NORC at the University of Chicago

CHRR AT THE OHIO STATE UNIVERSITY

TASK 1.3.4 ALTERNATIVE DATA SOURCES FINAL REPORT

May 27, 2022

Contract Deliverable

Presented by:
Roxanne Wallace, PMP

Document Change Log

Published Date	Document Version No.	Pages Affected	Description of Revision	Author
5/27/2022	v1	2,3	BLS Comments were addressed and clarifying information added	Erik Scherpf, Andrew Stern

Executive Summary

This report assesses alternative data sources (ADS) that could be integrated into a new NLS cohort as a way to improve the accuracy of survey data collected, reduce respondent burden, or expanding the scope of the survey content. Data for this report were gathered from a scan of publicly available sources, including secondary research using these alternative data sources.

Excluded from this final report were data sources that primarily would inform sampling and recruitment of the sample. Instead, this report focuses on those data sources that are both integral to the NLS mission and that could be used to enhance the NLSY26 data files, whether through direct replacement of survey items, edit and imputation following data collection, or as an auxiliary data file that could be used in conjunction with the NLSY26.

Drawing on the data quality framework recently proposed by the Federal Committee on Statistics and Methodology (FSCM): Each alternative data source was assessed along five dimensions: relevance, accuracy, coherence, feasibility, as well as confidentiality, consent, and accessibility. Under relevance, we have attempted to provide information on the content of the ADS that will inform BLS' decision of whether the data source would meet the needs of prospective data users of a new NLS cohort. This includes, to the extent possible, variables likely to be of interest, the timing with which data are collected and released, and the granularity of the data elements. Under accuracy, we have characterized the potential for linkage error, the coverage of the ADS as well as any other measurement and accuracy issues. Under coherence, we identify survey items from the NLSY97 that each ADS may bear on.

Under feasibility, we attempt to characterize how difficult it might be for NLS to gain access to the ADS. This report distinguishes between ADS available at the federal level and those available at the state level. Or more specifically, distinguishing between data that are available from a single, national repository versus those that would require data sharing agreements with multiple data owners (e.g., states). Evidence of prior data sharing and data linkage for research purposes also figures prominently in our assessment of the feasibility of linking that data source to the NLSY26. Data sources that have been part of data linkage efforts at Census or NCHS should be more feasible candidates for integration with the NLS.

Lastly, under confidentiality, consent, and accessibility of the ADS, we report any relevant publicly available information on laws or regulations governing data access and use.

Introduction

This report explores the ways that alternative data sources could be integrated into a new NLS cohort. Capturing information through record linkages can decrease the number of items collected during the survey and reduce respondent burden. Reducing respondent burden could help address declining response rates in national surveys. Alternative data sources (ADS) could also be used to impute missing survey data due to nonresponse or otherwise improve the accuracy of the data collected. ADS can also be used to expand the scope or content of a survey by providing data that respondents would not be able to provide themselves.

Data for this report were gathered from a scan of publicly available sources, including secondary research using these alternative data sources. We did not engage any data owners directly in preparing this report. Data elements are also summarized in the ADS Evaluation List, an Excel file that is also part of this deliverable.

We organized information from each ADS along five dimensions: relevance, accuracy, coherence, feasibility, as well as confidentiality, consent, and accessibility.[1] Under relevance, we have attempted to provide information on the content of the ADS that will inform BLS' decision of whether the data source would meet prospective users' needs. This includes, to the extent possible, variables likely to be of interest, the timing with which data are collected and released, and the granularity of the data elements. We have provided additional information on when in a NLSY26 respondent's lifecycle the prioritized ADS are likely to be relevant (see Appendix Table B).

Under accuracy, we have collected secondary evidence on dimensions of data quality of the ADS. To assess each data source's suitability for data linkage and the potential for linkage error, we documented other data linkage efforts that used the ADS and any relevant information on those efforts. Further, under accuracy, we have attempted to characterize the coverage of the ADS as well as any other measurement and accuracy issues, such as measurement error. However, without direct access to the data, we have had to rely on secondary sources for this information.

Under coherence, we identify survey items from the NLSY97 that each ADS may bear on. We note that the notion of coherence is perhaps less fixed here than it would be if the ADS were being assessed for an existing survey, which is why the discussion of this dimension tends to be brief for most data sources. We also provide links to any available data dictionaries and codebooks.

Under feasibility, we attempt to characterize how difficult it might be for NLS to gain access to the ADS. For this assessment, we look to other instances in which the data sources were shared. In particular, we highlight whether the data owner has previously shared the data source for other household survey linkage efforts, such as with the Census Bureau or with NCHS. These precedents would suggest that a similar linkage with the NLS would be feasible. Several administrative data sources are linked for

[1] We note that we have not covered two alternative data sources on the final: ACA Enrollment Data and Medicaid Analytic eXtract (MAX) data. We discovered that the former appears to be derived from the MIDAS data. The latter has not been released since 2015, and it's not clear whether it will be updated again. It looks like it may be superseded by the T-MSIS Analytic Files (TAF).

operational or program administrative purposes, but this does not necessarily mean that a linkage to a household survey like the NLS for research purposes would be likely.

Lastly, we have attempted to collect any relevant publicly available information on laws or regulations governing data access and use. Here we primarily present this information, without offering an interpretation or assessment. We note here that, because BLS does not operate under title 13—as Census does—and based on the information gleaned from the NCHS data linkage effort, consent by respondents of a new NLS cohort would be required for linkage with each of these data sources.

Of the roughly 40 ADS collected as part of the preliminary assessment, BLS identified 20 for further investigation in the final ADS assessment. Those data sources that were eliminated would have mainly informed sampling efforts. The remaining 20 ADS were deemed relevant for enhancing data files, either through direct substitution, editing and imputation, or as an auxiliary data file. The complete list of ADS initially considered is provided in appendix table A.

Of the ADS considered in the final assessment are organized in three groups. The first group, listed below, are those that, after NORC assembled the materials herein, BLS determined to conduct its own follow up, with the goal of producing a companion report that details those sources:

- National Directory of New Hires (NDNH)
- Social Security Administration (SSA) data
- Unemployment Insurance (UI) wage data

The second group of ADS, listed below, were those prioritized by BLS for the richness of information that each source can provide

- Criminal Justice Administrative Record System (CJARS)
- Housing and Urban Development (HUD) data
- Medicaid/CHIP data
- National School Clearinghouse (NSC)
- National Student Loan Data System (NSLDS)
- US Veteran Affairs (USVA) data

The final group of ADS were those for which, after the included information was assembled, BLS indicated that collection of further information on these sources should be de-prioritized for a number of reasons, such as sufficiency of section, low likelihood of BLS pursuing the data, and timing issues. These data sources included:

- All-Payers Claims Database (APCD)
- Credit Agency Data (Experian, Equifax, FICO)
- IBM MarketScan
- MIDAS
- National Death Index (NDI)
- SNAP/WIC data

Potential Data Uses

Even for alternative data sources for which there is national coverage of the target population, direct replacement of survey items is unlikely to be viable. An adaptive design for integrating alternative data sources might be one alternative. Under this design, rather than eliminating a survey item entirely (and replacing it with a value from an administrative record match), the survey instrument will only skip the question—and use the administrative record value—if a record match is present. If a record match in not present, the survey instrument will ask the respondent the survey item and the survey-reported value will be used. The adaptive design is generally easiest to integrated with internet and computer-assisted interview modes that use automated data collection.

Short of direct substitution or adaptive collection, alternative data sources could also be used to edit and impute NLSY survey items, even when coverage in the ADS is less than complete. This could entail a more straightforward editing of survey edits based on administrative data linkage post-data collection. For ADS with substantially less than national coverage, newer developments in imputation might make even these ADS potentially viable candidates for enhancing the accuracy of the NLSY26. For example, Rothbaum et al. (2021), used SNAP administrative records from only eight states—including some smaller states—to impute SNAP participation to the “rest of the country” in the CPS-ASEC. Mittag (2019) similarly employed a conditional distribution method to impute SNAP receipt in the ASC using administrative records.

