Research Using Large Healthcare Datasets

Lecture 2

John J. Chen, Ph.D.
Professor
Office of Biostatistics & Quantitative Health Sciences
UH JABSOM

Surgery Resident Conference Session
December 19, 2018

Outline

Lecture 1 (10/17/2018)

- The goal of statistics
- Introduction to descriptive biostatistics
- Some research design and data presentation issues

Lecture 2 (12/19/2018)

- Large databases

Lecture 3 (01/23/2019)

- Introduction to inferential statistics
- Some commonly used statistical approaches

Lecture Objectives

- Clarify the goal of statistics
- Grasp descriptive statistics
- Be familiar with various data presentation approaches
- Introduce key concepts of inferential statistics
- Survey some commonly used statistical approaches
- Understand basic research design principles
- Understand the pros and cons of large databases
- Some examples of research using large databases
- Build a foundation which will facilitate the active participation in clinical and translational research
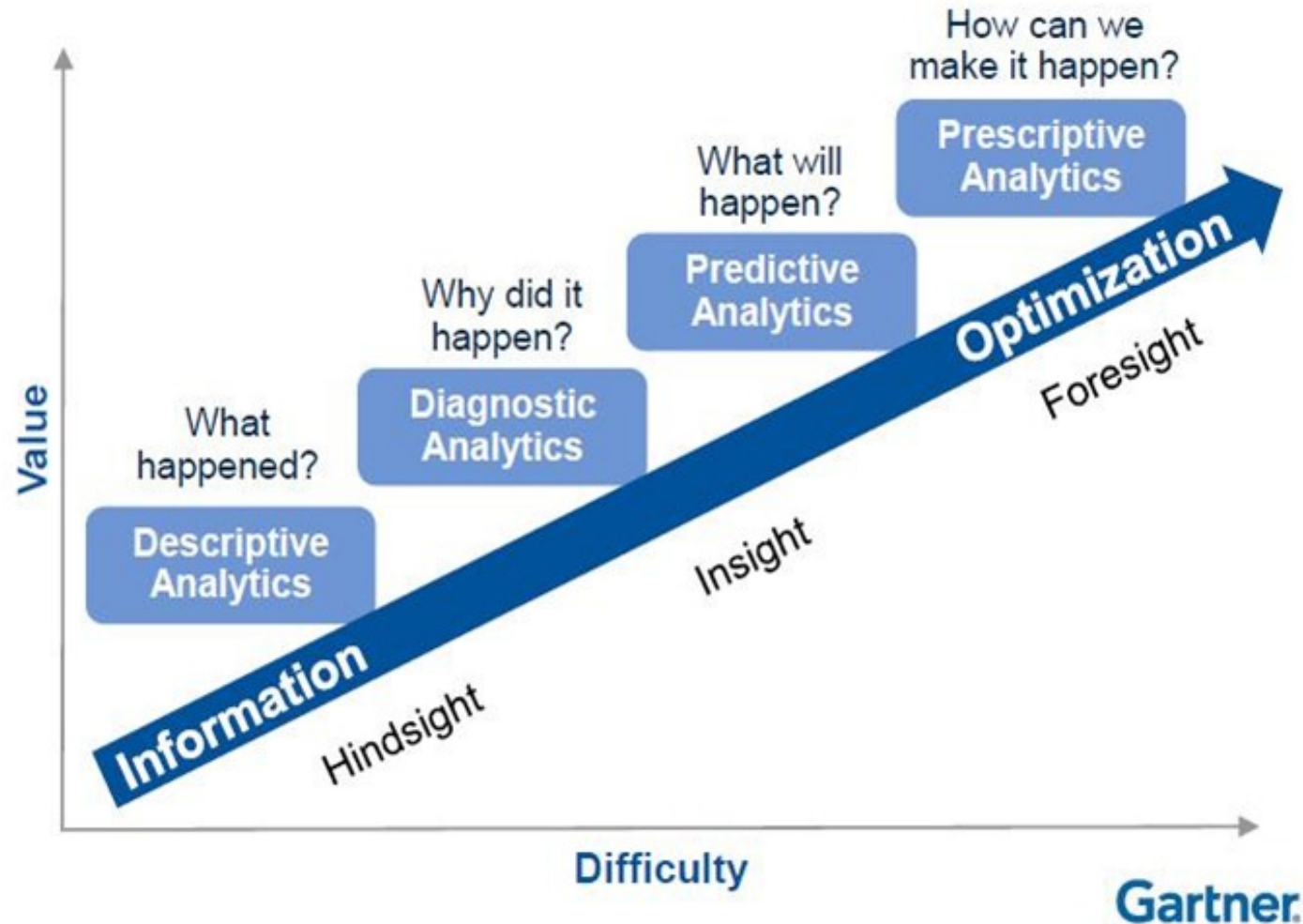
Growth of Healthcare Data
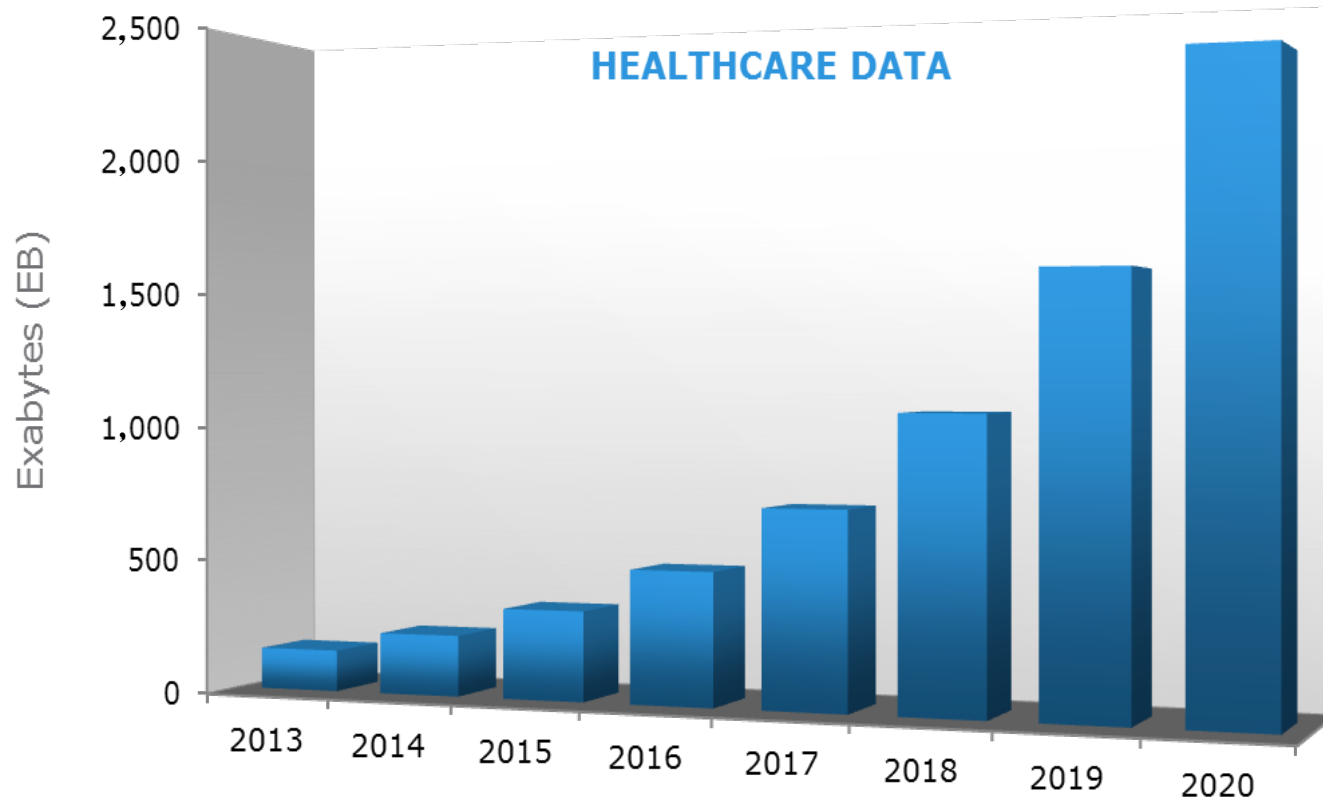
Enterprise Strategy Group 2011, Research report: North American Healthcare Provider
Information includes hospitals & Ambulatory Health Care Provider Market Size & Forecast
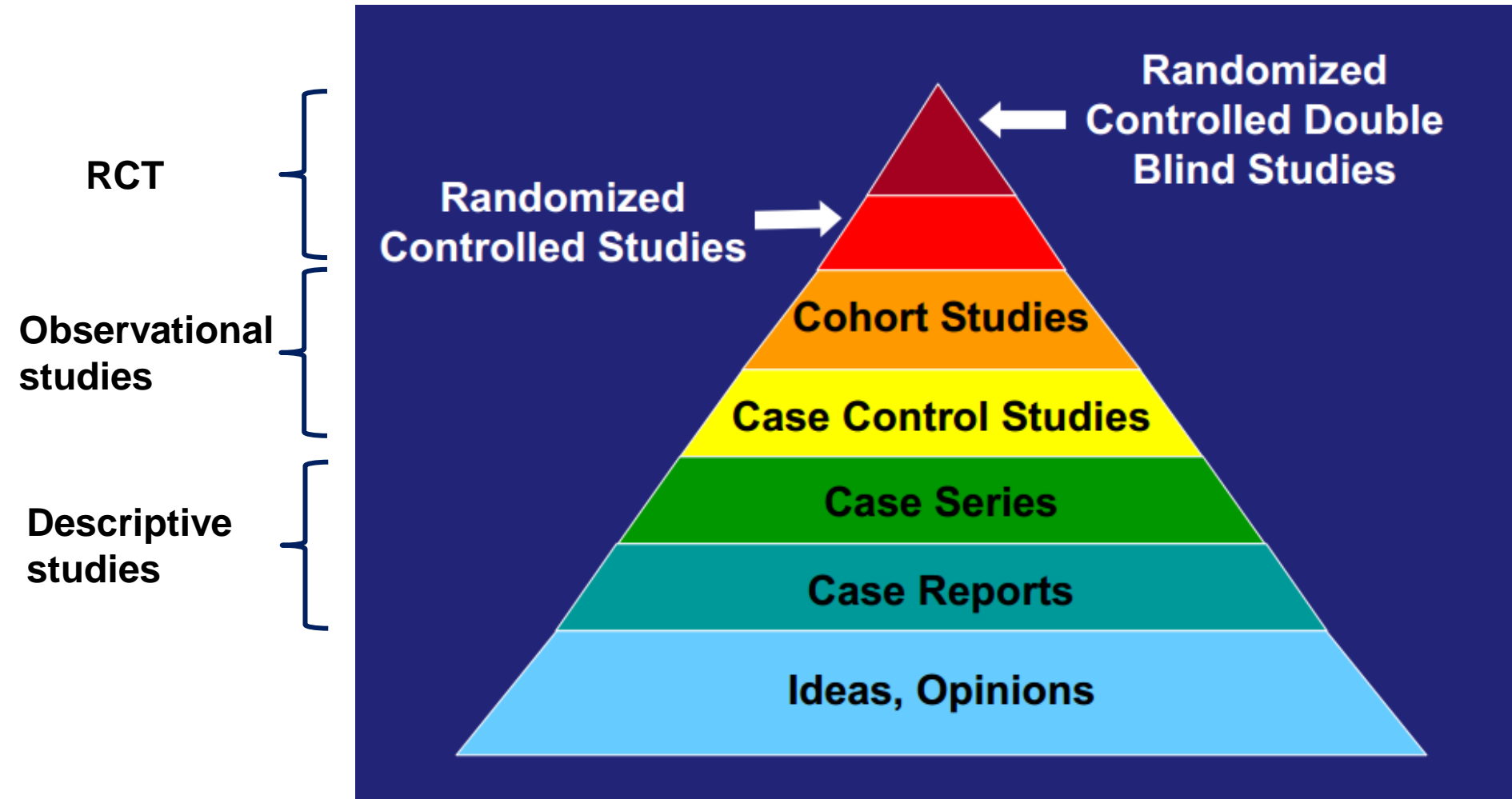
Source: EMC Digital Universe with Research & Analysis by IDC
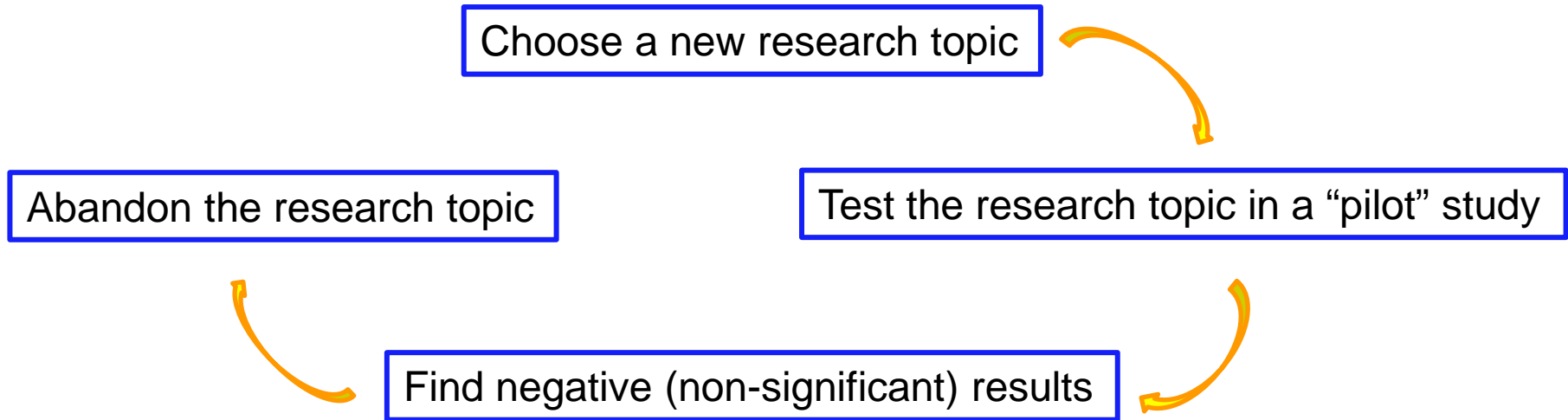
Analytics Are No Longer a Nice to Have

