share
State Health Access Reform Evaluation

Robert Wood Johnson Foundation

shadac
State Health Access Data Assistance Center

September 2011

Strategies for Leveraging SNACC Data for Policy and Evaluation: Barriers and Challenges to Linked Data Sets

-Second in the Series, "The Next Generation of Data Linkage Projects"

Michael J. O'Grady, Ayesha Mahmud

NORC • University of Chicago • Chicago, IL

Introduction

The Importance of Data Linkages

The demand for data for the purposes of policy analysis has increased dramatically, with health services researchers and policymakers alike poised to analyze and evaluate a range of challenges to the health care system. However, health services researchers and policymakers often face the challenge of having incomplete data. Health-related data, whether from surveys, claims, or administrative records, are often created and held by different public and private entities. To address this disconnect, individual data sets can be linked to one another, providing a more comprehensive overarching data set while avoiding the cost of duplicate data collection.

The process of linking data sets can take a number of different forms, but the common characteristic of the process is the application of statistical methods in order to identify and connect the same—or demographically similar—individuals within each of the data sets. For example, a researcher might match survey respondents with their actual claims and eligibility files in order to study the correlation between income level and service utilization.

Background

SNACC

SNACC—alternately referred to as the "Medicaid Undercount Project"—began as a collaborative effort to explain why discrepancies exist between survey estimates of Medicaid enrollment and the enrollment numbers reported in state and national administrative data. With the financial support of the Robert Wood Johnson Foundation, six organizations (see text box below) joined forces to conduct six data linkage projects to determine which data source provided the most robust estimates of the Medicaid population. These projects are listed in Table 1. In addition to the linkages established under SNACC, several other useful data linkage projects have been conducted. These non-SNACC projects are listed in Table 2.

SNACC Organizations

State Health Access Data Assistance Center (SHADAC)

National Center for Health Statistics (NCHS)

Agency for Healthcare Research & Quality (AHRQ)

Assistant Secretary for Planning & Evaluation (ASPE)

Centers for Medicare & Medicaid Services (CMS)

Census Bureau

SNACC has always had a focus on informing policy: The larger aim of identifying the most robust Medicaid estimates was to provide policy makers with more accurate approximations of the Medicaid and uninsured populations in order to facilitate the creation of effective policy. The project then evolved into a [powerful] set of analytic files uniquely positioned to inform policy development, implementation, and evaluation.

DATA LINKAGES UNDER THE ACA

Challenges Surrounding Linked Data Sets

With the Affordable Care Act (ACA) in place, both the federal and state governments need robust data in order to effectively meet the challenges of designing new programs, implementing those designs, and evaluating the outcomes. This brief is the second in a three-part series, "The Next Generation of Data Linkage Projects." The first brief explored priority areas for data linkages, looking at health benefit exchanges, the intersection of Medicare and Medicaid, and Medicare payment reform. The present brief will analyze three challenges to creating linked data sets: (1) methodological challenges; (2) privacy concerns; and (3) barriers to data access. The ideas outlined here were identified during interviews with key stakeholders both inside and outside of government.

Statistical Matching: An Example

The American Community Survey (ACS) has approximately three million records and is a preferred tool for calculating small area estimates. Suppose a state wants to estimate its Medicaid population, and the state's Medicaid claims file has 10 variables in common with the ACS. Using these common variables, the state can estimate (typically using regression techniques) ACS values that it would like to match onto its Medicaid claims data—e.g., educational attainment. If the state can accurately predict educational attainment using the 10 common variables, it can then use the coefficients developed from the ACS-based model to predict and project educational attainment values onto its Medicaid claims file.

Table 1:
The SNACC Project: Data Linkages to Date

The national-level CPS database
The Medicaid Statistical Information Statistic (MSIS) and the Current Population Survey (CPS)
The state-frame, household, and person MSIS data to the CPS
The MSIS and the National Health Interview Survey (NHIS)
The MSIS and the CPS annual Social and Economic Supplement (ASEC), 2003-2004
The MSIS and the Medical Expenditure Panel Survey (MEPS)