Lastly, an ADS could serve as an auxiliary source of data that would not otherwise be available in the NLSY26 or as a data source that could supplement data collected in the NLSY26 in some way. An example of the former might include information on respondent credit scores, an item that is not likely to be asked of respondents in the survey. The latter use case would be similar to the postsecondary transcript study in the NLSY97. The transcript study provided an administrative record analogue to a number of items in the NLSY97, and was useful even though it did not cover all NSLY97 respondents who attended a postsecondary institution. An example of an auxiliary data source for the NLSY26 might be linked CJARS data, which could be used in conjunction with the NLSY26. CJARS data, though not national in scope, could both expand the information on respondent encounters with the criminal justice system in the NLSY26 as well as provide administratively recorded alternatives to survey items in the NLSY26 for a subset of the sample.

Additional Resources

Rothbaum, Jonathan, et al. Fixing Errors in a Snap: Addressing Snap under-Reporting to Evaluate Poverty. U.S. Census Bureau, 2021.

Mittag, Nikolas. Correcting for Misreporting of Government Benefits. American Economic Journal: Economic Policy, 2019.

National Directory of New Hires (NDNH)

The NDNH is a central repository of employment data, unemployment insurance claimant data and quarterly wage data from state directories of new hires, state workforce agencies and federal agencies. The federal Office of Child Support Enforcement (OCSE) operates the NDNH, a database established by the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA) for the purposes of assisting state child support agencies in locating parents and enforcing child support orders.[2] In addition, Congress authorized specific state and federal agencies to receive information from the NDNH for authorized purposes. OCSE uses the NDNH primarily to assist states in administering programs that improve states' abilities to locate parents, establish paternity, and collect child support.

The NDNH is comprehensive and large. The NDNH includes information on (1) all newly hired and rehired employees, compiled from state reports (and reports from federal employers), (2) the quarterly wage reports of *existing employees* (in Unemployment Insurance [UI]-covered employment), and (3) UI applications and claims. HHS estimated that in fiscal year 2018, 793 million records were posted to the NDNH.

The NDNH is extensively linked to other data sources, though most of those linkages are operational. States match new-hire reports against their child support records to locate parents, to establish a child support wage-withholding order or to enforce an existing order. In addition to matching within a state, states transmit the new-hire reports to the NDNH.

Relevance

The NDNH contains the following files: 1) New Hire; 2) Quarterly Wage (QW); 3) Unemployment Insurance (UI). The New Hire file contains information on all newly hired employees reported by employers to each State Directory of New Hires (SDNH).[3] Employers are required to report the following seven data elements; however, many states require additional information, such as state of hire and date of birth.

- Employee name
- Employee Social Security number or Taxpayer Identification Number (TIN)
- Employee address
- Employer name
- Federal Employer Identification Number (FEIN); if a federal employer, then the Department of Defense code, if available, or employer identification number
- Employer address
- Date of hire

[2] The NDNH system itself is housed in the Social Security Administration's (SSA's) enterprise data infrastructure.

[3] Federal agencies report directly to the NDNH.

The QW file contains quarterly wage information on individual employees from -the state workforce agency (SWA) and federal agency records. When an individual is working more than one job during the reporting period, separate QW records are established for each job.

SWAs and federal agencies transmit the following QW data elements to the NDNH:

- Employee name (if collected by the state)
- Employee Social Security number or TIN
- Employee wage amount
- Reporting period (calendar quarter in which wages were paid)
- Employer name
- FEIN
- Employer address
- Employer optional address

The UI file contains unemployment insurance information on individuals who received or applied for unemployment benefits, as reported by SWAs. The states only submit claimant information that is already contained in the records of the state agency administering the UI program.

States transmit the following UI data elements to the NDNH:

- Claimant name
- Claimant Social Security number or TIN
- Claimant address
- Claimant benefit amount (gross amount before any deductions)
- Reporting period (calendar quarter in which the UI claim was filed)

The universe includes workers covered by unemployment insurance and federal workers. Data are released quarterly, and the lag time varies by the specific NDNH data set. The NDNH contains approximately 24 months (approximately 8 quarters) of data at any given time.

Self-employed workers, such as independent contractors, are generally not included in the NDNH. However, 16 states did require employers to report the income of independent contractors in some form.

With regard to non-wage employment, the NDNH generally does not gather information on self-employed workers (e.g., independent contractors), which are a growing proportion of the working adult population, according to various estimates. A 2019 study identified 16 states that require employers to report the income of independent contractors in some form (in many cases, to the SDNH).

Accuracy

The NDNH certainly meets the NLS timeliness requirements. The NDNH data are used by a range of other government agencies for eligibility determination and enforcement, where timeliness is crucial (e.g., child support enforcement). Because quick and accurate reporting of information on new hires in

NDNH is necessary to support its main purpose (e.g., locating a noncustodial parent who may have found temporary employment), there is probably no other ADS that provides information as timely as the NDNH.

In fact, there are mandatory timeframes for state and federal agencies to report data (new hire, quarterly wage, and unemployment insurance data) to the NDNH. The use of NDNH data by these agencies indicates that the NDNH data are both timely and of high quality.

Given the importance of the NDNH to a number of state and federal agencies, and the robustness of the database, the risk of discontinuation would appear to be very low.

Coherence

As a *national* database of wage and employment information, the NDNH should meet the NLS coverage requirements. There are some gaps in coverage—common to other administrative data sources of wage and income—such as independent contractors and unreported work.

Wage and unemployment insurance benefit information in the NDNH are reported on a quarterly basis. It is not clear whether these reports include a finer disaggregation (e.g., by month).

Feasibility

Federal law provides that a state or federal agency that receives NDNH information must reimburse OCSE for the costs of obtaining, verifying, maintaining, and comparing the information at rates that OCSE determines to be reasonable. OCSE uses a standard methodology to calculate fees based on three components: 1. Access (a fee that is split evenly among NDNH users) 2. Frequency of matches 3. User-specific costs related to performing the match. New NDNH users and new matches under existing users will be charged a onetime new user start-up fee to cover costs incurred by OCSE to set up a new agreement and perform the work required to implement a new match.

Because the NDNH plays such a critical role in child support enforcement and eligibility determination for a number of programs, it would seem that the risk of discontinuation is low. The main threat to discontinuation would come from increasing concerns over privacy.

Confidentiality, Consent, and Accessibility

The NDNH is subject to security and privacy requirements under Sections 453(l) and (m) of the Social Security Act. It also is considered to be a system of records under the Privacy Act, and thus is subject to the requirements under that act for administrative, technical, and physical safeguards for both the records matched and any results of those matches (see 5 U.S.C. 552a).

In addition to meeting the requirements of the Social Security Act (specified information to an authorized agency for an authorized purpose), an agency must meet other requirements governing the information comparison as outlined in this section. Requests for research information must originate from a federal, state, or local government agency.

Statutory authority is required to receive NDNH information. OCSE cannot disclose NDNH information if the law does not specifically authorize an agency to receive specified NDNH information and the information or comparison being listed requested does not meet the purposes stated in the statutory authority.

OCSE enters into an agreement (MOU or CMA) with each agency that receives NDNH information. The agreement describes the purpose, legal authority, justification, expected results of the match, description of the records, retention and disposition of information, reimbursement, and performance reporting requirements. Each agency is required to sign the security addendum, which is a component of the agreement. The security addendum provides a detailed description of the security requirements and safeguards that an agency must have in place before receiving NDNH information.

A larger concern with the NDNH is the data deletion (e.g., after two years) and deidentification requirements.[4] It is not clear whether these requirements would apply to BLS. Moreover, the strict privacy protections and other security requirements attached to the use of the NDNH may be administratively burdensome.

