Conducting High-Value Secondary Dataset Analysis: An Introductory Guide and Resources

Alexander K. Smith, MD, MPH[1,2], John Z. Ayanian, MD, MPP[3,4,5], Kenneth E. Covinsky, MD, MPH[1,2], Bruce E. Landon, MD, MBA, MSc[3,4,6], Ellen P. McCarthy, PhD, MPH[3,6], Christina C. Wee, MD, MPH[3,6], and Michael A. Steinman, MD[1,2]

[1]Division of Geriatrics, Department of Medicine, University of California, San Francisco, San Francisco, CA, USA; [2]Veterans Affairs Medical Center, San Francisco, CA, USA; [3]Harvard Medical School, Boston, MA, USA; [4]Department of Health Care Policy, Harvard School of Public Health, Boston, MA, USA; [5]Division of General Medicine, Brigham and Women's Hospital, Boston, MA, USA; [6]Division of General Medicine and Primary Care, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA.

Secondary analyses of large datasets provide a mechanism for researchers to address high impact questions that would otherwise be prohibitively expensive and time-consuming to study. This paper presents a guide to assist investigators interested in conducting secondary data analysis, including advice on the process of successful secondary data analysis as well as a brief summary of high-value datasets and online resources for researchers, including the SGIM dataset compendium (www.sgim.org/go/datasets). The same basic research principles that apply to primary data analysis apply to secondary data analysis, including the development of a clear and clinically relevant research question, study sample, appropriate measures, and a thoughtful analytic approach. A real-world case description illustrates key steps: (1) define your research topic and question; (2) select a dataset; (3) get to know your dataset; and (4) structure your analysis and presentation of findings in a way that is clinically meaningful. Secondary dataset analysis is a well-established methodology. Secondary analysis is particularly valuable for junior investigators, who have limited time and resources to demonstrate expertise and productivity.

KEY WORDS: large datasets; secondary analysis; publicly available; guide; resources.
J Gen Intern Med 26(8):920–9
DOI: 10.1007/s11606-010-1621-5

INTRODUCTION

Secondary data analysis is analysis of data that was collected by someone else for another primary purpose. Increasingly, generalist researchers start their careers conducting analyses of existing datasets, and some continue to make this the focus of their career. Using secondary data enables one to conduct studies of high-impact research questions with dramatically less time and resources than required for most studies involving primary data collection. For fellows and junior faculty who need to demonstrate productivity by completing and publishing research in a timely manner, secondary data analysis can be a key foundation to successfully starting a research career. Successful completion demonstrates content and methodological expertise, and may yield useful data for future grants. Despite these attributes, conducting high quality secondary data research requires a distinct skill set and substantial effort. However, few frameworks are available to guide new investigators as they conduct secondary data analyses.[1–3]

In this article we describe key principles and skills needed to conduct successful analysis of secondary data and provide a brief description of high-value datasets and online resources. The primary target audience of the article is investigators with an interest but limited prior experience in secondary data analysis, as well as mentors of these investigators, who may find this article a useful reference and teaching tool. While we focus on analysis of large, publicly available datasets, many of the concepts we cover are applicable to secondary analysis of proprietary datasets. Datasets we feature in this manuscript encompass a wide range of measures, and thus can be useful to evaluate not only one disease in isolation, but also its intersection with other clinical, demographic, and psychosocial characteristics of patients.

REASONS TO CONDUCT OR TO AVOID A SECONDARY DATA ANALYSIS

Many worthwhile studies simply cannot be done in a reasonable timeframe and cost with primary data collection. For example, if you wanted to examine racial and ethnic differences in health services utilization over the last 10 years of life, you could enroll a diverse cohort of subjects with chronic illness and wait a decade (or longer) for them to die, or you could find a dataset that includes a diverse sample of decedents. Even for less dramatic examples, primary data collection can be difficult without incurring substantial costs, including time and money—scarce resources for junior researchers in particular. Secondary datasets, in contrast, can provide access to large sample sizes, relevant measures, and longitudinal data, allowing junior investigators to formu-

Received August 13, 2010
Revised December 9, 2010
Accepted December 17, 2010
Published online February 8, 2011

late a generalizable answer to a high impact question. For those interested in conducting primary data collection, beginning with a secondary data analysis may provide a “bird’s eye view” of epidemiologic trends that future primary data studies examine in greater detail.

Secondary data analyses, however, have disadvantages that are important to consider. In a study focused on primary data, you can tightly control the desired study population, specify the exact measures that you would like to assess, and examine causal relationships (e.g., through a randomized controlled design). In secondary data analyses, the study population and measures collected are often not exactly what you might have chosen to collect, and the observational nature of most secondary data makes it difficult to assess causality (although some quasi-experimental methods, such as instrumental variable or regression discontinuity analysis, can partially address this issue). While not unique to secondary data analysis, another disadvantage to publicly available datasets is the potential to be “scooped,” meaning that someone else publishes a similar study from the same data set before you do. On the other hand, intentional replication of a study in a different dataset can be important in that it either supports or refutes the generalizability of the original findings. If you do find that someone has published the same study using the same dataset, try to find a unique angle to your study that builds on their findings.

STEPS TO CONDUCTING A SUCCESSFUL SECONDARY DATA ANALYSIS

The same basic research principles that apply to studies using primary data apply to secondary data analysis, including the development of a clear research question, study sample, appropriate measures, and a thoughtful analytic approach. For purposes of secondary data analysis, these principles can be conceived as a series of four key steps, described in Table 1 and the sections below. Table 2 provides a glossary of terms used in secondary analysis including dataset types and common sampling terminology.

Define your Research Topic and Question

Case. A fellow in general medicine has a strong interest in studying palliative and end-of-life care. Building on his interest in racial and ethnic disparities, he wants to examine disparities in use of health services at the end of life. He is leaning toward conducting a secondary data analysis and is not sure if he should begin with a more focused research question or a search for a dataset.