Clinical Research & Scientific Evidence

Barnett Kramer (NIH)

Nonproductive Research Strategy

Choose a new research topic

Test the research topic in a "pilot" study

Find negative (non-significant) results

Abandon the research topic

A "pilot" study?

- Little or no funding?
- A vague and poorly developed research proposal?
- No time for detailed thought about subsequent studies?

Lancaster, GA. Pilot and feasibility studies come of age! *Pilot and Feasibility Studies* 2015; 1:1.

Pilot and Feasibility Studies

Home   About   Articles   Submission Guidelines

Aims and scope
Fees and funding
Language editing services
Copyright
∨ Preparing your manuscript
Prepare supporting information
Conditions of publication
Editorial policies
Peer-review policy
Manuscript transfers
Promoting your publication

Aims and scope

Aims and scope

Pilot and Feasibility Studies encompasses all aspects of the design, conduct and reporting of pilot and feasibility studies in biomedicine. The journal publishes research articles that are intended to directly influence future clinical trials or large scale observational studies, as well as protocols, commentaries and methodology articles. The journal also ensures that the results of all well-conducted, peer-reviewed, pilot and feasibility studies are published, regardless of outcome or significance of findings.

Pilot and feasibility studies are increasingly conducted prior to a full randomized controlled trial. However, these studies often lack clear objectives, many remain unpublished, and there is confusion over the meanings of the words "pilot" and "feasibility". *Pilot and Feasibility Studies* provides a forum for discussion around this key aspect of the scientific process, and seeks to ensure that these studies are published, so as to complete the publication thread for clinical research.

Standard Research Process

Identifying a research question and a hypothesis

↓

Designing study and developing research protocol

↓

Gathering preliminary data and
revising the protocol

↓

Conducting the study

↓

Analyzing data analysis and interpreting results

↓

Drawing conclusions and disseminating the results

Database Research Process

Identifying a research question and a hypothesis

↓

Designing study and developing research protocol

↓

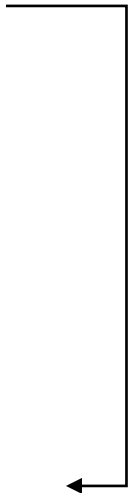Gathering preliminary data and
revising the protocol

↓

Conducting the study

↓

Analyzing data analysis and interpreting results

↓

Drawing and disseminating the conclusions

Primary versus Secondary Data

- Primary data: data collected specifically for the purpose of answering a research question and to meet specific research objectives

- Secondary data: data that have already been collected for some other purposes

Secondary data analysis: The practice of re-analyzing an existing data set for a purpose other than the purpose for which it was originally created.

