Investing in a Data Quality Research Program for Administrative Data Linked to Survey Data for Policy Research Purposes Is Essential

Michael Davern, Ph.D.
Assistant Professor
Health Policy and Management
University of Minnesota
2221 University Ave. SE Suite 345
Minneapolis, MN 55414
Email: daver004@umn.edu
Phone: 612-625-4835

Marc Roemer, MS
Man of Leisure and Research
Costa Rica
Formerly a Statistician with the US Census Bureau

Wendy Thomas
Director of Data Access
Minnesota Population Center
University of Minnesota
50 Willey Hall
Minneapolis MN 55414
Email: wlt@umn.edu

Abstract

Survey data have many limitations for policy research yet they are widely used because they are the only source for some data we need to predict, evaluate and learn about the impact of potential and actual changes in policy. Public program administrative data that are used to keep track of program enrollees and benefits received are often used in combination with survey data to create new linked data products. The potential for these linked data products for improving policy research are tremendous but many issues need to be dealt with. A drawback of the linked data files is that most of them will never be put into the public domain due to the sensitive nature of linking. In this paper we set out a research agenda for improving linked data files for policy research considering research that needs to be conducted concerning coverage error, non-response error, sampling error, measurement error, editing/imputation, documentation of metadata and production of timely linked data files.

Keywords: Administrative data, survey data, data linkage, linked data quality, policy research, US Census Bureau, public program participation

Acknowledgments:

Preparation of this paper was supported by grant no. 52084 from the Robert Wood Johnson Foundation. This document reports the results of research and analysis undertaken by the U.S. Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This document is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed herein are attributable only to the author(s) and do not represent the views of the U.S. Census Bureau, its program sponsors or data providers. Karen Soderberg at SHADAC provided expert editing assistance to this manuscript. Sally Obenski, Shelly Wilkie Martinez, Ron Prevost, Dean Resnick, Victoria Lynch (all of the U.S. Census Bureau), Dave Baugh (of the Centers for Medicare and Medicaid Services) and Chris Cox (of the National Center for Health Statistics) also provided excellent comments on the ideas expressed in this paper. All remaining problems with the paper are the fault of the authors only.

Introduction

High quality demographic and program participation data is crucial to public policy evaluation, simulation, and law making, and general population survey data have long met this need. However, survey data have several limitations for this type of policy research. Most notably, surveys often undercount participants in many of the most important public welfare programs in the United States relative to administrative enrollment data. Why this "undercount" occurs is a matter of multiple causes that have to deal with both survey and administrative data limitations (Davern et. al. 2009)) ; however, its existence calls the survey data usage in public policy analysis into question. Furthermore, administrative data have the significant limitation that they do not include information on persons eligible for a program but not enrolled, or those who would be eligible if the program rules changed. To help improve both sources of data for policy research, researchers are now linking survey and administrative data for policy research and analysis (Hotz et al. 1998).

Demographic survey data [1] are widely used for policy research because they cover entire populations of people and they contain policy-relevant information on people, families, and households. This information includes people's characteristics such as relationships to other household members, employment, income sources, income amounts, program participation, and much more. Surveys target people enrolled in public programs, people who are eligible but not participating, and people whose eligibility could change depending on program rules. Together this information can be used to simulate effects of a change in policy or program (e.g., does the Earned Income Tax Credit result in more labor force participation?), as well as to evaluate the

[1] Through this paper we shorten demographic household surveys to simply "surveys" and public program administrative data is often shortened to "administrative data".

effectiveness of a particular program at achieving its goals (e.g., does the State Children's Health Insurance Program result in lower rates of uninsurance for low income children?).

Survey data are crucial to policy makers and evaluators, and despite their many limitations, continue to be widely used because they are the only source of information rich enough to predict, evaluate, and understand the impact of potential and actual changes in policy on the entire population, both those currently affected by a policy and those who may be affected under a policy change. Data gleaned from public program administration that track payments to beneficiaries, such as Social Security checks, or to providers such as Medicare, are regularly combined with survey data to create linked data products with superior analytic capabilities than either source of data by itself. The potential uses of these linked data products in policy research are tremendous, but investigators need to understand that the potential for deriving incorrect inferences from these data is also great if the basic research into data quality is not conducted and disseminated.

Overview of the Limitations and Strengths of Survey Data

One of the great strengths of survey data, ironically, is the accumulated knowledge about their limitations. Much research has studied the components of survey error, including sampling and non-sampling error. We know that survey sampling frames have problems with population coverage (Groves 2004). The Current Population Survey (CPS), the U.S. Census Bureau's premier demographic survey, estimates sampling frame coverage to be only 93% (U.S. Census Bureau 2002). Additionally, survey data suffer from a growing problem of non-response. Over the last decade, response rates for personal interview surveys such as the CPS and the National Health Interview Survey (NHIS) have declined markedly. For the NHIS the household response

rate fell from 90% 1998 to a 86.5% response rate for the 2005 survey (National Center for Health Statistics, Division of Health Interview Statistics. 2000; 2006). Over the same period the response rate for the CPS Annual Social and Economic Supplement fell from 85.6% to 82.6% (U.S. Census Bureau 1998; 2005). For telephone surveys such as the Behavioral Risk Factor Surveillance System (BRFSS) and the Survey of Consumer Attitudes, response rates fell even faster (Centers for Disease Control and Prevention. 2006; Curtin et al. 2005).