Investigators new to secondary data research are frequently challenged by the question “which comes first, the question or the dataset?” In general, we advocate that researchers begin by defining their research topic or question. A good question is essential—an uninteresting study with a huge sample size or extensively validated measures is still uninteresting. The answer to a research question should have implications for patient care or public policy. Imagine the possible findings and ask the dreaded question: "so what?" If possible, select a question that will be interesting regardless of the direction of the findings: positive or negative. Also, determine a target audience who would find your work interesting and useful.

It is often useful to start with a thorough literature review of the question or topic of interest. This effort both avoids duplicating others’ work and develops ways to build upon the literature. Once the question is established, identify datasets that are the best fit, in terms of the patient population, sample size, and measures of

Table 1. A Practical Approach to Successful Research with Large Datasets

Steps	Practical advice
(1) Define your research topic and question	(1) Start with a thorough literature review (2) Ensure that the research question has clinical or policy relevance and is based on sound a priori reasoning. A good question is what makes a study good, not a large sample size (3) Be flexible to adapt your question to the strengths and limitations of the potential datasets
(2) Select a dataset	(1) Use a resource such as the Society of General Internal Medicine’s Online Compendium (www.sgim.org/go/datasets) (Table 3) (2) To increase the novelty of your work, consider selecting a dataset that has not been widely used in your field or link datasets together to gain a fresh perspective (3) Factor in complexity of the dataset (4) Factor in dataset cost and time to acquire the actual dataset (5) Consider selecting a dataset your mentor has used previously
(3) Get to know your dataset	(1) Learn the answers to the following questions: •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 your dataset to other large datasets? (2) Read everything you can about the database (3) Check to see if your measures have been validated against other sources (4) Get a close feel for the data by analyzing it yourself or closely reviewing outputs if someone else is doing the programming
(4) Structure your analysis and presentation of findings in a way that is clinically meaningful	(1) Think carefully about the clinical implications of your findings (2) Be cautious when interpreting statistical significance (i.e., p-values). Large sample sizes can yield associations that are highly statistically significant but not clinically meaningful (3) Consult with a statistician for complex datasets and analyses (4) Think carefully about how you portray the data. A nice figure sometimes tells the story better than rows of data

Table 2. Glossary of Terms Used in Secondary Dataset Analysis Research

Term	Meaning
Types of datasets (not mutually exclusive)	
Administrative or claims data	Datasets generated from reimbursement claims, such as ICD-9 codes used to bill for clinical encounters, or discharge data such as discharge diagnoses
Longitudinal data	Datasets that measure factors of interest within the same subjects over time
Clinical registries	Datasets generated from registries of specific clinical conditions, such as regional cancer registries used to create the Surveillance Epidemiology and End Results Program (SEER) dataset
Population-based survey	A target population is available and well-defined, and a systematic approach is used to select members of that population to take part in the study. For example, SEER is a population-based survey because it aims to include data on all individuals with cancer cared for in the included regions
Nationally representative survey	Survey sample that is designed to be representative of the target population on a national level. Often uses a complex sampling scheme. The Health and Retirement Study (HRS), for example, is nationally representative of community-dwelling adults over age 50
Panel survey	A longitudinal survey in which data are collected in the same panel of subjects over time. As one panel is at the middle or end of its participation, a panel of new participants is enrolled. In the Medical Expenditures Panel Survey (MEPS), for example, individuals in the same household are surveyed several times over the course of 2 years
Statistical sampling terms	
Clustering	Even simple random samples can be prohibitively expensive for practical reasons such as geographic distance between selected subjects. Identifying subjects within defined clusters, such as geographic regions or subjects treated by the same physicians, reduces cost and improves the feasibility of the study but may decrease the precision of the estimated variance (e.g., wider confidence intervals)
Complex survey design	A survey design that is not a simple random selection of subjects. Surveys that incorporate stratification, clustering and oversampling (with patient weights) are examples of complex data. Statistical software is available that can account for complex survey designs and is often needed to generate accurate findings
Oversampling	Intentionally sampling a greater proportion of a subgroup, increasing the precision of estimates for that subgroup. For example, in the HRS, African-Americans, Latinos, and residents of Florida are oversampled (see also survey weights)
Stratification	In stratification, the target population is divided into relatively homogeneous groups, and a pre-specified number of subjects is sampled from within each stratum. For example, in the National Ambulatory Medical Care Survey physicians are divided by specialty within each geographic area targeted for the survey, and a certain number of each type of physician is then identified to participate and provide data about their patients
Survey weights	Weights are used to account for the unequal probability of subject selection due to purposeful over- or under-sampling of certain types of subjects and non-response bias. The survey weight is the inverse probability of being selected. By applying survey weights, the effects of over- and under-sampling of certain types of patients can be corrected such that the data are representative of the entire target population