Table 2:
Other Data Linkage Initiatives

Health insurance data from the 2001 State and Local Area Integrated Telephone Survey's National Survey of Children with Special Health Care Needs (NS-CSHCN) linked to immunization status data from the 2000-2002 National Immunization Survey (NIS)
Cancer registry data from the Surveillance, Epidemiology, and End Results (SEER) linked to Medicare managed care enrollee survey data from the Medicare Health Outcomes Survey (MHOS)
Tract-level poverty data linked to vital records for infants born to American Indian women between 1990 and 1999
Administrative data from the Medicaid/State Children's Health Insurance Program linked to birth and death records
Air pollution data from the National Health Interview Survey (NHIS) linked to data from the National Hospital Discharge Survey (NHDS)
Survey data from the National Center for Health Statistics (NCHS) linked to death certificate records from the National Death Index
Survey data from NCHS linked to claims data from CMS
Survey data from NCHS linked to benefit records from the Old Age, Survivors and Disability Insurance (OASDI) and Supplemental Security Income (SSI)

Methodological Challenges to Data Linkage

Researchers face many methodological challenges when performing data linkage. One of these challenges is posed by the characteristics of the data sets that are available, which determines the appropriate linkage strategy. The first and more common type of data linkage strategy uses a unique identifier (when available) to find and select the same individual within both a survey and an administrative database, linking together the individual's data from both files in a new "linked" data set.[1] This strategy is considered the most rigorous form of matching because the linked data comes from the same individual in both databases.

The second form of data matching, sometimes referred to as "statistical" matching, takes two data sets with a series of variables in common and matches the common variables to predict values for variables found on one data set but not the other (see example in text box on right). The first of these two methods is by far the more common and preferred, given that the data from both files is for the same person. However, statistical matching is sometimes the only option available, since it can be difficult to find common unique identifiers, such as Social Security numbers, between data sets.

General Methodological Challenges

An additional methodological challenge facing data linkage is poor data quality. For example, linkage variables might be miscoded, or they might be inconsistently or incorrectly reported across data sources.

Significantly different sample sizes between data sets pose yet another challenge.

Data linkages also offer a new methodological challenge in the way we think about representative data. That is, data linkage efforts call for a re-examination of traditional statistical measures of variability and representativeness in order to address the problem of potential selection bias in claims data. The typical survey file to be linked will be a statistically representative sample with sampling weights carefully calculated and perhaps recalculated to account for practical issues like respondent non-response. Linking this type of file to an administrative file, such as a claims file, will seldom if ever result in a 100 percent match. This is because in almost every insured population, some reasonable percentage of the covered population has no claims during the year. The result is a situation where the more claims a person has, the more likely he or she is to be linked, thus introducing a selection bias towards those who are more likely to use services. In this case, analysts must consider the effect of this bias on the original sampling weights from the survey. Researchers are still exploring how to adjust weights, create standard errors, and provide a new merged data set with known statistical properties when faced with this scenario.

Privacy Concerns and Barriers to Accessing Linked Data

Because the steps that are taken to maintain data privacy almost always present barriers to data access, it is difficult to disentangle the two. Accordingly, privacy and access will be discussed together.

On-Site Data Access

The matching of unique identifiers and other linkage variables creates new personally identifiable data. It is often difficult to de-identify these data and to prevent re-identification in the event that de-identification is carried out. To address privacy concerns, policies have been enacted to prevent easily identifiable records. One such policy is to restrict physical access to linked files by, for example, requiring researchers to travel to a federal Research Data Center (RDC) where special security clearance is required in order to access sensitive data. Such centers are operated by a number of federal agencies, including the Census Bureau, the Agency for Healthcare Quality and Research (AHRQ) and the National Center for Health Statistics (NCHS). Then, after a researcher accesses on-site sensitive data, government officials must review any output from the linked files to ensure that identifiable information does not leave the facility.

While policies restricting physical access to linked files serve to protect data privacy, they in turn create barriers for researchers, who must spend time and money obtaining clearance and travelling to data centers. One step being taken to address these barriers is to expand the number of RDCs in order to facilitate easier access to linked data. However, with a few exceptions, most of these expansion efforts are focused in the Washington, DC, metropolitan area, leaving significant swathes of the country with limited access.

[1] Depending on how many exact matches can be found between the two databases, survey sampling weights may be adjusted to a lesser or greater extent.

Another option for expanding access to linked files is the use of data enclave technology. While not in use by the Census Bureau, other federal agencies—for example, the National Institute of Standards and Technology (NIST)—are using this technology to provide privacy protection in a virtual format. It is not clear at this time whether the same level of privacy protection can be provided virtually as can be provided on-site, so additional testing is being done. An additional issue is that data enclave technology does require that the user have access to enclave-provided hardware in order to gain access.

