Finding and Using Publicly Available Datasets for Secondary Data Analysis Research

SGIM Annual Meeting
April 30, 2010
Disclosures and acknowledgements

Disclosures:
None

Acknowledgements:
SGIM Research Committee
Bruce Landon, Ken Covinsky, John Ayanian, Christina Wee, Alex Smith, Sunil Kripalani
Learning objectives

- Appreciate key conceptual and methodologic issues involved in secondary data analysis
- Understand the range of resources and support required to successfully complete a secondary data analysis
- Identify and use online tools for locating and learning about publicly available datasets relevant to their research
Structure of session

- Working with secondary data
  - Conceptual and methodologic issues
  - J. Michael McWilliams, MD, PHD – Harvard Medical School

- Planning your dataset project
  - Practical advice
  - Ann Nattinger, MD, MPH – Medical College of Wisconsin

- SGIM Research Dataset Compendium and other web-based resources
  - Resources
  - Michael Steinman, MD – University of California, San Francisco

- Q&A
Overview

1. Types and uses of secondary data
2. Conceiving a project
3. Advantages
4. Challenges and pitfalls
5. Statistical considerations
6. Summary tips
(My) Definition of Secondary Data

Data that have been collected but not for you
Types of Secondary Data

- Survey (NHIS, NHANES, HRS, BRFSS)
- Administrative (Medicare claims)
- Discharge (HCUP SID and NIS)
- Medical chart / EMR
- Disease registries (SEER)
- Aggregate (ARF, US Census)
- Combinations and linkages
What kinds of research can be conducted with secondary data?

Anything but randomized trials

- By discipline:
  - Epidemiology
  - Outcomes research
  - Health services research

- By question:
  - Descriptive
  - Comparative
  - Causal
Conceiving a Project

- Which comes first: question or dataset?
  a. Research question first
  b. Dataset first
  c. Trick question: both

- Hybrid approach
  1. Identify research focus, broad question
  2. Consider candidate datasets
  3. Hone question
  4. Iterate between 2 and 3
What makes a good research question? (FINER)

- **Feasible**—data, variables, & resources accessible & available
- **Interesting**—to researcher and audience
- **Novel**—extends what is already known
- **Ethical**—upholds standards
- **Relevant**—to patient care, clinical outcomes, policy, etc.

Selecting a Database

- Compatibility with research question(s)
- Availability and expense
- Sample: representativeness, power
- Measures of interest present and valid
- Messiness and missingness
- Local expertise
- Linkages
Example of a Hybrid Approach

- General question: effects of near-universal Medicare on outcomes, implications for costs of care

- Explore candidate datasets
  - HRS: longitudinal, linked claims, but only self-reported health, cohort too young at first
  - NHANES: clinical measures of disease control, but cross-sectional data

- Hone questions: examine use of preventive services & mortality first, allow HRS to mature, examine variables, specify analytic plans, obtain restricted data

- New questions from familiarity with data
  - Effects of plan choice on enrollment in Medicare Advantage
  - Effects of Part D on hospitalizations and Part A & B spending
Advantages of Secondary Data

- They are not primary data!
  - Efficiency: fast and cheap
  - No regrets

- Scale and scope
  - Size and detail not otherwise feasible for individual research team
  - Generalizable

- Novel and creative research questions

- Often easier IRB review process
Challenges and Pitfalls

1. Data mining/overfitting
   - When the analysis precedes the question
   - “I just need to find my CRP”
   - Does urine cortisol predict Catholicism?

2. Causal inference
   - Inherently limited with observational data
   - But does not preclude quasi-experimental designs to recover causal effects
3. Validity of measures
   • Beware of assumptions
   • Problems: coding, reporting, recall biases
   • Solutions: direct validation in subgroup or another data source, literature review, factor analysis, sensitivity analyses

4. Complexity of file structure
   • Row in dataset may not be unit of analysis
   • Skip patterns, proxy respondents
Want to measure time preferences
  • Behavioral economics: people tend to overvalue the present
  • Explanation for unhealthy habits, underuse of cancer screening?

Have measures on financial planning horizons

Are the two equivalent?