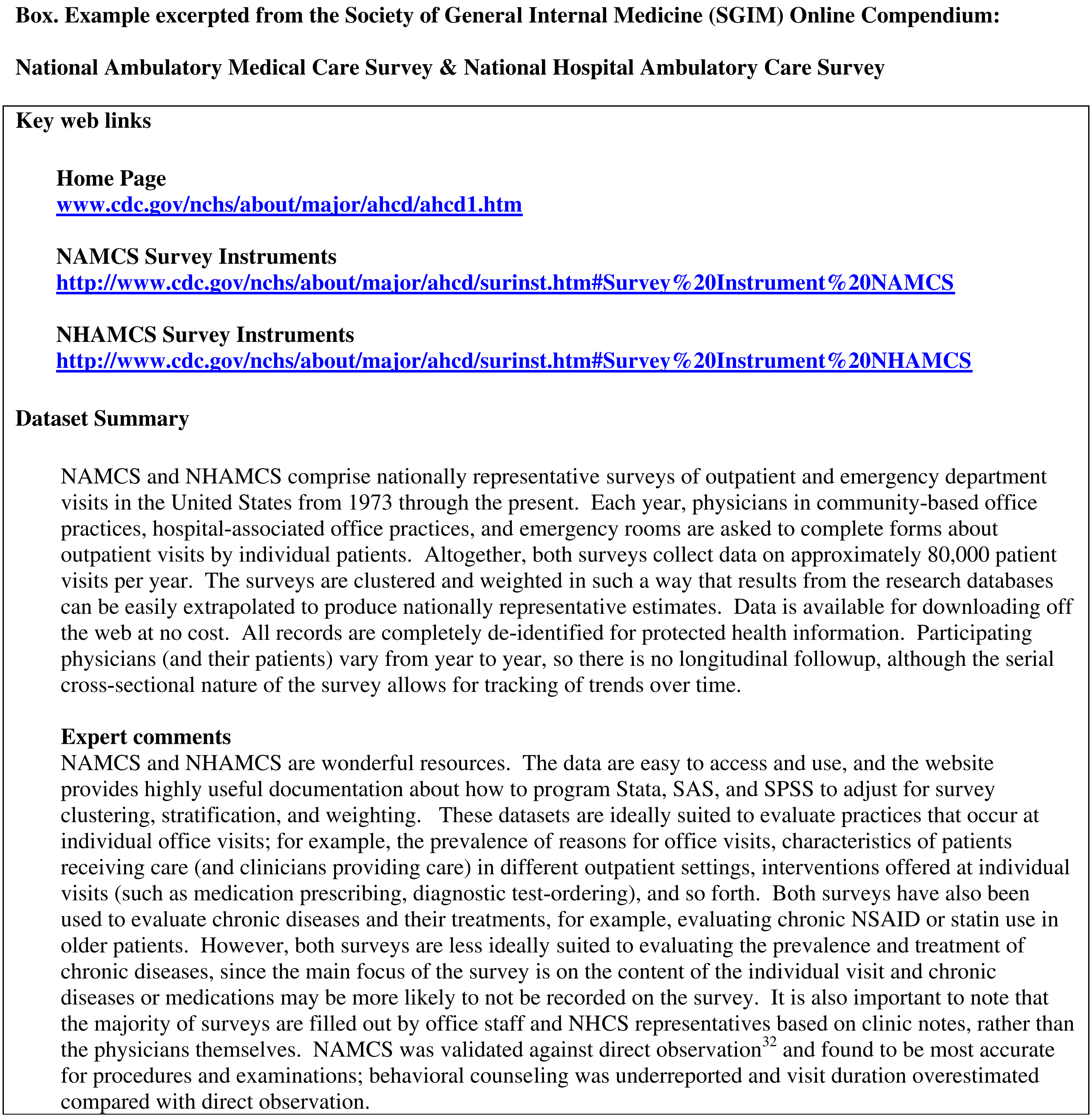
the variables of interest (including predictors, outcomes, and potential confounders). Once a candidate dataset has been identified, we recommend being flexible and adapting the research question to the strengths and limitations of the dataset, as long as the question remains interesting and specific and the methods to answer it are scientifically sound. Be creative. Some measures of interest may not have been ascertained directly, but data may be available to construct a suitable proxy. In some cases, you may find a dataset that initially looked promising lacks the necessary data (or data quality) to answer research questions in your area of interest reliably. In that case, you should be prepared to search for an alternative dataset.

A specific research question is essential to good research. However, many researchers have a general area of interest but find it difficult to identify specific research questions without knowing the specific data available. In that case, combing research documentation for unexamined yet interesting measures in your area of interest can be fruitful. Beginning with the dataset and no focused area of interest may lead to data dredging—simply creating cross tabulations of unexplored variables in search of significant associations is bad science. Yet, in our experience, many good studies have resulted from a researcher with a general topic area of interest finding a clinically meaningful yet underutilized measure and having the insight to frame a research question that uses that measure to answer a novel and clinically compelling question (see references for examples).[4–8] Dr. Warren Browner once exhorted, “just because you were not smart enough to think of a research question in advance doesn’t mean it’s not important!” [quote used with permission].

Select a Dataset

Case Continued. After a review of available datasets that fit his topic area of interest, the fellow decides to use data from the Surveillance Epidemiology and End Results Program linked to Medicare claims (SEER-Medicare).

The range and intricacy of large datasets can be daunting to a junior researcher. Fortunately, several online compendia are available to guide researchers (Table 3), including one recently developed by this manuscript’s authors for the Society of General Internal Medicine (SGIM) (www.sgim.org/go/datasets). The SGIM Research Dataset Compendium was developed and is maintained by members of the SGIM research committee. SGIM Compendium developers consulted with experts to identify and profile high-value datasets for generalist researchers. The Compendium includes a description of and links to over 40 high-value datasets used for health services, clinical epidemiology, and medical education research. The SGIM Compendium provides detailed information of use in selecting a dataset, including sample sizes and characteristics, available measures and how data was measured, comments from expert users, links to the dataset, and example publications (see Box for example). A selection of datasets from this Compendium is listed in Table 4. SGIM members can request a one-time telephone consultation with an expert user of a large dataset (see details on the Compendium website).



