition Data
A summary of HRS cognition data, including the new Harmonized Cognition Assessment Protocol (HCAP.)

Biomarker and Health Data
Sensitive health data files available are from the public data portal after a supplemental agreement is signed.

Restricted Data
HRS restricted data files require a detailed application process, and are available only through remote-virtual desktop or encrypted physical media.

Administrative Linkages
Links HRS data with Medicare and Social Security.

Genetic Data
Genetic data products derived from 20,000 genotyped HRS respondents.

More Info

Data Collection Path Diagram
A table of HRS data products arranged by data collection year.

Data Alerts
Notices of errors, corrections, or problems in HRS early and final public data releases and associated documentation.

Distribution and Replication Policy
Information for organizations interested in redistributing or archiving HRS data products.

File Merge Reference
Information on limitations when merging the various types of HRS data products.

CDC Centers for Disease Control and Prevention
 CDC 24/7: Saving Lives, Protecting People™

Search

Behavioral Risk Factor Surveillance System

2018 BRFSS Data Now Available
View the latest 2018 BRFSS Annual Data

BRFSS™

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world. [See More.](#)

About BRFSS

BRFSS Questionnaires

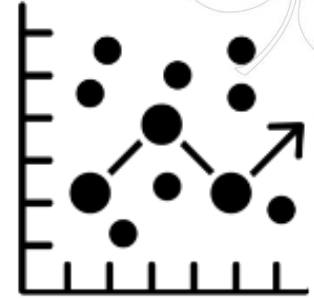
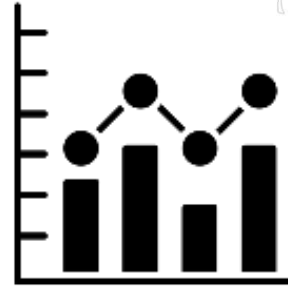
Publications & Resources

SHARP

THE OHIO STATE UNIVERSITY
 WEXNER MEDICAL CENTER

Analytic potential

- Epidemiology
- Outcomes research
- Social determinants of health
- Social network analysis
- Health behaviors
- Cost effectiveness





Step 1.

What is your research question?

Why do you need a large dataset to ask it?

What will you do with the findings?

Step 2.

Write your introduction – End with the why/so what

Template your tables

Step 3.

Understand the data – Review the data dictionary

What variables are collected?

How are they made available?

How can you identify your population of interest?

Step 4.

Which dataset can answer your research question?

What are the limitations of this choice dataset?

How can you acquire the data?

Does someone else on campus already own it?

Step 5.

Acquire the data

DUAs

IRBs

Data Security



Step 6.

Design your experiment

Step 7.

Conduct analyses

Fill your tables

Step 8.

Ask yourself did I find anything novel?

Did my findings support or refute the hypothesis?

Do not massage the data

Step 9.

Create compelling visuals

Write the results

Interpret the results

Step 10.

Craft conclusions, limitations, implications

So what?

THANK YOU!

heena.santry@osumc.edu

Questions?

SHARP