We also know that surveys with complex sample designs have sampling error that can be easily underestimated (Korn and Graubard 1999; Davern et al. 2006). Furthermore, surveys have serious measurement problems with key concepts. For example, they produce widely varying estimates of the number of people without health insurance. The estimate of full-year uninsurance in 2001 from the U.S. Census Bureau's Survey of Income and Program Participation (SIPP) is 22 million individuals, half the size of the 44 million estimate from the CPS (Peterson 2005). Surveys also tend to undercount enrollment in public programs such as Medicaid and the Food Stamp Program (U.S. Census Bureau 2004; Lewis, Elwood, and Czajka 1998), partially as a result of enrolled people not answering questions correctly. There are also problems with editing, imputing, and processing survey data (Davern et al. 2004; Davern et al. 2007; Bollinger and Hirsch 2006). Finally, policy makers wish survey data would be released more quickly and that broader access be given to the microdata, including key elements that are often suppressed (Blewett et al. 2004; Davern et al. 2006).

Knowledge of limitations from past research has led to significant improvement in survey data, as well as a better informed community of data users. In response to research conducted by government agencies and outside researchers, data providers have improved public-use products. Knowing the limitations and problems motivates and facilitates improvements, or at least more

appropriate use of the data if improvement is not possible or forthcoming. And as data users have demanded more consistent and complete survey documentation, there have been improvements in methodology and documentation (Data Documentation Initiative 2007). Survey data have improved greatly for analytic purposes as a result of well-researched limitations and high-quality documentation available in the public domain.

Overview of the limitations and strengths of administrative data

The strength of administrative data lies in the reuse of an existing data resource, full coverage of the target population, and accuracy of key data items within the data set. Transaction data reflects automatic capture of payments and service usage more accurately than self-reporting and self-recall. Information required for service eligibility or delivery receives quality assessment during the delivery of services and correction of service problems. The temporal coverage of data collection coincides with the period of agency activity providing a complete record of both individual transactions as well as agency development and change over time. Changes in technology and increased use of administrative data during the 1980's and 1990's have improved both data retention activities (archiving) and data quality (Holenbeck, 2003).

However, the original purpose of administrative data is to internally tract agency and program activities. Focus is on program eligibility, current and recent transactions, and information directly related to the provision of service and the efficiency of the agency. Depending on the program and its funding and review processes, incentives to under- or over-report activity may exist. This is also true of those applying for services. While some data may be more accurate than that captured in a survey, individual data items may tend toward over- or under-reporting. In addition, data items irrelevant to receipt of services such as background

socio-demographic information may show a high level of item non-response. This could effect the availability of data items such as educational level that are important covariates for the researcher. Inter-state and intra-state data availability and consistency may also vary (Holenbeck, 2003).

Data collected for nationally administered data encourages greater consistency than those that capture information for more locally administered programs. National programs tend to use consistant definitions of terminology and data items. Prior to the devolution of data collection to the states that occurred in the 1990's, centralization aided the consistency of data collected for national programs. However, devolution pushed data collection back to the states and often to the county increasing the variation in how data was collected, the level to which definitions were enforced, and the consistency and coverage of data collection. For example, data could be collected by a state agency in one state, by multiple agencies within each county in another state, or not collected at all. This type of inconsistency is shows up in a number of national data sets. Some data items used for linking or geocoding are collected in ways that result in high error rates due to the use of mail delivery addresses. Mail addresses include P.O. Box numbers and rural routes. These addresses cannot not be geocoded accurately and fall disproportionately in six states. Issues like these can be addressed if known, however, the assessment of coverage, accuracy, and reliability may often be done only through data analysis.

Lack of documentation on concepts, data item definitions, collection and cleaning processes, changes in data management systems over time, and local variations that may affect data accuracy and consistency is a major issue for administrative data (Reidy, 1998). It is difficult to identify and locate the individuals able to provide this information after the fact. Moreover, the movement of administrative data from local collection points, to local agency data

management systems, and from there to national level data collections introduces multiple points where actions affecting data quality assessment may take place. Ideally, programmatic documentation along with local variations on data collection, cleaning and processing information should be available to the researcher. This is rarely the case for administrative data. This makes the type of data quality assessment done for survey data extremely difficult to perform for administrative data (Reidy, 1998).

A Research Agenda into Linked Data File Quality

All data – whether administrative, survey, or linked – have limitations for answering public policy questions. As discussed earlier, survey data problems are well-known to the policy research community because these data are in the public domain and heavily used. The data themselves are better because the limitations are known – even quantified – and improvements often follow research findings. Moreover, if a research paper fails to consider important issues or limitations, peer review often informs the author of the omission. A problem with most linked data files is that they are usually constructed for a limited research question or set of questions, are available only to a limited number of researchers, and cannot be released to the public according the Inter-Agency Agreement (IAA) or Memorandum of Understanding (MOU) negotiated to allow the linkage (Cox, Berning, and Wilkie Martinez 2006; Obenski 2006; Hotz et al. 1998; National Research Council 2000).