Might financial planning also depend on:
  • Income
  • Source of income, employment status
  • Dependents
  • Inheritance
Ask: IF ((piRTab1X007AFinFam = FAMILYR) OR (piRTab1X007AFinFam = FINANCIAL_FAMILYR)) AND ((ACTIVELANGUAGE <> EXTENG) AND (ACTIVELANGUAGE <> EXTSPN)) AND (piInitA106_NumContactKids > 0) AND (piInitA100_NumNRKids > 0)

JE012 CHILDREN LIVE WITHIN 10 MILES
Section: E Level: Household Type: Numeric Width: 1 Decimals: 0
CAI Reference: SecE.KidStatus.E012_
2000 Link: G1980 2002 Link: HE012

IF {R DOES NOT HAVE SPOUSE/PARTNER and DOES NOT STILL HAVE HOME OUTSIDE NURSING HOME {{CS11/A028=1) and (CS26/A070 NOT 1)}} or {R & SPOUSE/PARTNER} LIVE IN SAME NURSING HOME (CS11/A028=1 and CS12/A030=1):
[Do any of your children who do not live with you/Does CHILD NAME] live within 10 miles of you (in R's NURSING HOME CITY, STATE (CS25b/A067))?  

OTHERWISE:  
[Do any of your children (who do not live with you)/Does CHILD NAME] live within 10 miles of you (in MAIN RESIDENCE [CITY/CITY, STATE STATE])?

6802 1. YES  
4720 5. NO  
32 8.DK (Don't Know); NA (Not Ascertained)  
4 9. RF (Refused)  
2087 Blank. INAP (Inapplicable)
5. **Representativeness of Sample**
   - External validity (generalizability)
   - Internal validity (selection bias)
   - Example: comparing outcomes for insured and uninsured patients using hospital discharge data
     - Must be hospitalized to enter sample
     - Not only limits generalizability (to outpatients)
     - But inferences about the sample may be wrong
       - Sample would need to include uninsured who would have been hospitalized if insured
Statistical Considerations: Missing Data

Sources
- Non-response: unit and item
- Variability in data collection (e.g. across states or over time, collected on subset due to expense)
- Incomplete linkages

Language
- MCAR: $M \perp Y \Rightarrow$ strong assumption, can ignore
- MAR: $M \perp Y | X \Rightarrow$ weaker assumption, can fix
- Non-ignorable, informative: $M$ predicts $Y \Rightarrow$ can’t fix
Approaches

- Listwise deletion, complete case (ok if MCAR)
- LVCF, ad hoc imputation (biased estimates)
- Mean imputation (biased standard errors)
- Multiple imputation (MAR)
- Weighting techniques (MAR)
- Random effects models (MAR)
Statistical Considerations: Analyzing Survey Data

Survey design
- **Stratification**: ↑representativeness (protects against a bad sample), ↑precision
- **Clustering**: convenience, ↓precision
- **Oversampling**: ↑representation and precision for subgroup of interest

Example multistage probability sample:
US divided into strata → sample of PSUs (MSAs) selected in each stratum → sample of SSUs (census blocks) in each PSU → sample of households in each SSU → all household members surveyed
Survey weights: affect point estimates
- Individuals may have unequal selection probabilities
- Need to apply weights to recover representativeness
  - \( W = \frac{1}{p({\text{selection}})} = \# \text{ people represented} \)
- \( W \)'s reflect sampling design, adjustments to match to census totals, non-response

Survey strata, clusters: affect se’s
- Need variance estimators that account for correlated data
Design–based estimation

- Survey packages in most statistical software (e.g. SAS: surveyreg with weight, strata, cluster statements)
- Based on sampling theory, not model assumptions (residuals~iid N)
- Robust: no distributional assumptions, data may be correlated
- Only need to specify PSUs in multistage design
Summary: 10 Tips for Success in Secondary Data Analyses

1. Start with a clear research question and hypothesis
2. Get to know your data source:
   - Why does the database exist?
   - Who reports the data?
   - What are the incentives for accurate reporting?
   - How are the data audited, if at all?
   - Can you link the data to other large databases?
3. Get good documentation of the cohort, variables, and data layout, then read the fine print
4. Consult or collaborate with researchers who have used the database