A one- or two-year time span provides a relatively limited window for observing earnings before, during, and after the time of program participation. Data of particular interest may have already been deleted before a research agreement can be reached. In the absence of identifiers, it is impossible for researchers to incorporate additional years or sources of administrative data into their research sample or correct problems with prior linkages once the de-identified file with NDNH data has been returned. While it is possible to construct a longitudinal research sample in the future, this requires greater involvement by OCSE (since only OCSE has access to the identifiers needed to continue updating the earnings data), increasing the cost and complexity of the project

Additional Resources

Overview of National Directory of New Hires

ACF Overview of National Directory of New Hires

Congressional Research Service National Directory of New Hires

A Guide to the National Directory of New Hires

National Directory of New Hires (NDNH) and State Directory of New Hires (SDNH) Guidance and Best Practices

National Directory of New Hires Guide

Compendium of Administrative Data Sources

[4] NDNH data that are used by HHS, ED, HUD, and USDA to conduct research or analyses of certain topics (as authorized under the Social Security Act) generally does not contain personal identifiers.

Social Security Administration (SSA) Data

Relevance

SSA data is comprised of several files that contain information on applicants, recipients, and beneficiaries of Supplemental Security Income (SSI) and Old Age, Survivor's, and Disability Insurance (OASDI), as well as individual earnings.

Master Earnings File. The Master Earnings File contains the individual lifetime records of wages and self-employment earnings. The file's primary sources of information are the W-2 form (for wages) and electronic files of form 1040, schedule SE (for self-employment income) from the Internal Revenue Service (IRS) in the Department of the Treasury. The most frequently used data elements are the individual's SSN, annual total wages (1978 to present), annual self-employment earnings, annual earnings used for OASDI contributions (1951 to present), and report year.

Master Beneficiary Record (MBR). The MBR is used to administer the OASDI program and contains beneficiary and payment history data. An MBR record is created whenever an individual applies for benefits and SSA adjudicates the application as an award, a denial, an abatement, or a withdrawal. Information maintained in the MBR includes the primary worker's SSN, the beneficiary's own SSN, benefit application date, benefit entitlement date, and type and amount of benefit.

Supplemental Security Record (SSR). The SSR contains information on individuals applying for SSI payments. SSA uses the income, resources, disabling condition, and living arrangement information from the application and other sources in determining eligibility for and administering the needs-based SSI program. SSR data elements include SSN, date of claim, citizenship status, income, resources, eligibility code, payment code, and payment amount.

NCHS survey data have been linked to five SSA Administrative Data Files: The Master Beneficiary Record (MBR) file, the Supplemental Security Record file (SSR), the Payment History Update System (PHUS) file, the 831 Disability Master File (831) and a special extract of summarized quarters of coverage (QOC) from the Master Earnings File. File layouts and data elements can be found here.

More than 56 million individuals received OASDI benefits in 2021 (on December 31); although OASDI receipt will likely not affect many NLS respondents until well into the panel. SSI receipt could be more relevant to respondents, but the program is relatively small in terms of participants (but not in terms of benefit payments). In February 2022, there were just over 5 million individuals receiving only SSI, roughly 4 million of which were under 65 and thus receiving SSI as a result of a disability. Another 2.5 million received SSI in conjunction with Social Security benefits, a number that was fairly evenly split between recipients above and below age 65.

Accuracy

The SSA data has been used in data linkage efforts both at Census and NCHS, suggesting that data quality is high.

In the case of the NCHS linkage, NCHS survey data was linked with SSA at the Social Security Administration (SSA) by a process in which individual NCHS survey respondents were matched with

SSA's Numident file.[5] NCHS provided SSA with as many of the following individual identifiers that were available on the survey record for all eligible survey respondents: SSN, first name, last name, middle initial, sex, father's surname (women only), state of birth, and zip code. NCHS survey participants were considered ineligible for matching to the Numident file if they refused to provide their SSN at the time of the interview. Additional ineligibility criteria included refused, missing, or incomplete information on last name or date of birth.

Match rates to linkage-eligible respondents in NCHS surveys has varied by age group and survey. In the National Health Interview Survey (NHIS), the match rate among the linkage-eligible respondents declined between 1994-2005, from 94 percent to 77 percent. Match rates were higher in each for respondents 65 and over. For the National Health and Nutrition Examination Survey (NHANES), match rates among linkage-eligible respondents remained constant over the same time period, at about 95 percent. More information on NCHS data linkage efforts can be found in this report.

The MBR, ME and SSR files are updated annually, with a lag time for release of one month for the MBR and SSR files (and 18 months for the ME file).

Coherence

Measurement of income is less granular than in, say, the UI wage data (i.e., annual vs. quarterly). However, the coverage of the SSA income data is national. The benefit receipt measures for SSI or OASDI would cohere with an annual receipt and benefit amount variable.

Feasibility

The SSA links their data with administrative data from a variety of government agencies and large surveys, including data from the Census Bureau (e.g., the Current Population Survey (CPS) and the Survey of Income and Program Participation (SIPP)), data from the Centers for Medicare and Medicaid Services (CMS), and the NCHS' National Health Interview Study. Data linked to the Census Bureau or the IRS are subject to additional restrictions.

Given the robustness and importance of the SSA administrative data platform, we believe that the risk of discontinuation for this ADS is low.

Confidentiality, Consent and Accessibility

As provided under the Privacy Act (5 U.S.C. § 552a), SSA is responsible for safeguarding the information maintained in its administrative files against an invasion of an individual's personal privacy. Other legal protections of the information SSA maintains or links to are provided by the Social Security Act and regulations, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA), Title 13 of the United States Code governing the Census Bureau, and the Internal Revenue Code covering earnings data that are considered to be tax return information (and would presumably fall under Title 26 restrictions).

[5] The Numident file is a numerically ordered master file for each Social Security number (SSN) ever issued and contains records for approximately 400 million SSNs, including personal identifying information.

The release of identified data outside of the SSA is restricted by legislation and policy. The SSA is responsible for protecting the information it maintains. SSA policy is to share identifiable data only with those *that have the legal authority to access the data and only if identifiable data are required to accomplish a specific research or statistical purpose*. Requestors must submit numerous documents that outline how they will keep the data secure, guarantee the data are not redisclosed, and restrict the data use for only the approved research or statistical purpose. Data linked to the Census Bureau, or the IRS are subject to additional restrictions.

SSA policy is to share identifiable data only with those having the legal authority to access data for a particular purpose, and only if identifiable data are required to accomplish a research or statistical purpose. The requestor must submit a proposal, a data protection plan, and confidentiality agreements. A Memorandum of Agreement must be approved by SSA's Office of the General Counsel. The user must guarantee to keep the data secure, not redisclose the data, and restrict the use of the data to the approved purpose. Access to SSA data that have been linked to Census Bureau data is subject to additional restrictions imposed by Title 13 of the U.S. Code, such as requiring users to obtain Special Sworn Status and permitting access only for Census-approved purposes and at a Census-approved site. Census Bureau procedures and regulations dictate how survey data can be used. SSA is not authorized to grant access to matched CPS or SIPP data. Additionally, the Internal Revenue Code provides its own restrictions, such as limiting access to earnings data to certain individuals and for certain purposes.

The privacy of all personal information SSA maintains is protected by a number of laws and regulations, including the Privacy Act of 1974, as amended; section 1106 of the Social Security Act, as amended; the E-Government Act of 2002, as amended; section 6103 of the Internal Revenue Code; related Social Security regulations and policies; and other federal statutes, rules, and regulations.

The Privacy Act and related legal authorities noted above allow SSA to disclose information from its program records to federal, state, and local agencies for certain "routine uses." These routine uses, defined in the Privacy Act at 5 U.S.C. 552a(a)(7), are permissive uses of information collected by SSA that, "with respect to the disclosure of a record, the use of such record for a purpose which is compatible with the purpose for which it was collected."