Sources & Types of Secondary Data

- – Government-funded datasets
- – Records/files
- – Internet
- – Archives
- – Previous studies including journal supplements

Some are restricted and others are publicly available

Some Large Clinical Databases

Database	Description
American Hospital Association Annual Survey	Annual survey data from US hospitals with focus on organiztaional characteristics and health care utilization
AMA Physician Masterfile	Database of US physicians with information on education, training, and professional certification
Area Resource File (ARF)	County-level data on healthcare professionals, health facilities and utilization, and census data
Atherosclerosis Risk in Communities (ARIC)	Population cohort study of late-middle-age subjects and c... program of ca...
Behavioral Risk Factor Surveillance System (BRFSS)	Serial cross-se... health risk beh... practices, and
California Health Interview Survey (CHIS)	Serial cross-se... socioeconomic... environmental... children in Cal
Canadian Institute for Health Information databases	Series of datal...
Cardiovascular Health Study (CHS)	Population coh... on cardiovasc... issues
Community Tracking Study (CTS)	Longitudinal c... patients and p
Consumer Assessment of Healthcare Providers and Systems (CAHPS)	Family of surv... experience wit

Database	Description
Coronary Artery Risk Development in Young Adults (CARDIA)	Popul... focus
Health and Retirement Study (HRS) & Asset and Health Dynamics among the Oldest Old (AHEAD)	Longit... and e... socioe
HCUP State Inpatient Databases (SID)	Clinic... from i...
HCUP State Ambulatory Surgery Databases (SASD)	Clinic... from a... partici
HCUP State Emergency Department Databases (SEDD)	Clinic... from e... partici
HCUP Nationwide Inpatient Sample (NIS)	Natior... utiliza
HCUP Nationwide Emergency Department Sample (NEDS)	Natior... utiliza
Health Information National Trends Survey (HINTS)	Biannual telephone survey of the public's knowledge, attitudes, and behavior regarding cancer
HMO Cancer Research Network	Consortium of integrated health care del... sites that collect data on a wide range of aspects of cancer care

Database	Description
Medical Expenditure Panel Survey (MEPS)	Nationwide panel surveys of individuals, families, health care providers, and employers covering a variety of topics
Medical School Graduation Questionnaire and Matriculating Student Questionnaire	Nationwide surveys of medical students about career choices, medical school admissions process, and educational experiences during medical school
Medicare Claims Data	Claims d... demogra... wide var... settings
Medicare Current Beneficiary Survey (MCBS)	Panel su... use, hea... demogra... expendit
Midlife in the United States (MIDUS)	Longitud... adults in... socioder... characte... specializ... Health s... resident... physical... function
Minimum Data Set (Long-Term Care)	Clinic... from e... partici
National Ambulatory Medical Care Survey (NAMCS) & National Hospital Ambulatory Care Survey (NHAMCS)	Natior... sectiona... emerger... department visits

Database	Description
National Longitudinal Survey of Adolescent Health	National longitudinal survey of causes and outcomes of health-related behaviors in adolescents and young adults
National Program of Cancer Registries (NPCR)	Data assembled from state cancer registries, including information about patient demographics as well as cancer rates and counts by geographic area
National Registry of Myocardial Infarctions (NRMI)	Nationwide database of treatment for patients hospitalized with myocardial infarction
National Surgical Quality Improvement Program (NSQIP)	Ongoing data collection on major surgical procedures from U.S. hospitals including data on preoperative patient characteristics, intraoperative variables, and morbidity and mortality
National Survey of Ambulatory Surgery (NSAS)	Nationally-representative survey of ambulatory surgery in hospital-based and freestanding ambulatory surgery centers
Surveillance, Epidemiology and End Results Program (SEER), and SEER-Medicare data	National cancer registry database, available with or without linkages to Medicare data
U.S. Census data	Data tools provided by the U.S. Census Bureau; census data can be linked to other datasets, for example to obtain median income for each U.S. zip code
U.S. Renal Data System (USRDS)	Extensive nationwide data on patients with end-stage renal disease, collected on an ongoing basis