Box. Example excerpted from the Society of General Internal Medicine (SGIM) Online Compendium:

National Ambulatory Medical Care Survey & National Hospital Ambulatory Care Survey

Key web links

Home Page
www.cdc.gov/nchs/about/major/ahcd/ahcd1.htm

NAMCS Survey Instruments
http://www.cdc.gov/nchs/about/major/ahcd/surinst.htm#Survey%20Instrument%20NAMCS

NHAMCS Survey Instruments
http://www.cdc.gov/nchs/about/major/ahcd/surinst.htm#Survey%20Instrument%20NHAMCS

Dataset Summary

NAMCS and NHAMCS comprise nationally representative surveys of outpatient and emergency department visits in the United States from 1973 through the present. Each year, physicians in community-based office practices, hospital-associated office practices, and emergency rooms are asked to complete forms about outpatient visits by individual patients. Altogether, both surveys collect data on approximately 80,000 patient visits per year. The surveys are clustered and weighted in such a way that results from the research databases can be easily extrapolated to produce nationally representative estimates. Data is available for downloading off the web at no cost. All records are completely de-identified for protected health information. Participating physicians (and their patients) vary from year to year, so there is no longitudinal followup, although the serial cross-sectional nature of the survey allows for tracking of trends over time.

Expert comments

NAMCS and NHAMCS are wonderful resources. The data are easy to access and use, and the website provides highly useful documentation about how to program Stata, SAS, and SPSS to adjust for survey clustering, stratification, and weighting. These datasets are ideally suited to evaluate practices that occur at individual office visits; for example, the prevalence of reasons for office visits, characteristics of patients receiving care (and clinicians providing care) in different outpatient settings, interventions offered at individual visits (such as medication prescribing, diagnostic test-ordering), and so forth. Both surveys have also been used to evaluate chronic diseases and their treatments, for example, evaluating chronic NSAID or statin use in older patients. However, both surveys are less ideally suited to evaluating the prevalence and treatment of chronic diseases, since the main focus of the survey is on the content of the individual visit and chronic diseases or medications may be more likely to not be recorded on the survey. It is also important to note that the majority of surveys are filled out by office staff and NHCS representatives based on clinic notes, rather than the physicians themselves. NAMCS was validated against direct observation[32] and found to be most accurate for procedures and examinations; behavioral counseling was underreported and visit duration overestimated compared with direct observation.

Dataset complexity, cost, and time to acquire the data and obtain institutional review board (IRB) approval are critical considerations for junior researchers, who are new to secondary analysis, have few financial resources, and limited time to demonstrate productivity. Table 4 illustrates the complexity and cost of large datasets across a range of high value datasets used by generalist researchers. Dataset complexity increases by number of subjects, file structure (e.g., single versus multiple records per individual), and complexity of the survey design. Many publicly available datasets are free, and others can cost tens of thousands of dollars to obtain. Time to acquire the datasets and obtain IRB board approval vary. Some datasets can be downloaded from the web, others require multiple layers of permission and security, and in some cases data must be analyzed in a central data processing center. If the project requires linking new data to an existing database, this linkage will add to the time needed to complete the project and probably require enhanced data security. One advantage

Table 3. Online Compendia of Secondary Datasets

Compendium	Web address	Description
Society of General Internal Medicine (SGIM) Research Dataset Compendium	www.sgim.org/go/datasets	Designed to assist investigators conducting research on existing datasets, with a particular emphasis on health services research, clinical epidemiology, and research on medical education. Includes information on strengths and weaknesses of datasets and the insights of experienced users about making best use of the data
National Information Center on Health Services Research and Health Care Technology (NICHSR)	http://www.nlm.nih.gov/nichsr/index.html	This group of sites provides links to a wide variety of data tools and statistics, including research datasets, data repositories, health statistics, survey instruments, and more. It is sponsored by the National Library of Medicine
Inter-University Consortium for Political and Social Research (ICPSR)	https://www.icpsr.umich.edu	World's largest archive of digital social science data, including many datasets with extensive information on health and health care. ICPSR includes many sub-archives on specific topic areas, including minority health, international data, substance abuse, and mental health, and more
Partners in Information Access for the Public Health Workforce	http://phpartners.org/health_stats.html	Provides links to a variety of national, state, and local health and public health datasets. Also provides links to sites providing a wide variety of health statistics, information on health information technology and standards, and other resources. Sponsored by a collaboration of US government agencies, public health organizations, and health sciences libraries
Canadian Research Data Centres	http://www.statcan.gc.ca/rdc-cdr/index-eng.htm	Links to datasets available for analysis through Canada's Research Data Centres (RDC) program
Directory of Health and Human Services Data Resources (US Dept. of Health and Human Services)	http://aspe.hhs.gov/datacncl/DataDir/index.shtml	This site provides brief information and links to almost all datasets from National Institutes of Health (NIH), Centers for Disease Control and Prevention (CDC), Centers for Medicare and Medicaid Services (CMS), Agency for Healthcare Research and Quality (AHRQ), Food and Drug Administration (FDA), and other agencies of the US Department of Health and Human Services
National Center for Health Statistics (NCHS)	http://www.cdc.gov/nchs/index.htm	This site links to a variety of datasets from the National Center for Health Statistics, several of which are profiled in Table 4. These datasets are available for downloading at no cost
Medicare Research Data Assistance Center (RESDAC); and Centers for Medicare and Medicaid Services (CMS) Research, Statistics, Data & Systems	www.resdac.umn.edu/Available_CMS_Data.asp	These sites link to a variety of datasets from the Centers for Medicare and Medicaid Services (CMS)
Veterans Affairs (VA) data	www.virec.research.va.gov/index.htm	A series of datasets using administrative and computerized clinical data to describe care provided in the VA health care system, including information on outpatient visits, pharmacy data, inpatient data, cost data, and more. With some exceptions, use is generally restricted to researchers with VA affiliations (this can include a co-investigator with a VA affiliation)

of most secondary studies using publicly available datasets is the rapid time to IRB approval. Many publicly available large datasets contain de-identified data and are therefore eligible for expedited review or exempt status. If you can download the dataset from the web, it is probably exempt, but your local IRB must make this determination.