Such restrictions are necessary because of the risk of disclosing the identities of sampled persons. The linking process most often uses identifiers such as name, address, and Social Security number, so when the U.S. Census Bureau links a file they remove all this information

immediately and replace it with a Protected Identity Key (PIK). Even with this step, only authorized researchers working on an approved project can view the linked data, and only for the approved purposes. Furthermore, the U.S. Census Bureau requires the research to demonstrate a clear benefit to ongoing statistical programs. Extensive research into the quality of the linked data file for broad-based policy research is normally not undertaken because it is not part of the core research question the data were linked to answer. Although there are some exceptions where researchers can apply to work on the data in a secure environment, such as the Surveillance Epidemiology and End Results (SEER) cancer registry linked to administrative Medicare data (National Cancer Institute 2007), most linked data sets remain highly restricted and accessible only for very specific research questions. This situation prevents the research community from gaining familiarity with the quality issues associated with linked data files, and therefore from conducting adequate peer reviews.

How the federal statistical system balances the competing interests of providing access to data on one hand, and ensuring the privacy of individuals on the other, creates uncertainty about how much the research community will be able to learn about linked data sets. Under pressures of privacy concerns, it seems that more and more data fall under access restrictions and linked data files are unlikely to enter the public domain, hindering the growth of knowledge (Lane 2006). At the same time, because these data sets constitute such a rich and novel resource, a researcher who does gain access risks misusing them if he or she undertakes substantive research before conducting a thorough quality review.

We will attempt to outline a research agenda for assessing the quality of linked data files using the perennial sources of survey error as a guide. These concerns about data quality, familiar to all researchers who use survey data, will navigate us through the following potential

problems, and some possible solutions, associated with linked data files: sampling frame coverage error, sampling error, non-response error, measurement error, data processing issues, imputation procedures, editing rules, documentation, and timeliness of the data.

Coverage Error in the Sampling Frame

Sample coverage error in a linked data set is largely a function of the coverage error in the survey data that persists in the linked data file. To the extent that the sampling frame fails to properly represent the population of interest, the linked data file will inherit this problem. For example, survey data tend to have poor coverage of minority members, males, and young adults; the group with the lowest sample coverage ratio in the CPS is black males aged 20-29, who only have a 0.66 coverage ratio, while people in the 60 and over category have coverage ratios much closer to 1.0 (U.S. Census Bureau 2002). Weighting and post-stratification adjustments to the survey weights attempt to minimize the impact of these coverage problems. But to the extent that these adjustments are inadequate, the linked data files will also have issues of sampling frame coverage.

The administrative data themselves do not have to contend with coverage issues because everyone covered or "administered" by a program should have a record on the administrative data file. However, not everyone administered by a program is part of the survey's sampling frame. Biased conclusions about aggregate discrepancies between survey estimates and administrative data can result from a poorly-defined set of individuals who are eligible to be linked to the survey because they are known to be in the survey sampling frame. Even if the sampling frame is available, enforcing a survey universe on administrative data is difficult

because the administrative data may lack the geographic detail necessary for linking to the sampling frame.

Two examples demonstrate this problem. First, administrative data and survey sampling frames can entail different concepts of group quarters or institutions. People living in such places are excluded from some survey sampling frames. Individuals in the administrative data may be thought to be within the survey sampling frame because they are not in "group quarters" according to the administrative agency's definition, but they could, in fact, be living in group quarters by the survey's definition. The reverse is possible as well. For example, someone's address in administrative data could be a mailing address for a guardian or relative, while the individual actually lives in an Institutional Group Quarter such as a residential nursing facility. Second, the administrative data could contain more than one address for an individual, and it is difficult to determine from the information available when the person lived at each address, or where they lived at the time of the survey.

An important measure of quality is to establish a "linkable universe ratio." This statistic is the ratio of the weighted number of linked individuals to the number of people in the administrative data who are in the survey's sampling frame. This linkable universe ratio is imperfect as there often will not be enough information on the administrative data to reconstruct the survey sampling frame and target population. This is a problem in linked data which needs further investigation by researchers who have access to the data.

Sampling Error

Sampling error is not typically a problem for research on administrative records because all the records are available and a sample is not necessary. However, when linking survey data

to administrative records, the linked data file carries with it the limitations of the survey's complex sample design. Estimates derived from the linked file need to take the complex sample design into account when producing standard errors and variances for statistical inferences (Korn and Graubard 1999). Procedures for estimating design-based sampling errors are available in most statistical packages such as SAS, STATA, SUDAAN, and SPSS, and should be invoked.

Non-response Error

Item non-response is a major barrier to working with linked survey data and administrative data files. Surveys have unit and item non-response, which imputation and post-stratification weighting adjustments account for in order to make the responding population more representative of the population as a whole. Imputation and post-stratification weighting techniques have been evaluated for survey data and have been found to perform well when the critical assumptions are met (Little and Rubin 2002). The most critical assumption is that the responding population is representative of the non-responding population after controlling for key demographic characteristics (i.e., Missing at Random), through post-stratification for unit and person non-response, and multivariate modeling and imputation for item non-response.