Provided by John Ayanian, MD, MPP, Ellen McCarthy, PhD, “Research with Large Databases”, Harvard School of Public Health
Summary: 10 Tips for Success in Secondary Data Analyses

5. Line up computing resources before data arrive
6. Allow time to receive data if not publicly available
7. Learn SAS, Stata, or other statistical software so you can analyze data yourself (or collaborate)
8. Assess data quality (e.g., outliers & missing data) with plots or frequency tables
9. Consult or collaborate with a statistician on your analysis plan, especially for complex surveys with sampling weights
10. Use clinical intuition to interpret results and consult experts as needed

Provided by John Ayanian, MD, MPP, Ellen McCarthy, PhD, “Research with Large Databases”, Harvard School of Public Health
Planning Your Dataset Project

Ann B. Nattinger, MD, MPH
SGIM Annual Meeting
April 30, 2010
Relate Conceptual Model and Available Measures
Conceptual Model

- **Predisposing Characteristics** (Age, race, ethnicity, marital status)
- **Need** (extent of disease, presence and severity of health conditions)
- **Enabling Resources** (SES, other drug coverage, neighborhood characteristics)

- **Part D Enrollment**
- **Access to and use of prescribed drugs**
- **Mortality**
- **Subsequent use of health services**

- Information and availability of alternatives (number of Area Networks on Aging, presence and generosity of state drug plans).
### Table 8. Information to be Obtained and Source

<table>
<thead>
<tr>
<th>PATIENT BASELINE CHARACTERISTICS</th>
<th>SOURCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socio-demographics: age, race/ethnicity, marital status</td>
<td>Medicare claims</td>
</tr>
<tr>
<td>Health status and comorbid conditions</td>
<td>Medicare claims</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PATIENT SOCIOECONOMIC STATUS</th>
<th>SOURCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual level poverty indicator based on dual enrollee status; enrollment in state buy-in low income subsidy programs (QMBs/SLMBs)</td>
<td>Medicare denominator file; published States program eligibility rules (Green Book)</td>
</tr>
<tr>
<td>Per capita income in neighborhood</td>
<td>Medicare Denominator and Medicare Beneficiary Address Files geocoded to Census Tracts; Census Tract data</td>
</tr>
</tbody>
</table>

| Proportion of persons in neighborhood in specific Federal Poverty groups (≤ 100% FPL, 100-150% FPL; 150-200% FPL; 200-300% FPL, >300% FPL) | Medicare Denominator and Medicare Beneficiary Address Files geocoded to Census Tracts; Census Tract data |
| Proportion of persons in neighborhood belonging to specific racial/ethnic groups: African American/Black; Hispanics; Other non-Caucasian | Medicare Denominator and Medicare Beneficiary Address Files geocoded to Census Tracts; Census Tract data |

<table>
<thead>
<tr>
<th>MEDICARE PART D ELIGIBILITY, ENROLLMENT, and UTILIZATION</th>
<th>SOURCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part D enrollment/coverage; Source of Part D drug coverage; Availability and source of other creditable drug coverage; Part D adjuvant HT claims, and number of overall prescription drug claims</td>
<td>Medicare Enrollment Beneficiary Database and Part D claims</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SCREENING MAMMOGRAM RATES by SES groups</th>
<th>SOURCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Medicare claims for 5% controls</td>
</tr>
</tbody>
</table>
Data Source(s)

- Includes needed measures
  - Direct measure or require algorithm?

- Validity of information
Data Validity: AMA Database

- Data from primary sources
- Data from survey of doctors
- Some data incomplete → how to deal with this
- How to merge to other dataset
Data Availability

- Available to public or restricted
- Approval and review policies
IRB Issues

- Exempt if no identifiers
  - Still need IRB determination of exempt status
- Limited data set – DUA
  - Usually expedited for IRB
- Full data use agreement
  - Still may be expedited for IRB
  - May need to explain lack of consent to funders/IRB
Resources Needed