Linking datasets can be a powerful method for examining an issue by providing multiple perspectives of patient experience. Many datasets, including SEER, for example, can be linked to the Area Resource File to examine regional variation in practice patterns. However, linking datasets together increases the complexity and cost of data management. A new researcher might consider first conducting a study only on the initial database, and then conducting their next study using the linked database. For some new investigators, this approach can progressively advance

Table 4. Examples of High Value Datasets

Cost, availability, and complexity	Dataset	Description	Sample publications
Free. Readily available. Population-based survey with cross-sectional design. Does not require special statistical techniques to address complex sampling	Surveillance, Epidemiology and End Results Program (SEER) http://www.seer.cancer.gov/	Population-based multi-regional cancer registry database. SEER data are updated annually. Can be linked to Medicare claims and files (see Medicare below)	Trends in breast-conserving surgery among Asian Americans and Pacific Islanders, 1992–2000[12] Treatment and outcomes of gastric cancer among US-born and foreign-born Asians and Pacific Islanders[13]
Free. Readily available. Requires statistical considerations to account for complex sampling design and use of survey weights	National Ambulatory Medical Care Survey (NAMCS) & National Hospital Ambulatory Care Survey (NHAMCS) http://www.cdc.gov/nchs/ahcd.htm	Nationally-representative serial cross-sectional surveys of outpatient and emergency department visits. Can combine survey years to increase sample sizes (e.g., for uncommon conditions) or evaluate temporal trends. Provides national estimates The NAMCS and NHAMCS are conducted annually. Do not link to other datasets	Preventive health examinations and preventive gynecological examinations in the US[14] Primary care physician office visits for depression by older Americans[15]
	National Health Interview Survey (NHIS) http://www.cdc.gov/nchs/nhis.htm	Nationally-representative serial cross-sectional survey of individuals and families including information on health status, injuries, health insurance, access and utilization information. The NHIS is conducted annually. Can combine survey years to look at rare conditions Can be linked to National Center for Health Statistics Mortality Data; Medicare enrollment and claims data; Social Security Benefit History Data; Medical Expenditure Panel Survey (MEPS) data; and National Immunization Provider Records Check Survey (NIPRCS) data from 1997–1999	Psychological distress in long-term survivors of adult-onset cancer: results from a national survey[16] Diabetes and Cardiovascular Disease among Asian Indians in the US[17]
	Behavioral Risk Factor Surveillance System (BRFSS) http://www.cdc.gov/brfss/	Serial cross-sectional nationally-representative survey of health risk behaviors, preventative health practices, and health care access. Provides national and state estimates. Since 2002, the Selected Metropolitan/Micropolitan Area Risk Trends (SMART) project has also used BRFSS data to identify trends in selected metropolitan and micropolitan statistical areas (MMSAs) with 500 or more respondents. BRFSS data are collected monthly. Does not link to other datasets	Perceived discrimination in health care and use of preventive health services[18] Use of recommended ambulatory care services: is the Veterans Affairs quality gap narrowing?[19]
Free or minimal cost. Readily available. Can do more complex studies by combining data from multiple waves and/or records. Accounting for complex sampling design and use of survey weights can be more complex when using multiple waves—seek support from a statistician. Or can restrict sample to single waves for ease of use	Nationwide Inpatient Sample (NIS) http://www.hcup-us.ahrq.gov/databases.jsp	The largest US database of inpatient hospital stays that incorporates data from all payers, containing data from approximately 20% of US community hospitals. Sampling frame includes approximately 90% of discharges from US hospitals NIS data is collected annually. For most states, the NIS includes hospital identifiers that permit linkages to the American Hospital Association (AHA) Annual Survey Database and county identifiers that permit linkages to the Area Resource File (ARF)	Factors associated with patients who leave acute-care hospitals against medical advice[20] Impact of hospital volume on racial disparities in cardiovascular procedure mortality[21]
	National Health and Nutrition Examination Survey (NHANES) http://www.cdc.gov/nchs/nhanes.htm	Nationally- representative series of studies combining data from interviews, physical examination, and laboratory tests NHANES data are collected annually. Can be linked to National Death Index (NDI) mortality data; Medicare enrollment and claims data; Social Security Benefit History Data; and Medical Expenditure Panel Survey (MEPS) data; and Dual Energy X-Ray Absorptiometry (DXA) Multiple Imputation Data Files from 1999–2004	Demographic differences and trends of vitamin D insufficiency in the US population,1988-2004[22] Association of hypertension, diabetes, dyslipidemia, and metabolic syndrome with obesity: findings from the National Health and Nutrition Examination Survey, 1999 to 2004[23]
	The Health and Retirement Study (HRS) http://hrsonline.isr.umich.edu/index.php	A nationally-representative longitudinal survey of adults older than 50 designed to assess health status, employment decisions, and economic security during retirement HRS data is collected every 2 years. Can be linked to Social Security Administration data; Internal Revenue Service data; Medicare claims data (see Medicare below); and Minimum Data Set (MDS) data	Chronic conditions and mortality among the oldest old[24] Advance directed and surrogate decision making before death[25]
	Medical Expenditure Panel Survey (MEPS) http://www.meps.ahrq.gov/mepsweb/	Serial nationally-representative panel survey of individuals, families, health care providers, and employers covering a variety of topics. MEPS data are collected annually Can be linked by request to the Agency for Healthcare Research and Quality to numerous datasets including the NHIS, Medicare data, and Social Security data	Loss of health insurance among non-elderly adults in Medicaid[26] Influence of patient-provider communication on colorectal cancer screening[27]

(continued on next page)

Table 4. (continued)