In addition to the Privacy Act, SSA also refers to The Computer Matching and Privacy Protection Act of 1988 (CMPPA) (and its amendments in 1990), 5 U.S.C 552a (a)(8)-(13), (3)(12), (o), (p), (q), (r), & (u), which establishes requirements that federal agencies must follow when matching information on individuals with information held by other federal, state or local agencies. The CMPPA, as interpreted by the Office of Management and Budget, also states certain guidelines for computer matches related to verification, notification, data accuracy, etc., to ensure that the federal government conducts computer matches uniformly and provides protections to the individual as provided under the Privacy Act.

In addition, matches covered under the CMPPA must meet certain stringent requirements. Generally, if a match will have an adverse effect on an individual or can reveal personally identifiable information, then certain provisions of the CMPPA will govern the content, format, processing, administration, and length of the life of the match. Certain administrative or enforcement actions that require specific information such as medical records or involve other confidential information may require the consent of the individual.

Additional notes:

- SSA will exclude from the extracts and will not disclose Federal Tax Information or other data that is restricted by law, contract, or MoU
- Earnings data are restricted and can be shared only in aggregate form in certain circumstances. The SSA follows guidelines set by the Internal Revenue Service (IRS) concerning earnings data.
- Additional information on accessing the Master Death File can be found here.
- For the Supplemental Security Income Record and Special Veterans Benefits (SSR) (System Number60-0103), SSA will disclose the individual-level Identifiable SSR data
- For the Master Beneficiary Record (MBR) (System Number 60-0090), SSA will disclose the individual-level identifiable MBR data which includes Payment History Update System data.
- The first linkage of NCHS survey respondents to their Social Security benefit history records and Medicare enrollment and claims records was initiated in 2001 *upon approval of the NCHS Research Ethics Review Board (ERB)*.
- The SSA Data Exchange Request Form can be found here.

Additional Resources

Uses of Administrative Data at the Social Security Administration

Social Security Data Page

Social Security Administration Agreement Types

Social Security Administration Data Exchange Request

Social Security Administration Data from J-PAL

Linkages between NCHS Surveys to Social Security Administration and CMS

Social Security Research, Statistics & Policy Analysis February 2022 Monthly Snapshot

Social Security Beneficiary Statistics (OASDI Benefits)

Use of Social Security Administration Data for Research Purposes

Compendium of Administrative Data Sources

UI Wage Records

Relevance

The content of UI wage records varies by state, but generally contain, at a minimum, employee PII, such as name and SSN, as well as the employee's quarterly wages, unemployment insurance benefits and the employer's name.

The LEHD Employment History File (EHF) contains the complete in-state work history for each individual that appears in the UI wage records. The EHF for each state contains one record for each employee-employer combination–a job–in that state in each year. Both annual and quarterly earnings variables are available in the EHF.

The Individual Characteristics File (ICF) for each state contains one record for every person who is ever employed in that state over the time period spanned by the state's unemployment insurance records.

Accuracy

The UI wage records will only capture UI-covered employment, which excludes agricultural workers, railroad workers, private household workers, student workers, the self-employed, and unpaid family workers. In New York State, for example, UI records cover about 97 percent of New York's nonfarm employment. However, some estimates of the coverage of UI records are more on the order of 90 percent of all jobs, and might even be lower for low-wage workers.

The presence of SSNs, and other PII, suggests that linkage error should be low. In general, linkages between the different files are created using deterministic match-merge techniques. Person, firm, and establishment identifiers allow users to link all LEHD Infrastructure files. LEHD reports that about 3 percent of the Protected Identification Keys (PIKs) in the UI wage records do not link to the PCF.

The LEHD undertakes major integrated quality control checks.[6] This post-processing, which includes QC and harmonization, means that the UI wage data maintained by the LEHD will be more research-ready than the raw data extracts from individual states. Though this likely comes at the cost of some timeliness.

Coherence

The quarterly wage and UI benefit measures would align well temporally with prior measures from NLS cohorts. Apart from the well-known gaps in coverage of certain wage earners (i.e., non-UI-covered workers)—which affect most other administrative wage data—the difficulty of achieving national coverage poses a major drawback of this data source.

Feasibility

UI wage records are state-owned data, and as such individual data sharing agreements would need to be negotiated with each state (see a sample MOU for New York State). The Census Bureau's LEHD project is a central repository of sorts for UI wage records, having collected UI wage data from all states except Kansas (in negotiation), Alaska, Arkansas, and Mississippi. But it is not clear whether an agreement with LEHD would be possible; even it if it were, it would still require sign-off in most, if not all, cases from the individual states. For individual researchers, access to LEHD data will only be granted to qualified researchers on approved projects with authorization to use specific data sets, and access to the restricted-use data must occur at an FSRDC.

[6] Abowd et al. (2005) note, for example, that sometimes "In particular, the state's archival historical UI wage record and ES-202 data are sometimes permanently damaged or defective."

Given the need to engage with individual states, the risk of discontinuation is moderate to high. Turnover in agency staff or leadership could lead to shifting priorities or capacity to continue to provide data over a long period of time.

Confidentiality, Consent, and Accessibility

State UI wage records are covered by state legislation. In New York, the relevant regulations around data sharing are the UI Data Sharing Bill (S5773A) and the New York State Labor Law §537Data Quality. Most states require that shared data is stored in a highly secure environment, and some states require evidence of informed consent forms and in some cases require researchers to collect signed state waivers that authorize UI wage and benefits to be released. Other restrictions and safeguards that attach to UI wage data varies by state; these are typically specified in each state's memorandum of understanding (MOU). For example, in New York, all individuals with access to confidential data must sign a nondisclosure agreement annually and take part in the UI Confidentiality Training developed by NYSDOL.

Additional Resources

Technical Assistance Guide for Using Unemployment Insurance Wage Data

Quarterly Census of Employment and Wages

Compendium of Administrative Data Sources

Criminal Justice Administrative Records System (CJARS)

CJARS is a data infrastructure initiative at the University of Michigan that seeks to improve public administration of the U.S. criminal justice system through data-driven research and statistical reporting to practitioners.

CJARS is the first integrated national research data repository that follows individual offenses from arrest to charge to conviction to sanction. Data come from multiple stages of the justice system and from a wide range of jurisdictions. In cooperation with the U.S. Census Bureau, the collected records are linked to confidential social, economic, and demographic data held by the federal government to further enhance the value of CJARS.

Relevance

CJARS is a longitudinal, multi-jurisdictional data source that is harmonized and linked to track individuals over time and jurisdictions. It is "built for integration with socio-economic survey and administrative data held by the Census Bureau." CJARS does not capture minors involved in the juvenile justice system, reported crimes that do not result in an arrest, or other events that might be tangential to the criminal justice system, such as child welfare investigations.

The CJARS data is comprised of six separate databases. The six databases include a roster and one database for each of the five types of events that are covered (Entry; Legal Proceedings; Institutional corrections; Community Corrections).

CJARS covers more than 20 states (including over 33 million unique individuals and over 78 million criminal justice events).

Variables include:

Roster: roster table contains the unique list of person-level CJARS IDs.

Arrest and booking: The arrest table contains information regarding the arrest and booking date, as well as the offense that led to the arrest.

Adjudication: The adjudication table contains detailed information about the offense the person was charged with, disposition information, and sentencing.

Probation: The probation table contains information on probation conditions, probation begin status and date, and probation end status and date.

Incarceration: The incarceration table contains information about the facility an individual is/was housed, entry and exit dates, as well as the current status of the person.

Parole: The parole table contains information on parole begin/end dates and exit status when available.

Accuracy

At the Census Bureau and in the FSRDCs, CJARS data may be linked at the person-level to other socioeconomic survey and administrative records using an anonymous identifier called a Protected Identification Key (PIK). Staff at the Census Bureau attempt to use all available PII to assign a PIK. However, no biometric identifiers are transferred to the Census Bureau.[7]

CJARS has developed an algorithm that probabilistically matches records to individuals when no unique identifiers available by using names and dates of birth to identify individuals. Once an individual has been identified, they receive an anonymized individual identifier (cjars_id). CJARS has also developed a method of probabilistically matching criminal justice events to an episode. This linkage is created so that researchers can trace every event associated with a single criminal justice episode. Once the PIK assignment process has occurred, sensitive PII is removed from the research files and the anonymized files with PIKs attached are transferred to a secure computing environment that is available at the Census Bureau headquarters and in the FSRDCs. There are some discordances between cjars_ids and PIKS.