Some Education & Social Sciences Datasets

- Common Core of Data (CCD)
- Current Population Survey (CPS)
- Early Childhood Longitudinal Study (ECLS): Birth (ECLS-B) and Kindergarten (ECLS-K) Cohort
- General Social Survey (GSS)
- Head Start Family and Child Experiences Survey (FACES)
- Monitoring the Future (MTF)
- National Assessment of Educational Progress (NAEP)
- National Education Longitudinal Study (NELS)
- National Household Education Surveys (NHES)
- National Longitudinal Study of Adolescent Health (Add Health)
- National Longitudinal Survey of Youth (NLSY)
- National Survey of American Families (NSAF)
- National Survey of Child and Adolescent Well-Being (NSCAW)
- National Survey of Families and Households (NSFH)
- NICHD Study of Early Child Care and Youth Development (SECCYD)
- Programme for International Student Assessment (PISA)
- Progress in International Reading Literacy Study (PIRLS)
- Trends in International Mathematics and Science Study (TIMSS)
- U.S. Panel Study of Income Dynamics (PSID): Child Development Supplement (CDS)

Useful Websites for Secondary Databases

- Office of Minority Health: https://www.minorityhealth.hhs.gov/Default.aspx
- Society of General Internal Medicine: http://www.sgim.org/communities/research/dataset-compendium/proprietary-datasets
- UCSF – Large Dataset Inventory: http://ctsi.ucsf.edu/research/celdac
- Institute for Social Research University of Michigan: http://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp
- JABSOM CIM: http://oitwp02.jabsom.hawaii.edu/cim/research/healthcare-databases/

Behavioral Risk Factor Surveillance System (BRFSS)

- A large, cross-sectional annual survey data from 1984 (~500,000/year)
- Emphasize health-related risk behaviors, chronic health conditions, and use of preventive services
- Conducted at the state level and complied for public use
- Hawaii data identifiable

https://www.cdc.gov/brfss/

National Health and Nutrition Examination Survey (NHANES)

- A national biennial cross-sectional survey (~5,000 persons / year)
- A rich assortment of data based on
  - Interviews: demographics, socioeconomic status, dietary, and health behavior questions
  - Direct examination: medical, dental, physiological measurements, and lab tests

https://www.cdc.gov/nchs/nhanes/
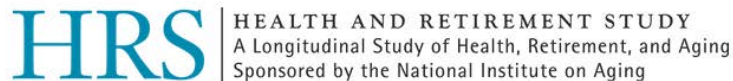
Health and Retirement Survey (HRS)

- A representative, longitudinal national study of aging, beginning at age 50
- Multiple cohorts of spouses, maintaining ~26,000 active participants
- Biennial examinations on all survivors, started in 1992

HRS | HEALTH AND RETIREMENT STUDY
A Longitudinal Study of Health, Retirement, and Aging
Sponsored by the National Institute on Aging

A public resource for data on aging in America since 1990

For more than 20 years, the Health and Retirement Study (HRS) has been a leading source for information on the health and economic well-being of adults over age 50 in the United States. We do this by asking people across the country to participate in the study and tell us about their lives.

As the world around us changes, the HRS follows the impact of these changes on our older population. This information has been used by scientists to publish over 3,000 books, articles and papers.

The HRS is a longitudinal project sponsored by the National Institute on Aging (NIA U01AG009740) and the Social Security Administration. The study director is Dr. David R. Weir of the Survey Research Center at the University of Michigan's Institute for Social Research.

http://hrsonline.isr.umich.edu/
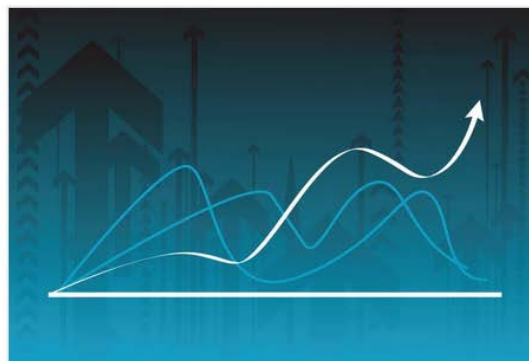
Hawaii Hospitalization Database

- 26 hospitals and medical centers, with over 100,000 hospitalizations per year

- Over 90 variables (e.g., age, payer, hospital and location of residence, admission/discharge dates, LOS, total charges, disposition)