Cost, availability, and complexity	Dataset	Description	Sample publications
Data costs are in the thousands to tens of thousands of dollars. Requires an extensive application and time to acquire data is on the order of months at a minimum. Databases frequently have observations on the order of 100,000 to >1,000,000. Require additional statistical considerations to account for complex sampling design, use of survey weights, or longitudinal analysis. Multiple records per individual. Complex database structure requires a higher degree of analytic and programming skill to create a study dataset efficiently.	Medicare claims data (alone), SEER-Medicare, and HRS-Medicare http://www.resdac.org/Medicare/data_available.asp	Claims data on Medicare beneficiaries including demographics and resource utilization in a wide variety of inpatient and outpatient settings. Medicare claims data are collected continually and made available annually. Can be linked to other Medicare datasets that use the same unique identifier numbers for patients, providers, and institutions, for example, the Medicare Current Beneficiary Survey, the Long-Term Care Minimum Data Set, the American Hospital Association Annual Survey, and others. SEER and the HRS offer linkages to Medicare data as well (as described above)	Long-term outcomes and costs of ventricular assist devices among Medicare beneficiaries[28] Association between the Medicare Modernization Act of 2003 and patient wait times and travel distance for chemotherapy[29]
	Medicare Current Beneficiary Survey (MCBS) http://www.cms.gov/MCBS/	Panel survey of a nationally-representative sample of Medicare beneficiaries including health status, health care use, health insurance, socioeconomic and demographic characteristics, and health expenditures. MCBS data are collected annually. Can be linked to other Medicare Data	Cost-related medication nonadherence and spending on basic needs following implementation of Medicare Part D[30] Medicare beneficiaries and free prescription drug samples: a national survey[31]

programming skills and build confidence while demonstrating productivity.

Get to Know your Dataset

Case Continued. The fellow's primary mentor encourages him to closely examine the accuracy of the primary predictor for his study—race and ethnicity—as reported in SEER-Medicare. The fellow has a breakthrough when he finds an entire issue of the journal Medical Care dedicated to SEER-Medicare, including a whole chapter on the accuracy of coding of sociodemographic factors.[9]

In an analysis of primary data you select the patients to be studied and choose the study measures. This process gives you a close familiarity with study subjects, and how and what data were collected, that is invaluable in assessing the validity of their measures, the potential bias in measuring associations between predictors and outcome variables (internal validity), and the generalizability of their findings to target populations (external validity). The importance of this familiarity with the strengths and weaknesses of the dataset cannot be overemphasized. Secondary data research requires considerable effort to obtain the same level of familiarity with the data. Therefore, knowing your data in detail is critical. Practically, this objective requires scouring online documentation and technical survey manuals, searching PubMed for validation studies, and closely reading previous studies using your dataset, to answer the following types of questions: Who collected the data, and for what purpose? How did subjects get into your dataset? How were they followed? Do your measures capture what you think they capture?

We strongly recommend taking advantage of help offered by the dataset managers, typically described on the dataset's website. For example, the Research Data Assistance Center (ResDAC) is a dedicated resource for researchers using data from the Centers for Medicare and Medicaid Services (CMS).

Assessing the validity of your measures is one of the central challenges of large dataset research. For large survey datasets, a good first step in assessing the validity of your measures is to read the questions as they were asked in the survey. Some questions simply have face validity. Others, unfortunately, were collected in a way that makes the measure meaningless, problematic, or open to a range of interpretations. These ambiguities can occur in how the question was asked or in how the data were recorded into response categories.

Another essential step is to search the online documentation and published literature for previous validation studies. A PubMed search using the dataset name or measure name/type and the publication type "validation studies" is a good starting point. The key question for a validity study relates to how and why the question was asked and data were collected (e.g., self-report, chart abstraction, physical measurements, billing claims) in relationship to a gold standard. For example, if you are using claims data you should recognize that the primary purpose of those data was not for research, but for reimbursement. Consequently, claims data are limited by the scope of services that are reimbursable and the accuracy of coding by clinicians completing encounter forms for billing or by coders in the claims departments of hospitals and clinics. Some clinical measures can be assessed by asking subjects if they have the condition of interest, such as self reported diagnosis of hypertension. Self-reported data may be adequate for some research questions (e.g., does a diagnosis of hypertension lead people to exercise more?), but inadequate for others (e.g., the prevalence of hypertension among people with diabetes). Even measured data, such as blood pressure, have limitations in that methods of measurement for a study may differ from methods used to diagnose a disorder in the clinician's office. In the National Health and Nutrition Examination Survey, for example, subject's blood pressure is based on the average of several measures in a single visit. This differs from the standard clinical practice of measuring blood pressure at separate office visits before diagnosing hypertension. Rarely do available measures capture exactly what you are trying to study. In our experience measures in existing datasets are often good enough to answer the research question, with proper interpretation to account for what the measures actually assesses and how they differ from the underlying constructs.

Finally, we suggest paying close attention to the completeness of measures, and evaluating whether missing data are random or non-random (the latter might result in bias, whereas the former is generally acceptable). Statistical approaches to missing data are beyond the scope of this paper, and most statisticians can help you address this problem appropriately. However, pay close attention to "skip patterns"; some data are missing simply because the survey item is only asked of a subset for which it applies. For example, in the Health and Retirement Study the question about need for assistance with toileting is only asked of subjects who respond that they have difficulty using the toilet. If you were unaware of this skip pattern and attempted to study assistance with toileting, you would be distressed to find over three-quarters of respondents had missing responses for this question (because they reported no difficulty using the toilet).

Fellows and other trainees usually do their own computer programming. Although this may be daunting, we encourage this practice so fellows can get a close feel for the data and become more skilled in statistical analysis. Datasets, however, range in complexity (Table 4). In our experience, fellows who have completed introductory training in SAS, STATA, SPSS, or other similar statistical software have been highly successful analyzing datasets of moderate complexity without the on-going assistance of a statistical programmer. However, if you do have a programmer who will do much of the coding, be closely involved and review all data cleaning and statistical output as if you had programmed it yourself. Close attention can reveal all sorts of patterns, problems, and opportunities with the data that are obscured by focusing only on the final outputs prepared by a statistical programmer. Programmers and statisticians are not clinicians; they will often not recognize when the values of variables or patterns of missingness don't make sense. If estimates seem implausible or do not match previously published estimates, then the analytic plan, statistical code, and measures should be carefully rechecked.