- Your effort
- Computer resources and security
- Programmer and/or statistician effort
- PhD statistical support – complex sampling or analyses
- Coordinator if merging datasets
- Time timeline
<table>
<thead>
<tr>
<th>Task</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obtain, process and clean INP, OPT, Part B claims</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apply algorithm for identification of breast cancer cases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohorts selection: 2001-2006; 2007-08 incidence cohorts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop comorbidity measures, including DCG and HCC methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obtain and process 5% Medicare data for mammography rates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obtain and process Part D enrollment, Part D claims and third-party</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>payer data: 2006, 2007 and 2008 cohorts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conduct quality control and consistency checks – Part D data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geocode address data to census tracts of residence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop SES measures at census tract level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop database on potential instrumental variables for 50 states</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obtain, process and clean National Death index data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop cause of death algorithm based on National Death Index</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop mortality measures at 36, 60 months, and survival time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Construct hormone therapy utilization measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combine data from multiple sources</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Construct variables and developing analytical files</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finalizing analytical plans for HT use and adherence measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop and validate mortality models at 36, 60 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop and validate hormone therapy utilization models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop and validate SES mortality by HT adherence models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prepare manuscripts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disseminate findings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SGIM Research Dataset Compendium and Other Web-Based Research Resources

A Quick Tour

Michael Steinman, MD
SGIM Annual Meeting
April 30, 2010
• Amita is a junior faculty member interested in doing a secondary data analysis project on association between race/ethnicity and the prevalence and outcomes of atrial fibrillation. No prior experience and limited local mentorship.

• Eric is a mid–level faculty member with past experience. Wants to find new dataset around which write grant on association between SES and ADL function in elders.
Amita – Getting Started

- Amita
  - Get acquainted with basics
  - Find dataset and assess merit and feasibility
  - Find a mentor / get expert help

- www.sgim.org/go/datasets
How to use this site:

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

Featured datasets: Descriptions and expert evaluations of high-value datasets.

User’s guide to working with secondary data: Tips for working with secondary data.

Other dataset compendia, repositories, and resources: Brief descriptions and links to other dataset compendia, data repositories, and miscellaneous resources.

Mentor / mentee connect: Venue for connecting mentors and mentees for research involving secondary data.

Request a consultation: Request a brief individualized consultation about a dataset with an expert user. This service is available to SGIM members only.

About: Acknowledgments and information about this compendium.
Get Acquainted with Basics

How to use this site:

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

Featured datasets: Descriptions and expert evaluations of high-value datasets.

User’s guide to working with secondary data: Tips for working with secondary data.

Other dataset compendia, repositories, and resources: Brief descriptions and links to other dataset compendia, data repositories, and miscellaneous resources.

Mentor / mentee connect: Venue for connecting mentors and mentees for research involving secondary data.

Request a consultation: Request a brief individualized consultation about a dataset with an expert user. This service is available to SGIM members only.

About: Acknowledgments and information about this compendium.
User's Guide to Working With Secondary Data

Please click on any of the following links. We encourage readers to read each of the sections in sequential order.

- Which comes first - the dataset or the research question?
- Opportunities, pitfalls, and common mistakes in working with secondary data
- Issues in data coding and validation
- Statistical issues in working with secondary data
- IRB issues
- Other general tips
- Resources for secondary data analysis of specific topics and datasets
How to use this site:

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

**Featured datasets:** Descriptions and expert evaluations of high-value datasets.

**User’s guide to working with secondary data:** Tips for working with secondary data.

**Other dataset compendia, repositories, and resources:** Brief descriptions and links to other dataset compendia, data repositories, and miscellaneous resources.

**Mentor / mentee connect:** Venue for connecting mentors and mentees for research involving secondary data.

**Request a consultation:** Request a brief individualized consultation about a dataset with an expert user. This service is available to SGIM members only.

**About:** Acknowledgments and information about this compendium.
# Featured Datasets: Description

For a list of datasets by topic, [click here](#).

Click on the dataset name for more detailed information about the dataset.