This issue is complicated when linking because survey and administrative data often lack linking identifiers, sometimes systematically. For example, Figure 1 shows results from a recent data linkage study conducted by the U.S. Census Bureau, the Centers for Medicare and Medicaid Services, the State Health Access Data Center, and the Department of Health and Human Services Assistant Secretary for Planning and Evaluation. Nationwide, 90 percent of the identifiers (i.e., SSNs) in the Medicaid Statistical Information System (MSIS) are verified, but this figure is much lower in some states and counties. California and Montana, in particular,

stand out with respect to missing identifiers (U.S. Census Bureau 2007). This means people from these states who are enrolled in Medicaid will not have the same probability of being linked to a U.S. Census Bureau survey.

--Insert Figure 1 about here--

In this study, the MSIS data were matched to the CPS data for 2001 and 2002. In both of these years, over 26 percent of the CPS persons were missing linking identifiers. This happens for two reasons. First, respondents refuse to provide the linking information (the Social Security number) and are interpreted as refusing to have their data linked, and consequently the U.S. Census Bureau is disallowed from linking their data to administrative records. This accounts for about 20 percentage points of the missing identifiers in the CPS, which can be dealt with in the usual manner assuming that those that refuse to provide the identifier are missing at random. The remaining 6 percent are cases in which the SSN provided could not be validated and the correct identifier could not be found (Davern 2007). It is quite possible this group is not "missing at random" in Little and Rubin's (2002) terminology, because some people, namely children under 1 year of age and immigrants, simply do not have an SSN. This group is not well represented by those in the data set with validated SSNs.

When working with linked data files, it is essential to understand whether systematic differences exist between those cases that were linked and those that were not linked. Both the survey and the administrative data likely have systematically missing linking identifiers. Agencies that have access to linked data files must endeavor to understand these sources of sample loss and how they can influence or bias the research that is performed using the linked

cases. In the case of the CPS-MSIS linked data file, 10 percent of the MSIS cases could not be linked because they were missing SSNs and 20 percent of the CPS data could not be linked because they refused the SSN or one could not be found to match the person. This is a large sample loss and needs investigation to better delineate the limitations of the linked data file for research purposes (U.S. Census Bureau 2007; Davern 2007).

Under some circumstances, if we understand how the linking identifier from the survey and administrative files is missing, we can control for minor differences between the populations represented by the linked and unlinkable cases (i.e., those cases with and without linking identifiers). It may be appropriate to create a post-stratified "linked" record weight that adjusts for sample loss. One way to accomplish this is to increase the survey weights of those linkable cases (i.e., those with verified linkable information, not just those that are linked) to the full weighted sample size for the survey, treating the unlinkable cases similarly to survey non-respondents, and stratifying by the characteristics of greatest interest.

Measurement Error

Survey data have extensive measurement error issues that have been the subject of many investigations (Groves 2004; Dillman 2000). This work has shown that survey measurement errors can be impacted by mode of interview, question wording, question appearance on the page, survey question order, reference period of the question, incentives for completion, interviewer effects, and proxy response. All of these errors could impact administrative data as well, and investigation into this possibility is critical for working with and understanding linked data files.

As stated earlier, there is little problem taking the core administrative portions of the administrative data as truth. For example, there is little reason to dispute that an SSI recipient or an SSA recipient was sent a check in the amount on the administrative data file on the date specified. However, the administrative data contain many more items than just pure programmatic details such as timing and amount of payments made. They contain information on age, sex, marital status, addresses, telephone numbers, race, ethnicity, and other program-relevant variables. For example, Medicaid data contain information on participation in complementary public programs such as the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC), Food Stamps, and TANF. In addition, the values reported to some administrative agencies for program purposes, such as the IRS for tax reporting, may systematically differ from a value the same person would report when asked in a survey. The incentives associated with reporting income are different. Reporting less income to the IRS or to the Medicaid office reduces taxes owed or can make someone eligible for a program, whereas there is no direct incentive to report lower income in a survey. Therefore, income amounts from the two sources on a linked data file could represent systematically different things.

Administrative data can be collected through many modes (i.e., interviewer or self-administered), during more than one wave of interviewing, and using several instruments. Very little information is kept on the origin of variable values, but this information is an important issue that managers of administrative data systems should address. In a survey, information on where each data element came from is called "paradata" and it is useful for research (Couper 2005). Survey research has demonstrated that the answers to questions can vary by survey mode, and this could be the case for administrative data as well. Furthermore, some interviewers collecting administrative data may have a great deal of training. For example, a hospital

employee whose job it is to help patients fill out forms for Medicaid is an expert in eligibility, as are tax accountants who fill out tax forms for clients. Forms filled out by professionals may vary in systematic ways from those filled out by someone who completes their own personal application for a program such as Medicaid, or files their own taxes. The difference may be even more pronounced if the first language of the applicant is not English or the applicant cannot read. How the information enters the administrative database, or paradata, should be carefully tracked so researchers can look for variation by the mode of data collection.

Self-administered data collection forms often fail to be user-friendly or to incorporate the advances made in survey questionnaire design developed over many years (Dillman 2000). Administrative data collectors are not as concerned about usability of their forms, as there is generally a strong incentive to complete a quirky form in order to collect benefits such as Medicaid, childcare assistance, or TANF. However, the forms can cause measurement errors with key pieces of demographic information, and research should be conducted on to create user-friendly self-administered instruments that ensure high-quality supplemental demographic data on age, date of birth, sex, marital status, race and ethnicity. This is an area where the research on surveys greatly surpasses the research on administrative data collection. Knowledge about the quality of administrative data is essential for properly exploiting linked data files, and a necessary first step is to collect better paradata in administrative systems.