A second way to bypass the need for on-site data access is a new technique called data perturbation. Data perturbation is a process that manipulates individual-level data in such a way that the individual-level data is changed, but the control totals and other summary measures remain the same. The idea is that researchers can use the perturbed file for modeling and other forms of analysis, but the individual-level data will not reflect any actual individuals. As an additional service, and to reassure researchers that using perturbed data is still accurate, some agencies (e.g., NCHS) will take a researcher's model and run it again with the unperturbed data for them. In this way, the researcher can see if the results change in any way when run against the real data without ever having direct access to the real data.

Regulatory Limitations

When considering the Census Bureau in particular, we encounter another indirect barrier to greater data linkage. The Census Bureau operates under a different privacy statute than the other federal statistical agencies and has access to a wider range of data.[2] Consequently, depending on the files to be linked, the Census Bureau may be the only federal agency authorized to do the linking. This can place a heavy workload on the Census Bureau, which must not only collect the necessary data, but must also carry out much of the data processing, editing, and linkage on Census premises. The different federal statistical agencies with the help and encouragement of the federal Office of Management and Budget (OMB), have been working to streamline the interagency sharing of data through enhanced interagency agreements. Despite the Census Bureau commitment to this work, the potential for a "bottleneck" of projects at Census is quite real. Given that current funding levels for data linkage projects are quite low, expecting Census to carry out the vast majority of the work may be unrealistic.

[2] The chief legal authorities governing surveys are Title 13 U.S.C., Section 8(b), and Title 15, U.S.C., Sections 1525 and 1526. The Census Bureau operates under Title 13, while other statistical agencies typically operate under Title 15, unless an agency has the Census Bureau conduct a survey on its behalf. A survey conducted under Title 15 is not permitted access to Census address lists and is not permitted access to listing information obtained for surveys that operate under Title 13 government regulations. Title 13 surveys are permitted to use census address lists, but the surveys' home organizations normally have very limited access to non-public-use survey microdata.

Conclusion

This report discusses a number of different concerns and barriers surrounding efforts to create linked datasets. The good news is that almost all of these barriers and concerns have solutions. Some will take more time and research, such as the development of methodologies for "new" standard errors and weights. Others will take more money, such as the expansion of research data centers. Finally, others will require policy changes, such as the regulations surrounding federal survey data collection.

The staffs of the federal statistical agencies are currently carrying the vast majority of the burden of creating linked data sets with little political support and therefore little budgetary support. While policymakers are the most significant potential beneficiaries of data linkages, they have not really discovered the vast richness of the data provided through linkages and as a result have not realized that data linkages could answer some of their most vexing questions.

The ACA is moving forward at a rapid pace. The design decisions that have to be made during this implementation phase will make or break the ACA. Data linkages can inform these decisions so that critical design errors are avoided. The challenge facing us now is how to keep moving forward with SNACC's data linkage efforts.

This brief is a companion to the brief titled, "The Next Generation of Data Linkage Projects: Priority Areas for SNACC under the ACA," which is available at http://www.shadac.org/files/shadac/publications/SNACCDataLinkageBrief_1of3.pdf.

ACKNOWLEDGMENTS

The authors would like to thank the following people for taking the time to be interviewed for this study: Richard Kronick, PhD, ASPE; Linda Bilheimer, PhD, NCHS/CDC; Chris Peterson and April Grady MACPAC; William Scanlon, PhD. National Committee of Vital and Health Statistics (NCVHS); Christine Cox, Jon Poisal, David Baugh, Steven Heffler and Kimberly Lochner, CMS; Amy and Bret O'Hara. Census; and Sharon Arnold and Cara Lesser, CCIIO/CMS.

ABOUT SHARE

The State Health Access Reform Evaluation (SHARE) is a Robert Wood Johnson Foundation (RWJF) program that supports rigorous research on health reform issues, specifically as they relate to the state implementation of the Affordable Care Act (ACA). The program operates out of the State Health Access Data Assistance Center (SHADAC), an RWJF-funded research center in the Division of Health Policy and Management, School of Public Health, University of Minnesota. Information is available at www.statereformevaluation.org.

State Health Access Data Assistance Center
2221 University Avenue, Suite 345
Minneapolis, MN 55414
Phone (612) 624-4802