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Hospital Association Annual Survey</td>
<td>Annual survey data from US hospitals with focus on organizational characteristics and health care utilization</td>
</tr>
<tr>
<td>AMA Physician Masterfile</td>
<td>Database of US physicians with information on education, training, and professional certification</td>
</tr>
<tr>
<td>Area Resource File (ARF)</td>
<td>County-level data on healthcare professionals, health facilities and utilization, and census data</td>
</tr>
<tr>
<td>Atherosclerosis Risk in Communities (ARIC)</td>
<td>Population cohort study of late-middle age subjects and community surveillance program of cardiac events</td>
</tr>
<tr>
<td>Behavioral Risk Factor Surveillance System (BRFSS)</td>
<td>Serial cross-sectional national survey of health risk behaviors, preventative health practices, and health care access</td>
</tr>
<tr>
<td>California Health Interview Survey (CHIS)</td>
<td>Serial cross-sectional survey of health, socioeconomic, behavioral, and environmental characteristics of adult and children in California</td>
</tr>
<tr>
<td>Canadian Institute for Health Information databases</td>
<td>Series of databases from Canada</td>
</tr>
<tr>
<td>Cardiovascular Health Study (CHS)</td>
<td>Population cohort study of elders with focus on cardiovascular disease and geriatric issues</td>
</tr>
</tbody>
</table>
## Featured Datasets: Topic Grid

For a list of datasets with brief descriptions, [click here](#).

Column headings highlight key content areas of each dataset:

- Health Systems - health care utilization and organization, physicians and health professional workforce
- Medical Education
- Health Behaviors
- Access, Costs, Social - health insurance and access, health care costs, social status
- Health Status and Health Care - health status, ambulatory, and inpatient care

<table>
<thead>
<tr>
<th>NAME</th>
<th>Health System</th>
<th>Medical Education</th>
<th>Health Behaviors</th>
<th>Access, Costs, Social</th>
<th>Health Status &amp; Health Care</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Hospital Association Annual Survey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMA Physician Masterfile</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area Resource File (ARF)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atherosclerosis Risk in Communities (ARIC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Behavioral Risk Factor Surveillance System (BRFSS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>California Health Interview Survey (CHIS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X, X</td>
</tr>
<tr>
<td>Canadian Institute for Health Information Databases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X, X, X</td>
</tr>
<tr>
<td>Cardiovascular Health Study (CHS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Coronary Artery Risk Development in Young Adults (CARDIA)

Population cohort study of young adults with focus on cardiovascular disease

Coronary Artery Risk Development in Young Adults

- Key web links
- Dataset summary and expert comments
- Dataset details
  - Owner / manager
  - Study and sample characteristics
  - Major foci
  - Special supplements and resources
  - Links to other datasets
  - Papers published
  - Dataset accessibility and cost
  - Help desk
- Request a consultation (SGIM members only)
Study and sample characteristics

Longitudinal population cohort study of 5,115 black and white adults from 4 study sites in the United States. Subjects were age 18-30 at enrollment in 1985-86, and have been followed in 6 subsequent evaluations, the most recent of which was 2005-06.

Major foci

Major topic areas covered by CARDIA are listed below. Please note that within each topic area, certain questions have been asked only for certain years.

Blood pressure
Chemistries
Anthropometry
Medical history (including sociodemographics and health behaviors)
Family history
Physical activity/fitness
Nutrient intake/dietary history
Obesity questionnaires
Psychosocial parameters (including a wide variety of data including social support, job strain, discrimination, caregiving stress, neighborhood cohesion, and many others)
Pulmonary function
Electrocardiogram and echocardiography
Coronary calcium
Carotid intimal medial thickness
Genetic studies
Dataset accessibility and cost

Access to CARDIA data is granted through the NHLBI, which provides free limited-access datasets for researchers upon completion of an application process.

See: http://www.nhlbi.nih.gov/resources/deca.descriptions/cardia.htm

CARDIA is a unique epidemiological study that is well suited for answering questions of how heart disease develops in young adults. Three features are important to note. First, these are young adults at the start of the study (18-30 years old) who are community dwelling and presumed free of disease. This cohort is about to enter its year 25 exam and participants have had regular study visits over this relatively long interval. Second, the cohort is ½ black, ½ white, ½ men, ½ women, ½ low SES, ½ high SES - because of these design features, the cohort is unique in its ability to address questions at the intersection of race/sex/SES in young adults. Third, study visits have include multiple in-depth questionnaires related to behavior, psychosocial issues, physiological measurements, and adjudicated cardiovascular outcomes. The number of clinical events thus far in CARDIA is relatively small, because the cohort is young. If you are looking for a large number of individuals with a particularly clinical condition, CARDIA is probably not the right dataset for you. On the other hand, if you are interested in issues of prevention and the period before the development of clinical disease, CARDIA is ideal.
Get Expert Help

How to use this site:

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

Featured datasets: Descriptions and expert evaluations of high-value datasets.