Data Processing, Editing Rules, and Imputation Procedures

Survey data and administrative data are routinely edited and imputed, but unlike survey data there is little documentation in the public domain regarding the editing and imputation procedures applied to administrative data. Because imputation and editing can influence

estimates substantially it is important to know how the procedures work. A process that makes perfect sense from one point of view (for example, administering a program) may not be appropriate for specific research purposes (for examples of such instances see, Davern et al. 2004; Davern et al. 2007; Bollinger and Hirsch 2006). Custodians of administrative data need to produce appropriate metadata on how items have been edited or imputed in order for linked data files to fulfill their potential. The assumption we make is that administrative data are edited and imputed in a way that optimizes their use in administering programs but may not be appropriate for researchers to use the edited or imputed administrative data for analytic purposes.

When working with linked data files it is especially important to recognize that data missing from the survey may have been imputed or edited in a sensible way from the survey processing perspective, but not when directly comparing administrative data and survey data on individual linked records. The purpose of imputation is to make population and subpopulation estimates accurate, not the individual-level values. Probabilistic, model-based, imputed values in the survey will agree much less with the administrative data than will the values reported by respondents. A linked MSIS-CPS case that had Medicaid coverage imputed in the CPS is less likely to be in the MSIS data set than a case with a reported value (Davern 2007). Imputation is not supposed to accurately put individuals into the "correct" category; rather, it is an attempt to make the estimates for certain classes of people correct. That is, imputation may not correctly classify a single 15 year-old Hispanic girl, but a statistic for 15 year-old Hispanic girls should be accurate in the aggregate, derived from reported and imputed data. Distinguishing reported and imputed data is essential in work with linked data files, depending on the purpose of the research.

Furthermore, administrative data can often systematically lack information in some non program-specific fields. For example, in the 2003 MSIS, race and ethnicity information was

listed as “unknown” for more than one in five Medicaid enrollees in Rhode Island, New York, and Vermont. A recent National Academy of Sciences panel that reviewed the quality of data available to study race and ethnic health disparities concluded that the "CMS does not yet have any information on the quality of the racial and ethnic data collected through the MSIS” (National Research Council 2004b, p.83). In fiscal year 2000, the race and ethnicity of three million of the total 44.5 million Medicaid enrollees (about seven percent) were reported as "unknown" (National Research Council 2004b). In addition, some cross-reference program material can be systematically missing from administrative data, depending on how the data were processed. For example, the variable for the receipt of TANF in the MSIS shows no one in Los Angeles County, California receiving TANF, clearly an anomaly of the data set (Centers for Medicare and Medicaid Services 2007). Using a variable without taking such anomalies into account can obviously lead to incorrect conclusions.

Data Access

There are significant time delays, availability, and disclosure issues associated with linked data files. For example, the U.S. Census Bureau received the 2003 MSIS data file in the summer of 2006. As discussed earlier, linked data files are highly protected due to the sensitive nature of both the survey data and the administrative data, and only a few people in a restricted environment typically have access. The IRS-U.S. Census Bureau linked data file, for example, has especially tight restrictions on who can access to it and for what purposes it can be used. Most data linkage projects are for a specific purpose only and use of the linked data files is tightly controlled due to negotiated IAA and MOUs among federal agencies (Hotz 1998; National Research Council 2000; Cox, Berning, and Wilkie Martinez 2006; Obenski 2006).

Moreover, in many cases tables produced using the linked data files also need to be cleared through both the agency that conducted the survey and the agency that provided the administrative data. So not only are the data not recent and generally inaccessible to outside researchers, but also subject to lengthy clearance procedures.

It is possible that as agencies begin to understand better ways of protecting confidentiality some access restrictions may be lifted. For example, the U.S. Census Bureau is currently investigating ways to enhance the Survey of Income and Program Participation through routine linking of survey and administrative data files into a Dynamics of Economic Well-being System. And even though public use of these data files may not be possible without sophisticated disclosure-proofing procedures, perhaps it could be possible to make the the data available at the U.S. Census Bureau's Research Data Centers (RDCs). In the RDCs, researchers could access the linked data files in a protected environment, and the collaborating agencies (for example, the U.S. Census Bureau and SSA) could develop a set of standards to be enforced by the RDC during disclosure review. The RDCs also require research to provide a direct benefit to the U.S. Census Bureau. Perhaps researchers could meet this requirement by evaluating the quality of the linked data, producing a report about the data quality issues they encountered in their research and how they overcame them. To the extent that the issues raised in this paper pertain to a particular linked data file, a researcher at an RDC can justify a project by investigating them in more detail as part of the research agenda.

Potential Uses of Linked Data files

Administrative data typically entail records collected or maintained by federal, state, tribal, or local government agencies, or commercial entities; not for demographic statistics or policy analysis, but for administering programs or providing services. For this reason, an administrative data set such as the Medicaid Statistical Information System (MSIS) contains basic information about persons enrolled in state Medicaid programs, but lacks the socioeconomic and demographic details provided by surveys and needed for policy research. Conceptually, therefore, combining the survey and the enrollment information from the MSIS records might provide a richer, more complete, and more accurate view of the Medicaid program, its participants, and the eligible but unenrolled population (Hotz et al. 1998). Specific possible applications include correcting bias in imputation, direct editing of survey items, quality evaluation of response and editing and imputation procedures, and improving the coverage of sampling frames (e.g., Chappell, Obenski, and Farber 2005; National Research Council 2004a).

