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
Nothing to disclose
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
Editorial

June 2018

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

Amy H. Kaji, MD, PhD1,2; Alfred W. Rademaker, PhD2,4; Terry Hyslop, PhD3,4

Author Affiliations

Resources - Training

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 content of the data is essential to designing a successful research project. This workshop will provide an overview of Medicare benefits and payment design, describe available research files, and provide 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.

Workshop Dates: August 6 to August 7, 2019
ResDAC Faculty:
- Sarah Zavala, PhD, MHS
- Nicola Farnan, PhD, MHS
- Julie Moro, PhD, MHS
- Stephanie Antrim, PhD, EDD

Agenda:
- CMS 101: An Overview of Medicare

Location: Carter School of Management, 3W Auditorium

Genomics & Precision Health

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 genomics and other molecular technologies helped launch Big Data, the field now offers emerging...
Types of data

- Administrative
- Registries
- Quality programs
- Survey
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 and participating sites with a data resource they can use to investigate and achieve 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 1,025,511 cases submitted from 702 NSQIP-participating sites. Twelve other separate NSQIP PUFs, containing a rich database of more than 8 million cases, are also available. Only cases included in corresponding Semiannual Report (SAR) risk-adjustment calculations are in the PUF database. 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 Guide. The data files are made available in a defined text, SAS, and SPSS file type.

Previously Available PUFs

2017—1,028,713 cases submitted by 708 hospitals (2017 PUF User Guide)
2016—1,020,062 cases submitted by 678 hospitals (2016 PUF User Guide)
Survey data

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 U.S. Program on Global Aging, Health, and Policy.
- Contributed and Replication Data
  Data products (unsupported by the HRS) provided by researchers sharing their contributions.
- 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 access or encrypted physical media.

Administrative Links
Links to 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 Merga Reference
Information on limitations when merging the various types of HRS data products.
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
Craft conclusions, limitations, implications

So what?
THANK YOU!

heena.santry@osumc.edu

Questions?