- 13,000 ICD-9 codes (primary and 19 secondary diagnoses or procedures)

- 32 different ethnic categories, including over 20 Asians/PIs, besides NH

- Gathered in Hawaii: before 2017 by HHIC; after 2017 by Laulima Data Alliance

- State and National Healthcare Cost and Utilization Project (HCUP) data

https://www.hcup-us.ahrq.gov/

Pros of Large Clinical Databases

- All subjects, large population, real world patients and settings, ideal for comparative-effectiveness and cost-effectiveness research

- Often contain longitudinal data across multiple years

- Relatively low-cost and quick alternative to generating own data

- Convenient for exploratory analyses and developing research hypothesis

- Large sample sizes, allowing subgroup and multivariable analysis

- Education and training use

Cons of Large Clinical Databases

- Data may not be of research quality
- No control over what data is collected
- Lack of compatibility between data and research questions
- Usually very large and complex
- Require not only analytical and technical expertise, but familiarity and experience with specific databases

QHS 650: Secondary Data Analysis
QHS 651: Secondary Data Analysis Practicum

Spring 2019 (2-Credits Each)
Tuesday Afternoons (1:00-4:30 pm) at Kaka'ako Campus
in Health Sciences Library Computer Lab (MEB 107D)

Why Take This Course?

- Secondary data analysis takes advantage of data originally collected for other purposes in order to answer additional research questions.

- Many secondary data sources are readily available but underutilized. As such, they provide a rich opportunity for research and subsequent publications.

- QHS 650 provides a didactic overview of issues in secondary data analysis. This course will allow students new to using secondary data to become comfortable with accessing the data, forming hypotheses, and designing study proposals. In addition, the course introduces both basic and advanced statistical approaches for analyzing secondary data.

- The practicum course (QHS 651) is intended for students who plan to analyze secondary datasets in their own research. The course adds the hands-on experience needed to conduct individual research using secondary databases.

Course Objectives

- Learn how to locate and download public-release data;

- Learn the steps involved in conducting research using secondary data;

- Provide hands-on training for students to perform proper secondary data analysis using R (only for QHS 651).

Prerequisites: QHS 650: None; QHS 651: QHS 601 or equivalent, and QHS 651 (which can be taken concurrently with QHS 650)

Instructors: Eunjung Lim, Ph.D., & James Davis, Ph.D., Office of Biostatistics & Quantitative Health Sciences, Department of Complementary & Integrative Medicine, John A. Burns School of Medicine, University of Hawaii

Class Time: 1:00-2:40 pm for QHS 650; 2:50-4:30 pm for QHS 651

Master of Science in Clinical and Translational Research

The Clinical and Translational Research (CTR) graduate program will prepare graduates with skills for successful careers in clinical and translational research and research support.

Clinical Research (CR) Track

Develop knowledge and skills to investigate clinical research topics though coursework and research projects focused on research design, methodologies, quantitative methods, scientific writing, ethical issues, and the capacity in obtaining research funding.

Quantitative Health Sciences (QHS) Track

Courses and research projects focus on biostatistical and bioinformatic methods development and application to improve population and individual health. Students will acquire big data skills and master the scientific principles and methodologies that underlie basic science, clinical, and translation research.

Career

Research, research support, data analyst positions at:
- Academia
- Hospitals
- Government agencies
- Healthcare organizations
- Pharmaceutical companies

Program Curricula

- 2-year 34 total credit hours graduate program

- Plan A (Thesis): 24 credits of didactic courses

- Plan B (Capstone Project): 28 credits of didactic courses

How to Apply

Visit http://manoa.hawaii.edu/graduate/content/clinical-research to either fill out an application or download a PDF form
Application Deadline: May 30

For more information

Phone: (808) 692-1840
Email: GradCTR@hawaii.edu
Web: http://msctr.jabsom.hawaii.

MSCTR Graduate Program Website

https://msctr.jabsom.hawaii.edu/

Lecture Objectives

- Clarify the goal of statistics
- Grasp descriptive statistics
- Be familiar with various data presentation approaches
- Introduce key concepts of inferential statistics
- Survey some commonly used statistical approaches
- Understand basic research design principles
- Understand the pros and cons of large databases
- Some examples of research using large databases
- Build a foundation which will facilitate the active participation in clinical and translational research