Keep in mind that "the perfect may be the enemy of the good." No one expects perfect measures (this is also true for primary data collection). The closer you are to the data, the more you see the warts—don't be discouraged by this. The measures need to pass the sniff test, in other words have clinical validity based primarily on judgement that they make sense clinically or scientifically, but also supported where possible by validation procedures, reference to auditing procedures, or in other studies that have independently validated the measures of interest.

Structure your Analysis and Presentation of Findings in a Way that Is Clinically Meaningful

Case continued. The fellow finds that Blacks are less likely to receive chemotherapy in the last 2 weeks of life (Blacks 4%, Whites 6%, $p < 0.001$). He debates the meaning of this statistically significant 2% absolute difference.

Often, the main challenge for investigators who are new to secondary data analysis is carefully structuring the analysis and presentation of findings in a way that tells a meaningful story. Based on what you've found, what is the story that you want your target audience to understand? When appropriate, it can be useful to conduct carefully planned sensitivity analysis to evaluate the robustness of your primary findings. A sensitivity analysis assesses the effect of variation in assumptions on the outcome of interest. For example, if 10% of subjects did not answer a "yes" or "no" question, you could conduct sensitivity analyses to estimate the effects of excluding missing responses, or categorizing them as all "yes" or all "no." Because large datasets may contain multiple measures of interests, co-variates, and outcomes, a frequent temptation is to present huge tables with multiple rows and columns. This is a mistake. These tables can be challenging to sort through, and the clinical importance of the story resulting from the analysis can be lost. In our experience, a thoughtful figure often captures the take-home message in a way that is more interpretable and memorable to readers than rows of data tables.

You should keep careful track of subjects you decide to exclude from the analysis and why. Editors, reviewers, and readers will want to know this information. The best way to keep track is to construct a flow diagram from the original denominator to the final sample.

Don't confuse statistical significance with clinical importance in large datasets. Due to large sample sizes, associations may be statistically significant but not clinically meaningful. Be mindful of what is meaningful from a clinical or policy perspective. One concern that frequently arises at this stage in large database research is the acceptability of "exploratory" analyses, or the practice of examining associations between multiple factors of interest. On the one hand, exploratory analyses risk finding a significant association by chance alone from testing multiple associations (a false-positive result). On the other hand, the critical issue is not a statistical one, but rather whether the issue is important.[10] Exploratory analyses are acceptable if done in a thoughtful way that serves an a priori hypothesis, but not if merely data dredging looking for associations.

We recommend consulting with a statistician when using data from a complex survey design (see Table 2) or developing a conceptually advanced study design, for example, using longitudinal data, multilevel modeling with clustered data, or surivival analysis. The value of input (even if informal) from a statistician or other advisor with substantial methodological expertise cannot be overstated.

CONCLUSIONS

Case Conclusion. Two years after he began the project the fellow completes the analysis and publishes the paper in a peer-reviewed journal.[11]

A 2-year timeline from inception to publication is typical for large database research. Academic potential is commonly assessed by the ability to see a study through to publication in a peer-reviewed journal. This timeline allows a fellow who began a secondary analysis at the start of a 2-year training program to search for a job with an article under review or in press.

In conclusion, secondary dataset research has tremendous advantages, including the ability to assess outcomes that would be difficult or impossible to study using primary data collection, such as those involving exceptionally long follow-up times or rare outcomes. For junior investigators, the potential for a shorter time to publication may help secure a job or career development funding. Some of the time "saved" by not collecting data yourself, however, needs to be "spent" becoming familiar with the dataset in intimate detail. Ultimately, the same factors that apply to successful primary data analysis apply to secondary data analysis, including the development of a clear research question, study sample, appropriate measures, and a thoughtful analytic approach.

Acknowledgments:
Contributors: *The authors would like to thank Sei Lee, MD, Mark Freidberg, MD, MPP, and J. Michael McWilliams, MD, PhD, for their input on portions of this manuscript.*

Grant Support: *Dr. Smith is supported by a Research Supplement to Promote Diversity in Health Related Research from the National Institute on Aging (R01AG028481), the National Center for Research Resources UCSF-CTSI (UL1 RR024131), and the National Palliative Care Research Center. Dr. Steinman is supported by the National Institute on Aging and the American Federation for Aging Research (K23 AG030999). An unrestricted grant from the Society of General Internal Medicine (SGIM) supported development of the SGIM Research Dataset Compendium.*

Prior Presentations: *An earlier version of this work was presented as a workshop at the Annual Meeting of the Society of General Internal Medicine in Minneapolis, MN, April 2010.*

Conflict of Interest: *None disclosed.*

Corresponding Author: *Alexander K. Smith, MD, MPH; Division of Geriatrics, Department of Medicine, University of California, San Francisco, 4150 Clement St (181G) 94121, San Francisco, CA, USA (e-mail: aksmith@ucsf.edu).*