Linked administrative data files can identify and correct problems with the survey data.[2] For example, the U.S. Census Bureau has linked its survey data to Internal Revenue Service (IRS) and Social Security Administration (SSA) administrative data in order to compare survey reports of income to the IRS and SSA records. Now in the CPS, the U.S. Census Bureau adjusts interest income upwardly for individuals for whom the amount was imputed (not reported),

[2] In this paper we are dealing with after-the-fact linking. The survey and administrative data were independently collected and linked after the fact. This paper is still relevant in some areas for cases where administrative data themselves were used as the survey sampling frame (such as is done with the Medicare Current Beneficiaries Survey) but not all the issues discussed will apply to these data collection efforts.

based on the relationships observed in a CPS and IRS linked data file (Nelson 1985). With linked data sets, data providers could develop similar adjustments for other survey items.

Linked data files can greatly improve simulation models for policy evaluation and review purposes. Research has documented that a significant proportion of enrollees in Medicaid and the Food Stamp Program do not report being enrolled in response to survey questions asked within the correct reference period (U.S. Census Bureau 2004; Call et al. 2002; Klerman, Ringel, and Roth 2005).[3] This discrepancy could be attenuated through the use of linked data by editing each person's self-reported status to be consistent with the administrative data. Both the Urban Institute's TRIM 3 model and the Actuarial Research Corporation's model of Medicaid receipt could benefit from a linked data file created by the U.S. Census Bureau from CPS and MSIS. The resulting linked data would better simulate the impact of Medicaid program changes (Urban Institute 2007; Callahan, Mays, and Brenner 2005).

Administrative data cannot be used to directly edit the survey data (because of disclosure issues) (Chappell, Obenski, and Farber 2005). Despite these limitations, research on linked data files which evaluates the quality of survey responses about public program participation is informative; for example, these files can help assess how well respondents answer questions about participation in Medicaid, Food Stamps, Temporary Assistance for Needy Families (TANF), and Supplemental Security Income (SSI). Such research could lead to better survey questions that extract more accurate information from respondents. Furthermore, comparing survey and administrative data can lead to better informed imputation methodologies for survey

[3] Furthermore, some people who cannot be found on the administrative data file report receiving some of these benefits. These respondents might have given an incorrect answer to the question, although it could also be a problem with the linking information.

data, and perhaps even model-based imputation methodologies to generate more accurate estimates regarding program participation.[4] The bottom line is that research with these linked data files will ultimately facilitate improvement in the accuracy of surveys. They will provide better knowledge about who is likely to make errors and how these errors can influence estimates, not only of the number and characteristics of people enrolled in a program, but also the number and characteristics of people eligible and not enrolled.

Linked data files have great potential to help the U.S. Census Bureau improve data collection by building better survey sampling frames; for example, improving the Master Address File used for many sampling purposes at the Census Bureau (National Research Council 2004a). Administrative data that include address information can be especially informative if they contain addresses of individuals who are likely to be missed on the current sampling frames. Research into sample coverage error shows that the sampling frames are likely to miss low-income people, and these are people who are eligible for public programs such as Medicaid, Food Stamps, or TANF (U.S. Census Bureau 2002). The U.S. Census Bureau sampling frames could be systematically missing these low-income households, and therefore undercount the number of poor people and the number of people participating in public programs. Linking addresses contained in such programs' administrative data files to the sampling frames would allow the U.S. Census Bureau to identify addresses that may have otherwise remained missing from the frames.

---

[4] Linked data files are not likely to produce an edit where, for example, a linked respondent who is enrolled in Medicaid actually has their value changed in the public use file because of disclosure issues. The disclosure risk comes from the administrative data agency that can more easily identify the survey record of individual people using the linked data file.

Linked data files can also help improve administrative data. For example, many administrative data elements such as race, ethnicity, education, and marital status are of suspect quality because they are not essential fields for the administration of programs, and therefore not collected consistently. Survey data could shed light on the nature of the data missing from administrative data sets within these domains. In particular, survey data could improve the outdated race and ethnicity information in some systems that was collected before Office of Management and Budget Directive 15 went into effect in 1997. For example, the Social Security Administration has collected race and ethnicity information for many years, during which time official definitions have changed repeatedly. Race and ethnicity information conforming to current definitions and applied to SSA data and Centers for Medicare and Medicaid Service (CMS) data sets could greatly improve research in the field of health disparities (National Research Council. 2004b, p. 83).

Conclusions

Linked survey and administrative data constitute a valuable means of policy-relevant research. These files are being created by a variety of federal agencies such as the U.S. Census Bureau, the Veterans Administration, the Social Security Administration, the National Center for Health Statistics, the Internal Revenue Service and the Centers for Medicare and Medicaid Services, to name a few (National Committee on Vital and Health Statistics 2006). Some of these projects are ongoing partnerships with strong linked data research agenda (e.g., the linked National Health Interview Survey and Medicare Claims data, or the LEHD program at the US Census Bureau) while many others are more *ad hoc*. Each source of data has limitations by itself that the other can potentially help overcome. Survey data are very strong because their

limitations are well-understood. Similar research into issues with administrative data is necessary to enhance our understanding of the limitations and possibilities for policy research to use linked survey and administrative data.