As is typical of state-level data, there is substantial post-processing of the individual state agency data to standardize and edit administrative records from more than 20 states. CJARS has developed a national data schema that attempts to strike a balance between capturing local complexity and differences while still achieving consistency across jurisdictions. To the extent possible, CJARS preserves the source values (e.g., offense descriptions, sentencing fields) but also provides the harmonized variables created from

[7] CJARS employed machine learning models on fingerprint-validated identifiers.

the source values. CJARS employs matching learning-augmented harmonization strategies. CJARS has also developed a Text-based Offense Classification (TOC) Tool, which translates over 4 million unique offense descriptions to 271 standardized offense codes.

CJARS has been validated against official federal statistical series.

Coherence

The various NLS cohorts varied significantly in how they cover criminal justice events. With data available from about 20 states, coverage is an issue in CJARS. However, if several large states can be matched to the NLS, it could expand the use of the NLS for research in this area, given that criminal justice events may not be accurately self-reported. For example,

- Labor market outcomes after a criminal justification
- Neighborhood environment and criminal justice involvement
- Criminal justice contact as an outcome for a non-criminal justice intervention

Feasibility

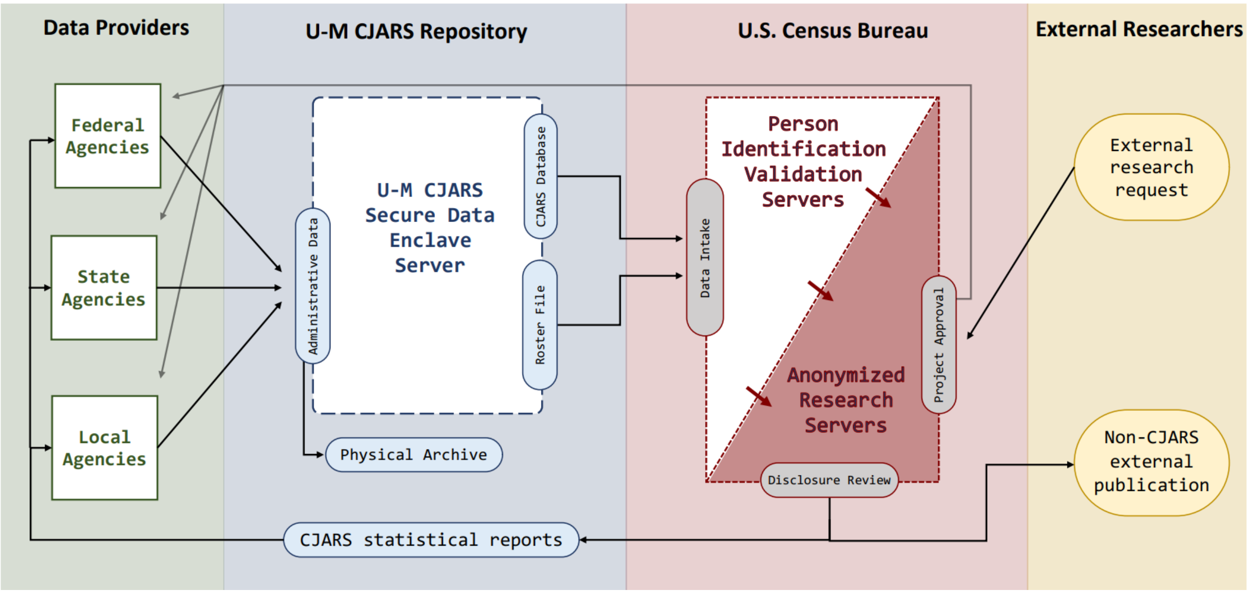
At the Census Bureau and in the FSRDCs, CJARS data may be linked at the person-level to other socioeconomic survey and administrative records using an anonymous identifier called a Protected Identification Key (PIK).

Confidentiality, Consent, and Accessibility

We (CJARS) are solely authorized to distribute CJARS data through the U.S. Census Bureau's Federal Statistical Research Data Center (FSRDC) network to qualified researchers on approved projects. FSRDCs are Census Bureau facilities, housed in partner institutions, that meet all physical and information security requirements for access to restricted-use microdata of participating agencies.

Data requests require a research proposal submitted to the U.S. Census Bureau. All proposals require approval before data is accessible for research. CJARS has developed a proposal guide to assist researchers interested in requesting CJARS data.

The following diagram illustrates the interaction of federal, state, and local data providers; the University of Michigan, the Census Bureau, and external researchers (taken from the following webinar).



Data are collected, cleaned, and harmonized at the University of Michigan and then integrated into U.S. Census Bureau data systems and made anonymous and available through the Federal Statistical Research Data Center (FSRDC) network. Qualified researchers can use the standard Census Bureau FSRDC proposal process to request use of the restricted-access CJARS data. **The data cannot be requested directly from the University of Michigan.**

At the Census Bureau, data are protected by 13 USC §9a. Only those individuals working on record linkage have access to the PII. CJARS recommends that research users should consult with their respective institutional review boards to determine if they must apply for approval to cover specific research projects using the CJARS data.

Additional Resources

Criminal Justice Administrative Records System

CJARS Data Access

CJARS Data Documentation Paper

CJARS Data Holdings

CJARS Data Schema

CJARS contact email: cjars-data-users@umich.edu

Housing and Urban Development (HUD) Data

Relevance

HUD administrative data systems contain housing, income, and program participation data for recipients of Housing Choice Voucher (HCV), Public Housing (PH), and privately-owned, subsidized Multi-Family Housing (MF) programs for all states, the District of Columbia, and some territories (e.g., Puerto Rico and the U.S. Virgin Islands). The data collected through the administration of HUD's housing assistance programs are stored in two information management systems, the Public & Indian Housing Information Center (PIC) and the Tenant Rental Assistance Certification System (TRACS).[8]

Housing Choice Vouchers (HCV). HCV is the federal government's largest housing assistance program. The HCV program also includes the homeownership voucher, project-based voucher, Section 8 Moderate Rehabilitation, and Section 8 Rental Certificate programs. The HCV program accounts for just over half of the NHIS and NHANES participants that linked to HUD.

Public Housing (PH) program was established to provide safe rental housing for eligible low-income families, the elderly, and people with disabilities. HUD provides capital subsidies and operating subsidies to local Public Housing Agencies that manage public housing for eligible low-income residents.

Multifamily (MF) programs – The MF program category encompasses a number of separate, distinct HUD programs, including Project-Based Section 8 Voucher Assistance in Multifamily Housing (the largest MF program), Section 221(d)(3) Below Market Interest Rate, Section 236 Multifamily Housing, Rental Assistance, Section 202 Supportive Housing for the Elderly Program, Section 202/162—Project Assistance Contract, Section 811 Supportive Housing for Person with Disabilities, and Rent Supplement.

Because each of the remaining MF programs lacked sufficient sample size on an individual basis in the linked file, they were combined into a single MF program category. In all MF programs, subsidies are paid directly to private property owners who provide a certain percentage of their housing units at affordable rates for low-income persons who qualify. MF program assistance is tied to the property, unlike tenant-based rental assistance programs (e.g., HCVs), and tenants cannot take their rental housing assistance subsidy elsewhere. Overall, among NHIS and NHANES participants that linked to HUD administrative data, slightly less than half were participating in a MF program.

These programs, the data systems in which they are stored (along with the federal forms that are the source of these data), the universe, and linking variables available are summarized in the table below:

HUD Program	Data System - Form	Universe	Linking Variables
Housing Choice Vouchers (HCV) Program	PIC (IMS/PIC) – Form HUD-50058	All persons (and households) in HUD's HCV programs	SSN, name, DoB, address
Public Housing (PH) Program	PIC (IMS/PIC) – Form HUD-50058	All persons (and households) in HUD's PH programs.	SSN, name, DoB, address

[8] Currently, the PIC system is evolving into the PIH Inventory Management System (IMS). During the transition, the system is being referred to as IMS/PIC. The system facilitates more timely and accurate exchanges of data between PHAs and HUD.