REFERENCES

1. **Mainous AG 3rd, Hueston WJ.** Using other people's data: the ins and outs of secondary data analysis. Fam Med. 1997;29(8):568–571.
2. **Doolan DM, Froelicher ES.** Using an existing data set to answer new research questions: a methodological review. Res Theory Nurs Pract. 2009;23(3):203–215.
3. **Shlipak M, Stehman-Breen C.** Observational research databases in renal disease. J Am Soc Nephrol. 2005;16(12):3477–3484.
4. **Williams BA, Lindquist K, Moody-Ayers SY, Walter LC, Covinsky KE.** Functional impairment, race, and family expectations of death. J Am Geriatr Soc. 2006;54(11):1682–1687.
5. **Steinman MA, Sands LP, Covinsky KE.** Self-restriction of medications due to cost in seniors without prescription coverage. J Gen Intern Med. 2001;16(12):793–799.
6. **Lindenberger EC, Landefeld CS, Sands LP, et al.** Unsteadiness reported by older hospitalized patients predicts functional decline. J Am Geriatr Soc. 2003;51(5):621–626.
7. **Linder JA, Ma J, Bates DW, Middleton B, Stafford RS.** Electronic health record use and the quality of ambulatory care in the United States. Arch Intern Med. 2007;167(13):1400–1405.
8. **Lee SJ, Steinman MA.** Tan EJ. Driving Status and Mortality in US Retirees: Volunteering; 2010.
9. **Bach PB, Guadagnoli E, Schrag D, Schussler N, Warren JL.** Patient demographic and socioeconomic characteristics in the SEER-Medicare database applications and limitations. Med Care. 2002;40(8):IV-19–25.
10. **Browner WS, Newman TB.** Are all significant P values created equal? The analogy between diagnostic tests and clinical research. JAMA. 1987;257(18):2459–2463.
11. **Smith AK, Earle CC, McCarthy EP.** Racial and Ethnic Differences in End-of-Life Care in Fee-for-Service Medicare Beneficiaries with Advanced Cancer. J Am Geriatr Soc. Nov 21 2008.
12. **Goel MS, Burns RB, Phillips RS, Davis RB, Ngo-Metzger Q, McCarthy EP.** Trends in breast conserving surgery among Asian Americans and Pacific Islanders, 1992-2000. J Gen Intern Med. 2005;20(7):604–611.
13. **Byfield SA, Earle CC, Ayanian JZ, McCarthy EP.** Treatment and outcomes of gastric cancer among United States-born and foreign-born Asians and Pacific Islanders. Cancer. 2009;115(19):4595–4605.
14. **Mehrotra A, Zaslavsky AM, Ayanian JZ.** Preventive health examinations and preventive gynecological examinations in the United States. Arch Intern Med. 2007;167(17):1876–1883.
15. **Harman JS, Veazie PJ, Lyness JM.** Primary care physician office visits for depression by older Americans. J Gen Intern Med. 2006;21(9):926–930.
16. **Hoffman KE, McCarthy EP, Recklitis CJ, Ng AK.** Psychological distress in long-term survivors of adult-onset cancer: results from a national survey. Arch Intern Med. 2009;169(14):1274–1281.
17. **Mohanty SA, Woolhandler S, Himmelstein DU, Bor DH.** Diabetes and cardiovascular disease among Asian Indians in the United States. J Gen Intern Med. 2005;20(5):474–478.
18. **Hausmann LR, Jeong K, Bost JE, Ibrahim SA.** Perceived discrimination in health care and use of preventive health services. J Gen Intern Med. 2008;23(10):1679–1684.
19. **Ross JS, Keyhani S, Keenan PS, et al.** Use of recommended ambulatory care services: is the Veterans Affairs quality gap narrowing? Arch Intern Med. 2008;168(9):950–958.
20. **Ibrahim SA, Kwoh CK, Krishnan E.** Factors associated with patients who leave acute-care hospitals against medical advice. Am J Public Health. 2007;97(12):2204–2208.
21. **Trivedi AN, Sequist TD, Ayanian JZ.** Impact of hospital volume on racial disparities in cardiovascular procedure mortality. J Am Coll Cardiol. 2006;47(2):417–424.
22. **Ginde AA, Liu MC, Jr Camargo CA.** Demographic differences and trends of vitamin D insufficiency in the US population, 1988-2004. Arch Intern Med. 2009;169(6):626–632.
23. **Nguyen NT, Magno CP, Lane KT, Hinojosa MW, Lane JS.** Association of hypertension, diabetes, dyslipidemia, and metabolic syndrome with obesity: findings from the National Health and Nutrition Examination Survey, 1999 to 2004. J Am Coll Surg. 2008;207(6):928–934.
24. **Lee SJ, Go AS, Lindquist K, Bertenthal D, Covinsky KE.** Chronic conditions and mortality among the oldest old. Am J Public Health. 2008;98(7):1209–1214.
25. **Silveira MJ, Kim SY, Langa KM.** Advance directives and outcomes of surrogate decision making before death. N Engl J Med. Apr 1;362 (13):1211-1218.
26. **Sommers BD.** Loss of health insurance among non-elderly adults in Medicaid. J Gen Intern Med. 2009;24(1):1–7.
27. **Carcaise-Edinboro P, Bradley CJ.** Influence of patient-provider communication on colorectal cancer screening. Med Care. 2008;46(7):738–745.

28. **Hernandez AF, Shea AM, Milano CA, et al.** Long-term outcomes and costs of ventricular assist devices among Medicare beneficiaries. JAMA. 2008;300(20):2398–2406.
29. **Shea AM, Curtis LH, Hammill BG, DiMartino LD, Abernethy AP, Schulman KA.** Association between the Medicare Modernization Act of 2003 and patient wait times and travel distance for chemotherapy. JAMA. 2008;300(2):189–196.
30. **Madden JM, Graves AJ, Zhang F, et al.** Cost-related medication nonadherence and spending on basic needs following implementation of Medicare Part D. JAMA. 2008;299(16):1922–1928.
31. **Tjia J, Briesacher BA, Soumerai SB, et al.** Medicare beneficiaries and free prescription drug samples: a national survey. J Gen Intern Med. 2008;23(6):709–714.
32. **Gilchrist VJ, Stange KC, Flocke SA, McCord G, Bourguet CC.** A comparison of the National Ambulatory Medical Care Survey (NAMCS) measurement approach with direct observation of outpatient visits. Med Care. 2004;42(3):276–280.