User’s guide to working with secondary data: Tips for working with secondary data.

Other dataset compendia, repositories, and resources: Brief descriptions and links to other dataset compendia, data repositories, and miscellaneous resources.

Mentor / mentee connect: Venue for connecting mentors and mentees for research involving secondary data.

Request a consultation: Request a brief individualized consultation about a dataset with an expert user. This service is available to SGIM members only.

About: Acknowledgments and information about this compendium.
Getting Expert Help

- Request a consultation
  - 1 on 1 consultation
  - Clear, defined questions about dataset
    - “strengths and weaknesses about using XYZ to study patterns of medication use for heart failure”

- Mentor–mentee connect
  - Link mentors with mentees
  - May be phased out
Eric – Getting Down to Business

- Identify datasets relevant to his research interests
- Identify health statistics, validated instruments, funding sources

www.sgium.org/go/datasets
### Featured Datasets: Description

For a list of datasets by topic, [click here](#).

Click on the dataset name for more detailed information about the dataset.

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Hospital Association Annual Survey</td>
<td>Annual survey data from US hospitals with focus on organizational characteristics and health care utilization</td>
</tr>
<tr>
<td>AMA Physician Masterfile</td>
<td>Database of US physicians with information on education, training, and professional certification</td>
</tr>
<tr>
<td>Area Resource File (ARF)</td>
<td>County-level data on healthcare professionals, health facilities and utilization, and census data</td>
</tr>
<tr>
<td>Atherosclerosis Risk in Communities (ARIC)</td>
<td>Population cohort study of late-middle age subjects and community surveillance program of cardiac events</td>
</tr>
<tr>
<td>Behavioral Risk Factor Surveillance System (BRFSS)</td>
<td>Serial cross-sectional national survey of health risk behaviors, preventative health practices, and health care access</td>
</tr>
<tr>
<td>California Health Interview Survey (CHIS)</td>
<td>Serial cross-sectional survey of health, socioeconomic, behavioral, and environmental characteristics of adult and children in California</td>
</tr>
<tr>
<td>Canadian Institute for Health Information databases</td>
<td>Series of databases from Canada</td>
</tr>
<tr>
<td>Cardiovascular Health Study (CHS)</td>
<td>Population cohort study of elders with focus on cardiovascular disease and geriatric issues</td>
</tr>
</tbody>
</table>
Finding Additional Resources

- National Information Center on Health Services Research and Health Care Technology (NICHSR)
- Inter–University Consortium for Political and Social Research (ICPSR)
- Partners in Information Access for the Public Health Workforce
- Roadmap K–12 Data Resource Center (UCSF)
- List of datasets from the American Sociologic Association
- Canadian Research Data Centers – Data Sets and Research Tools (Canada)
- Directory of Health and Human Services Data Resources
- Publicly Available Databases from National Institute on Aging (NIA)
- Publicly Available Databases from National Heart, Lung, & Blood Institute (NHLBI)
- National Center for Health Statistics (NCHS) Data Warehouse
- Medicare Research Data Assistance Center (RESDAC); and Centers for Medicare and Medicaid Services (CMS) Research, Statistics, Data & Systems
- Veterans Affairs (VA) data
National Information Center on Health Services Research and Health Care Technology (NICHSR)

- Databases, data repositories, health statistics
- Fellowship and funding opportunities
- Glossaries, research and clinical guidelines
- Evidence-based practice and health technology assessment
- Specialized PubMed searches on healthcare quality and costs
Inter-University Consortium for Political and Social Research (ICPSR)

- World’s largest archive of social science data
- Searchable
- Many sub-archives relevant to HSR
  - Health and Medical Care Archive
  - National Archive of Computerized Data on Aging