Multi-Family (MF) programs	TRACS – Form HUD 50059	All	SSN, name, DoB, address

Most of the data collected by Public Housing Authorities (PHA) on households and listed in forms HUD-50058, Family Report, and MTW-50058, Family Report, are accessible to researchers[9], and would presumably be available to BLS. The HUD-50058 form includes information on household PII for data matching, demographics, and composition, sources and amounts of income for each person in the household, information about the subsidized unit, and housing subsidy information, such as the amount of housing subsidy and the amount the household pays toward rent. Additionally, information on the physical public housing stock is maintained, including address, building type, number of bedrooms, and unit occupancy status, to ensure appropriate operating subsidy and capital improvement funding.

The Tenant Rental Assistance Certification System (TRACS) is a system to collect and maintain certified tenant data from owners and management agents of MF housing programs. The TRACS data extract created for the NCHS-HUD data linkage was based on TRACS point-in-time quarterly extracts from the TRACS production system. These data capture transactions occurring within the 18 months immediately prior to the date of extract. Most HUD recipients are required to recertify each year, and consequently, a transaction is expected each year. However, some HUD programs (for instance, the Moving to Work (MTW) Demonstration Program) have longer intervals between recertification. Contained in this file is race/ethnicity, address, SSN, transaction date and type, total income, subsidy type, total tenant payment, and gross rent.[10]

Based on these transaction files, the NCHS-HUD data linkage also provides episode files that contain the start and end dates for participation episodes in the various HUD programs to permit longitudinal (or spell) analysis. The term “episode” refers to a single continuous period of enrollment in a HUD program based on dates of HUD transactions. The begin date of a participant’s first episode is the effective date on their first transaction record. Subsequent episodes for the participant are identified based on the interval between the effective dates on their transaction records. For more information on these files, see the NCHS-HUD Linkage Methodology and Analytic Considerations documentation.

The geographic coverage is national (including external territories that have funded housing, such as Puerto Rico). The universe covers all public housing and housing choice voucher (HCV) recipients. Data are collected daily from the housing agencies. Extracts are uploaded to the HUD data website and HUD User Portal on a quarterly basis. The lag time depends on the specific HUD data source, and ranges from one week to one quarter.

NCHS has linked 1999-2018 National Health Interview Survey (NHIS) and 1999-2018 National Health and Nutrition Examination Survey (NHANES) to administrative data through 2019 for the Department of Housing and Urban Development’s (HUD) largest housing assistance programs: the Housing Choice Voucher (HCV) program, public housing (PH), and privately owned, subsidized multifamily housing (MF). NCHS survey data have been linked to HUD administrative data containing information on the timing of

[9] HUD accepts data license applications and restricted-use data requests to researchers, provided that their “research aligns with HUD priorities.” More information is available on the HUD USER website here and here.

[10] A transaction refers to any activity for which a HUD form was completed (e.g., new admission to a HUD program, annual recertification, end of participation, etc.). These files are released four times a year.

receipt of housing assistance, type of housing assistance received, the structure of the housing unit, as well as other household characteristics. The NCHS-HUD linked data are comprised of a transaction file, a temporal alignment file, and an episodes file.[11] [12]

The HUD rental assistance programs are relatively large. In January 2022, the Center for Budget and Policy Priorities (CBPP) reported that there were 10.2 million individuals, in 5.2 million households, receiving rental assistance in the US.

Accuracy

HUD data contains detailed PII, including SSNs, for data linkage, suggesting that linkage error should be low. For example, records from NCHS participants were linked to the HUD enrollment database using the following identifiers: SSN, first name, last name, middle initial, month of birth, day of birth, year of birth, 5-digit ZIP code of residence, state of residence, and sex.

Coherence

HUD data could provide information on the timing of housing assistance receipt and the type and amount of housing assistance.

PIC contains household-level and person-level administrative records on persons and households participating in HUD's HCV and PH program types. The PIC data extract created for the NCHS-HUD data linkage was based on HUD's PIC point-in-time quarterly files, which capture a household's most recent transaction with HUD during the prior 18 months, though it would seem that any transactions over 18-month period could be made available.

Feasibility

That HUD is sharing data with both Census and NCHS bodes well for integrating HUD data in the NLS program. An NLS MOU with HUD could be modeled on the data sharing agreements already in place with Census and NCHS, potentially expediting the data acquisition process.

Risk of discontinuation is low.

Confidentiality, Consent and Accessibility

The linkage of NCHS-HUD data was conducted through a designated agent agreement between NCHS and HUD. Approval for the linkage was provided by NCHS' Research Ethics Review Board (ERB). The data linkage work was performed at NCHS. Only a subset of 1999-2018 NHIS and 1999-2018 NHANES participants were eligible for linkage with the HUD administrative data. NCHS survey participants who

[11] There is also a weight file.

[12] The episode files, created by NCHS, contain start and end dates for participation episodes in various HUD programs based on the transaction data and assumptions about reasonable intervals between transactions. Most HUD recipients are required to recertify each year, and consequently, a transaction is expected each year. However, some HUD programs (for instance, the Moving to Work (MTW) Demonstration Program) have longer intervals between recertification. The episode files are useful primarily for longitudinal analysis related to the duration and timing of housing assistance episodes, and conditions or outcomes that may have preceded or followed such episodes.

have provided consent as well as the necessary personally identifiable information (PII), such as name and date of birth, are considered linkage eligible. Linkage eligibility refers to the potential ability to link data from an NCHS survey participant to administrative data. Criteria for NCHS-HUD linkage eligibility vary by survey and year due to variability of questions across NCHS surveys, changes to PII collection procedures by the surveys and changes in which survey participants are asked specific questions over time. Non-aggregated tenant and physical inventory may contain PII and requires approval from the Privacy Officer.

NCHS survey participants under 18 years of age at the time of the survey are considered linkage eligible if the linkage eligibility criteria described above are met and consent is provided by their parent or guardian. However, the consent provided by the parent or guardian does not apply once the child survey participant becomes a legal adult and there is no opportunity for NCHS to obtain consent to link the child participant's survey data to administrative data based on their adult experiences. As a result, in accordance with NCHS ERB guidance, NCHS only includes administrative data that were generated for program participation, claims and other events that occurred prior to the survey participant's 18th birthday on the linked data files provided to researchers.

NLS would need to obtain consent from respondents again as they enter adulthood.

Due to confidentiality requirements, the Restricted-use NCHS-HUD Data are accessible only through the NCHS Research Data Center (RDC). All interested researchers must submit a research proposal to the RDC.

Additional Resources

HUD Programs and Associated Administrative Data

Linking NCHS Survey Data to HUD Administrative Data: Linkage Methodology

HUD Data Sets

Multifamily Assistance & Section 8 Contracts Data Dictionary

Center on Budget and Policy Priorities Federal Rental Assistance Fact Sheets

Federal Policy for the Protection of Human Subjects ('Common Rule')

Compendium of Administrative Data Sources

HUD contact email: helpdesk@huduser.gov

Medicaid and CHIP Data

Like SNAP, WIC and TANF data, Medicaid data is provided by individual states; however, unlike those programs, a central repository has been created to which states report their Medicaid data.

Transformed Medicaid Statistical Information System (T-MSIS)

T-MSIS data is the most current and complete Medicaid and CHIP data resource available. CMS claims that the data released through T-MSIS "provides timely and accurate information on utilization and spending under Medicaid and CHIP and enable research and analysis to improve quality of care, assess beneficiary care costs and enrollment, improve program integrity and monitor performance." The data is utilized by researchers, entrepreneurs, Congress, oversight agencies and others.