Although linked microdata itself cannot be released to the public without substantial alteration, publicly available documentation of the files would nevertheless maximize the usefulness of the research based on them. One way to achieve this level of documentation would be to consider a linked data file a "data product" and to systematically conduct research into its quality. The results of this research could be assembled into a coherent set of metadata in the public domain, allowing researchers to understand the file's usefulness and idiosyncrasies. The documentation should attempt to develop and use a systematic standard such as the Data Documentation Initiative developed for survey data (Data Documentation Initiative 2007). Creating an ongoing linkage program takes a considerable amount of resources and the costs of producing a high quality linked data product should not be under-estimated.

The issues presented in this paper serve as a summary of the opportunities and challenges associated with linking survey data and administrative data to produce resources for policy-relevant research. Extensive experience with public-use survey data has allowed the research community to develop a wealth of information on working within limitations, and similar knowledge is necessary for administrative data as a stand-alone source of information or in combination with a survey. This statement should not deter researchers from creating and analyzing linked files. Rather, it is an invitation to pursue data linkage projects and establish a system of sharing information on methodologies for exploiting them intelligently. Combined

with improvements to the quality of administrative data, such projects can provide robust analyses, high-quality information, and ultimately better public policy choices.

References

Blewett, L.A., Good, M.B., Call, K.T., and Davern, M. (2004). Monitoring the uninsured: A state Policy perspective." *Journal of Health Politics, Policy and Law*. 29(1):107-45.

Bollinger, C.R. and Hirsch, B.T. (2006). Match bias from earnings imputation in the Current Population Survey: The case of imperfect matching. *Journal of Labor Economics* 24(3):483-519.

Callahan, C.M., Mays, J.W., and Brenner, M. (2005). A Longitudinal Model of Health Insurance: Employer Sponsored Insurance, Medicaid and the Uninsured. Working Paper prepared for the Office of the Assistant Secretary for Planning and Evaluation. Washington DC: Department of Health and Human Services; March 31.

Call, K.T., Davidson, G., Sommers, A.S., Feldman, R., Farseth, P., and Rockwood,T. (2002). Uncovering the missing Medicaid cases and assessing their bias for estimates of the uninsured. *Inquiry* 38 (4):396-408.

Centers for Disease Control and Prevention (CDC). (2006). *2005 Behavioral Risk Factor Surveillance System Data Quality Report Handbook*. Atlanta GA: Centers for Disease Control and Prevention.

Centers for Medicare and Medicaid Services (CMA). (2007). Medicaid Data Sources – General Information, MSIS Data web page. http://www.cms.hhs.gov/MedicaidDataSourcesGenInfo/02_MSISData.asp.

Chappell, G., Obenski, S., and Farber, J. (2005). Research to Improve Census Imputation Methods: Item Results and Conclusions. Presentation at the Joint Statistical Meetings of the American Statistical Association, Survey Research Methods Section. Minneapolis MN, August 10.

Couper, M.P. (2005). Technology trends in survey data collection. *Social Science Computer Review* 23(4): 486-501.

Cox, C., Berning, M., and Wilkie Martinez, S. (2006). Data Policy and Legal Issues in Creating and Managing Integrated Data Sets. Presentation to the Federal Committee on Statistical Methodology, Statistical Policy Seminar, Washington DC, November 28.

Curtin, R., Presser, S., and Singer, E. (2005). Changes in Telephone Survey Nonresponse over the Past Quarter Century. *Public Opinion Quarterly* 69 (1): 87-98.

Data Documentation Initiative. (2007). http://www.ddialliance.org/index.html

Davern, Michael, Jacob Alex Klerman, David Baugh, Kathleen Call, and George Greenberg. 2009. "An Examination of the Medicaid Undercount in the Current Population Survey (CPS): Preliminary Results from Record Linking." *Health Services Research.* 44(3): 965-987

Davern, M., Blewett, L.A., Bershadsky, B., and Arnold, N. (2004). Missing the Mark? Examining Imputation Bias in the Current Population Survey's State Income and Health Insurance Coverage Estimates. *Journal of Official Statistics* 20(3): 519-49.

Davern, M., Jones Jr., A., Lepkowski, J., Davidson, G., and Blewett, L.A. (2006). Unstable Inferences? An Examination of Complex Survey Sample Design Adjustments Using the Current Population Survey for Health Services Research. *Inquiry* 43(3): 283-97.

Davern, M., Rodin, H., Call, K.T., and Blewett, L.A. (2007). Are the CPS Uninsurance Estimates too High? An Examination of Imputation. *Health Services Research* forthcoming, doi: 10.1111/j.1475-6773.2007.00703.x

Davern, M. (2007). Fitting Square Pegs into Round Holes: Linking Medicaid and Current Population Survey Data to Understand the 'Medicaid Undercount'. Presentation to the Office of Research, Development, and Information Seminar Series, Centers for Medicare and Medicaid Services, Baltimore MD, February 15.