CMS has worked with states to implement the updated version of the Medicaid Statistical Information System (MSIS). T-MSIS provides comprehensive information on state Medicaid programs, such as who is eligible to receive services, the services provided (e.g., ambulatory, long-term care, waiver services, and drugs) and to whom they are provided, the outcomes of care, and how much the care costs. This data is available for all states, beneficiary groups, and payment systems.

T-MSIS builds on the person-level and claims-level data previously available under MSIS to improve timeliness, reliability, and completeness of national Medicaid and CHIP data. The T-MSIS Analytic File (TAF) is made up of several component files: 1) The annual demographic and eligibility (DE) file; the 2) the Claims Files; 3) the Annual Managed Care Plan (APL) file; the Annual Provider (APR) file. Links to the documentation to each of these files can be found here.

In brief, the DE file contains information on the demographic, eligibility, and enrollment characteristics of beneficiaries who were enrolled in Medicaid or CHIP for at least *one day* in a given calendar year. The Claims files are comprised of (1) the IP file (institutional inpatient services and payments); the (2) the LT file (institutional long-term care services and payments); (3) the OT file ("Other" medical services and payments); and (4) the RX (prescription drug fills and pharmacy payments). The APL file contains detailed information about each managed care entity authorized to enroll Medicaid and CHIP beneficiaries (e.g., the plan's characteristics, type of reimbursement arrangement, and the location and type of beneficiaries the plan is allowed to enroll). The APR file captures information about all providers authorized by a state to render services to Medicaid and CHIP beneficiaries at any point in the calendar year.

Additionally, T-MSIS is designed to capture significantly more data and information. It includes additional variables and expands reporting options for many existing variables. A summary of the new data elements in T-MSIS, can be found in this brief.

All states are now submitting T-MSIS data. CMS takes each state's raw T-MSIS data and standardizes them into a research ready data set known as the T-MSIS Analytic Files (TAF). The TAF is further refined to remove certain personally identifiable information and proprietary information on managed care payment amounts to providers before the data are publicly released as the TAF research identifiable file (RIF). In addition, CMS has released updated versions of earlier TAF RIF files as states have addressed certain data quality issues.

CMS collects and stores the T-MSIS data in a relational database format that is updated by states *on a monthly basis*. When states update information in T-MSIS, both the old and new records are retained, requiring users to develop logic to identify and use only the most recently submitted information. Some variables undergo data cleaning and standardization to make them easier to use. Values are standardized across states. The following table summarizes the data cleaning and standardization undertaken in the T-MSIS analytical files (TAF) (the source of this table can be found here):

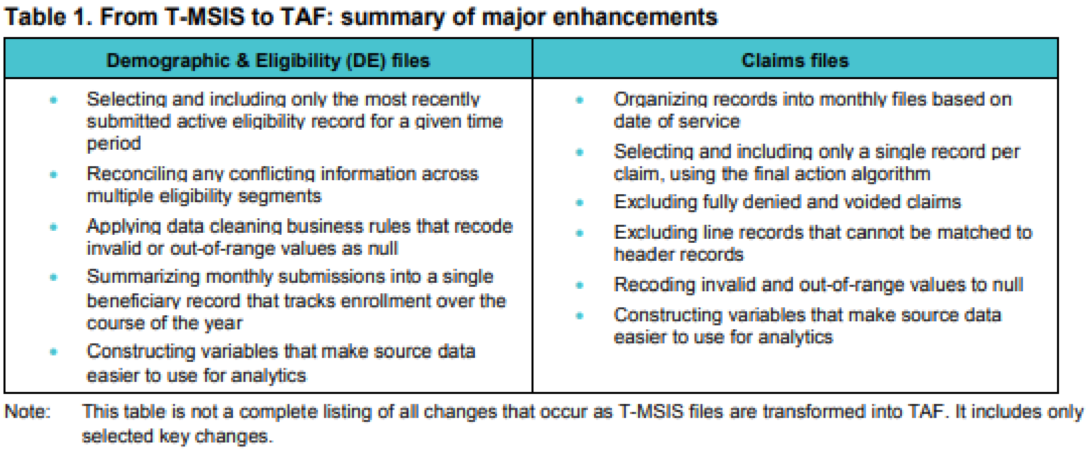


Table 1. From T-MSIS to TAF: summary of major enhancements

Demographic & Eligibility (DE) files	Claims files
• Selecting and including only the most recently submitted active eligibility record for a given time period • Reconciling any conflicting information across multiple eligibility segments • Applying data cleaning business rules that recode invalid or out-of-range values as null • Summarizing monthly submissions into a single beneficiary record that tracks enrollment over the course of the year • Constructing variables that make source data easier to use for analytics	• Organizing records into monthly files based on date of service • Selecting and including only a single record per claim, using the final action algorithm • Excluding fully denied and voided claims • Excluding line records that cannot be matched to header records • Recoding invalid and out-of-range values to null • Constructing variables that make source data easier to use for analytics

Note: This table is not a complete listing of all changes that occur as T-MSIS files are transformed into TAF. It includes only selected key changes.

Relevance

The NLSY97 asked about specific types of insurance coverage in the first survey round, but subsequently asked a general insurance coverage question that did not distinguish the types of insurance. The NLS-CYA tracks the Medicaid coverage of respondents more closely, including the reason for losing Medicaid coverage.

TAF RIFs contain beneficiary-level data that are available with an approved Data Use Agreement. These research files are used by CMS, oversight entities and researchers to answer key questions about the Medicaid and CHIP programs. The TAF RIFs include annual files that contain demographic and eligibility information for all Medicaid and CHIP eligible beneficiaries as well as claims files that contain service use and payment records. While CMS will not share with individual researcher's beneficiary names, addresses, or proprietary managed care payment information, it is not clear that the same would apply under an MOU with BLS. The data linkage effort with NCHS, as well as other state and federal agencies, suggests that PII would be made available.

A total of five file types are being released for calendar years 2014, 2015 and 2016 at this time:

- Annual Demographics and Eligibility (DE) File
- Inpatient Hospital (IP) Claims File
- Long-Term Care (LT) Claims File
- Pharmacy (RX) Claims File
- Other Services (OT) Claims File

Timeliness of the analytic files may be an issue, though it seems progress is being made in improving timeliness. In November 2019, CMS released TAF RIFs for calendar years 2014, 2015, and 2016. The fall 2020 release included data from calendar years 2017 and 2018, which is a significant step forward in the timeliness of available data. In addition to providing access to more timely data, the overall quality of the data has improved notably compared to the initial data released last November.

Medicaid is a very large program, especially in light of the Medicaid expansion in many states during the past decade. In November 2021, 78 million individuals were enrolled in Medicaid. CHIP is a much smaller program with just under 7 million individuals enrolled in November 2021.

Accuracy

T-MSIS is the most current and complete source of Medicaid and CHIP data. Only 3 U.S. territories do not submit T-MSIS data. States must meet “data quality” targets for data elements in three content categories: those classified as (1) critical priority, (2) high priority, and (3) expenditure. Twenty-two states (and Puerto Rico) met the data quality targets for all three data content categories; 16 states met the critical priority criterion but did not meet at least one of the targets for the high priority and/or expenditures data content category; 13 states and the Virgin Islands did not meet the target for critical priority criterion.

Over the last several years, CMS has been working with each state to improve the accuracy and completeness of its T-MSIS submissions. CMS has identified 32 T-MSIS Priority Items (TPIs) related to T-MSIS data quality and states have made significant progress in addressing these items. Information on the number of open TPIs per state for the first 23 TPIs can be found in the state maps here. Moving forward, CMS will continue to work with the states to improve the quality of their data.

One example of such an improvement is with the Medicaid and CHIP data. CMS is in the process of redesigning its system for the collection and management of Medicaid and CHIP data. Once fully implemented, the Transformed Medicaid Statistical Information System (T-MSIS) will provide the research community with a complete, accurate, and timely national database of detailed Medicaid and CHIP information. This new format may aid in future data linkage projects.