Dillman, D. (2000). *Mail and Internet Surveys: The Tailored Survey Design*, 2nd edition. John Wiley and Sons: New York.

Groves, R. (2004). *Survey Errors and Survey Costs*. John Wiley and Sons: New York.

Holenbeck, K., King, C.T., and Schroeder, D. (2003). Preliminary WIA Net Impact Estimates: Administrative Records Opportunities and Limitations. *New Tools for a New Era! Symposium.* Bureau of Labor Statistics & the Workforce Information Council: Washington, DC.

Hotz, J.V., Goerge, R., Balzekas, J., and Margolin, F., editors. (1998). Administrative Data for Policy-Relevant Research: Assessment of Current Utility and Recommendations for Development, A Report of the Advisory Panel on Research Uses of Administrative Data of the Northwestern University/University of Chicago Joint Center for Poverty Research. http://www.econ.ucla.edu/hotz/working_papers/adm_data.pdf

Klerman, J.A., Ringel, J.S., and Roth, B. (2005). Under-reporting of Medicaid and Welfare in the Current Population Survey. Working Paper. Santa Monica CA: RAND.

Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys.* John Wiley and Sons: New York.

Lane, J. (2006). Using Linked Data. Presentation at the Workshop on Data Linkages to Improve Health Outcomes, sponsored by the National Committee on Vital and Health Statistics, Subcommittee on Populations, Washington DC, September 18-19.

Lewis, K., Elwood, M., and Czajka, J.L. (1998). Counting the Uninsured: A Review of the Literature. Washington DC: Urban Institute. http://www.urban.org/url.cfm?ID=308032.

Little, R.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. John Wiley and Sons: New York.

National Committee on Vital and Health Statistics (NCVHS). (2006). Meeting of the Subcommittee on Populations. Transcript and presentation web page. Washington DC, November 18-19. http://www.ncvhs.hhs.gov/060918ag.htm

National Center for Health Statistics (NCHS), Division of Health Interview Statistics. (2000). National Health Interview Survey (NHIS) public use data release. 1998 NHIS survey description. Hyattsville, MD: Centers for Disease Control and Prevention, October. http://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/1998/srvydesc.pdf

National Center for Health Statistics 9NCHS), Division of Health Interview Statistics. (2006). National Health Interview Survey (NHIS) public use data release. 2005 NHIS survey description. Hyattsville, MD: Centers for Disease Control and Prevention, June. http://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2005/srvydesc.pdf

National Research Council. (2000). Improving Access to and Confidentiality of Research Data: Report of a Workshop. Washington, DC: The National Academics Press.

National Research Council. (2004a). *Reengineering the 2010 Census: Risks and Challenges*. Washington, DC: The National Academics Press.

National Research Council. (2004b). *Eliminating Health Disparities: Measurement and Data Needs.* Washington, DC: The National Academics Press.

Nelson, C. (1985). Adjusting Imputed Interest Amounts Based on Results of the CPS-IRS Exact Match. Unpublished memorandum for Chief of Income Statistics, U.S. Bureau of the Census.

Peterson, C. (2005). Survey Estimates of the Uninsured and of Medicaid/SCHIP Enrollees. Presentation at the American Enterprise Institute's event, "9 Million Less Uninsured?" Washington DC, April 8.

National Cancer Institute. (2007). SEER-Medicare Linked Database web page. Bethesda MD: National Cancer Institute. http://healthservices.cancer.gov/seermedicare/

Obenski, S. (2006). How Recent Advances Have Facilitated Comparing, Analyzing and Jointly Using Large-Scale Survey and Administrative Data to Answer Big Policy Questions. Presentation to the Federal Committee on Statistical Methodology, Statistical Policy Seminar, Washington DC, November 28.

Reidy, M., George, R., and Lee, B.J. (1998). Developing an Integrated Administrative Database, *Exploring Research Methods in Social Policy Research.* Asldershot, UK: The Ashgate Publishing Company.

Urban Institute. (2007). Transfer Income Model 3 (TRIM3) project website. http://www.trim3.urban.org.

U.S. Census Bureau. (1998). Current Population Survey March 1998: Technical Documentation. Washington, DC: U.S. Census Bureau.

U.S. Census Bureau. (2005). Current Population Survey 2005 Annual Social and Economic Supplement: Technical Documentation. Washington, DC: U.S. Census Bureau.

U.S. Census Bureau. (2002). Current Population Survey: Design and methodology. Technical Paper #63RV. Washington, DC: U.S. Census Bureau.

U.S. Census Bureau. (2004). Difference in estimates of Food Stamp Program participation between surveys and administrative records. U.S. Census Bureau: Washington DC. http://www.census.gov/acs/www/Downloads/ACS/FoodStampFinalReport.pdf

U.S. Census Bureau. (2007). Phase I research results: Overview of National Medicare and Medicaid Files. Report of the research project to understand the Medicaid undercount: The University of Minnesota's State Health Access Data Assistance Center, the Centers for Medicare and Medicaid Services, the Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation, and the U.S. Census Bureau. Washington DC: U.S. Census Bureau.

Figure 1: U.S. Counties by their Rate of Validated Social Security Numbers in the Medicaid Statistical Information System (MSIS): 2001