CMS has produced a series of 35 TAF DQ briefs that assess and summarize at a high level the reliability, accuracy, and usability of 2016 TAF data. The DQ snapshots provide topical and state-specific views of these data quality assessments. This summary DQ information is available for calendar year 2016 only, the most recent year of data in this release.

So, while there have been some data quality issues, it appears that data quality, if not already high, will be high by the time the NLSY26 launches.

There are four types of identifiers in the T-MSIS data: Medicare Health Insurance Numbers; Medicare Beneficiary Identifiers; Social Security Numbers, and T-MSIS Identification Numbers. The presence of SSNs suggests that linkage errors would be low if SSNs are available for linking in the survey data.

Under an interagency agreement between NCHS and the Centers for Medicare and Medicaid Services (CMS), data from several NCHS surveys have been linked to Medicaid enrollment and claims data. The resulting linked data files provide the opportunity to examine the administrative data during the year the survey was conducted, in years following the survey, as well as the years prior to the survey for some NCHS survey participants. The linked NCHS-Medicaid files, in particular, combine health and socio-demographic information from the surveys with enrollment and claims information from the Medicaid and Children’s Health Insurance (CHIP) programs.[13] However, the NCHS linkage appears to involve the dated Medicaid Analytic eXtract (MAX) data, which was last released in 2015.

[13] This linkage was to the Medicaid Analytic eXtract (MAX) files, research-ready calendar year person-level data files on eligibility, service utilization and payment information. The last year of MAX files available are in 2015.

Feasibility

The more centralized and standardized nature of Medicaid data collection, as well as NCHS's track record of data linkage with Medicaid data, suggest that this is a more viable alternative data source than other state-based data sources, such as SNAP, WIC, and TANF. CMS's efforts to consolidate state Medicaid and CHIP data does not have an analogue on the SNAP/WIC/TANF side. This leadership at the federal level points to a lower risk of discontinuation.

Accessibility, Consent and Confidentiality

CMS does make identifiable data files (IDFs) available to "certain stakeholders" as allowed by federal laws and regulations as well as CMS policy. IDFs contain protected health information (PHI) and/or personally identifiable information (PII) and so these would presumably fall under HIPAA protections. The process to request IDFs, add new IDFs to an existing request, or to re-use IDFs for a different project depends on the type of organization and the purpose for requesting the data. Researchers (including federal agencies) should contact ResDac.

Section 1902(a)(7) of the Social Security Act allows state Medicaid agencies to share information with other agencies only if it is directly related to administration of the state Medicaid plan. As implemented at 42 CFR 431.302, these purposes include establishing eligibility, determining the amount of medical assistance, and providing services for beneficiaries. Medicaid agencies wishing to exchange information with other agencies must execute a data exchange agreement restricting and safeguarding the types of information that can be released. **When releasing information to another agency, access to Medicaid information about applicants or beneficiaries must be restricted to persons or agency representatives who are subject to standards of confidentiality comparable to those of the Medicaid agency.** In addition, when Medicaid agencies agree to share data the agency must obtain consent from the individual before his/her data are shared. Whether BLS could obtain just top-level data on participation and costs, omitted any medical information on recipients, might facilitate data access, as some of the privacy protections attaching to medical information would no longer apply. But we were unable to confirm this without reaching out directly to CMS.

There are three exceptions to the requirement to obtain consent from the individual. This include cases in which the information will be used: 1) to verify income; 2) eligibility; or 3) the amount of medical assistance provided. If the information will be provided in an emergency situation, which does not permit obtaining consent before release of the information, the state must notify the family or individual immediately after release of the information.

The Medicaid agency must obtain consent before release of an individual's data, unless it is to verify income, eligibility, or the amount of medical assistance, under 42 CFR 431.306(d). This consent must come from the individual, a parent or guardian of the individual, or an authorized representative of the individual whose data would be exchanged. Determining who has the authority to consent to treatment and associated release of medical information when the title IV-E agency has placement and care responsibility for a child, varies by state.

For the title IV-E and Medicaid eligibility system exchange, the data exchanged must support the goals of serving clients and improving outcomes by sharing data required for purposes such as reporting, program administration, Medicaid eligibility determinations, and audits. CMS and ACF may review

Toolkit: Data Sharing for Medicaid and Child Welfare Agencies 15 the proposed data elements included in the bi-directional data exchange through the Advance Planning Document (APD) process.

Interagency Data Sharing Agreements. Both ACF and CMS strongly encourage title IV-E and Medicaid agencies to enter into interagency data sharing agreements to implement automated bi-directional data exchanges between the agencies' information systems. To establish a data exchange, leadership from both agencies should develop and enter into written agreements, for example a Memorandum of Understanding (MOU) or an Interagency Agreement (IAA). These agreements will encompass a data governance plan that will clarify each agency's responsibility about the sharing and use of case and child data, consistent with federal and state confidentiality provisions. These agreements should contain explicit rules governing consent, the intended benefit of the exchange, and which elements should be exchanged and when.

Regardless of the approach used, agencies should keep in mind that as laws, policies, and practices change, the data exchange, and any agreements related to the exchange, should be updated to reflect those changes. Agreements should, therefore, be made as 'living' documents that can be updated. To support the potential for change, each agency should document key personnel from IT and program offices responsible for updating the agreements. An Interagency Data Sharing Agreement is created through collaboration between two or more agencies. The document describes the terms of an agreement by defining certain data sharing terms, such as who, what, where, when, why and how the data shall be exchanged, and for how long the data should be retained. In this way, the agreement should address the business rationale for why the exchange is needed and the process to maintain the exchange. The rationale included in the exchange should support the business needs and administrative processes of both agencies and their common clients.

Additional Resources

Background on Medicaid and Children's Health Insurance Program (J-PAL)

CMS Virtual Research Data Center FAQs

Leverage Administration for Children and Families Administrative Data

Data Sharing for Child Welfare Agencies and Medicaid

Status of State Efforts to Integrate Health and Human Services Systems and Data

Medicaid and CHIP T-MSIS Analytic files Data Release

CMS Press Release Providing Medicaid and CHIP T-MSIS Data

Medicaid & CHIP Enrollment Data

T-MSIS Data Dictionary

Medicaid Analytic eXtract (MAX) General Information

Medicaid T-MSIS Analytic Files

CMS Contacts Database

Medicaid/CHIP contact email: DataConnectSupport@cms.hhs.gov

National Student Clearinghouse

Relevance

The National Student Clearinghouse (NSC) contains student-level data on nearly all enrollments at post-secondary, title IV, degree-granting institutions in the US. Institutions that do not participate with the Clearinghouse include most of the US military academies, most tribal colleges, and many very small institutions. International students and undocumented students (non-U.S. citizens) are also often not reported to the Clearinghouse, even when they are enrolled at participating U.S. institutions. In those cases when they are reported (estimated to be less than half of the time), these students are also more difficult to track if they change institutions.

Risk of discontinuation appears to be rather low. As mentioned in Dundar and Shapiro (2016), NSC uses the analytic power of the data to benefit institutions while consistently meeting data security and privacy standards. For this reason, despite the participation being voluntary, it is extremely rare for schools to discontinue their participation with NSC once they join.

Accuracy

The coverage is a near-census of students enrolled at post-secondary education institutions in the US. As Dynarski et al. (2013) note, participation in the NSC is voluntary, and although participation is very high, it is not complete. However, more recently, the NSC reports that over 3,600 colleges and universities — enrolling over 97% of all students in public and private U.S. institutions — regularly provide enrollment and graduation data to the Clearinghouse. Though their study is somewhat dated, they noted that at the time, the NSC did not cover postsecondary enrollment among black (and to a lesser extent Hispanic) students as it did among white students. A detailed analysis of the NSC's coverage of enrollments at all post-secondary, Title IV, and degree-granting institutions by state, sector, and level of institution (based on historical IPEDs institutional characteristics, can be found here.

Dynarski et al. (2013) report that the NSC matching algorithm relies primarily on name and date of birth to match students to their post-secondary records, and also tends to err on the side of false negatives (i.e., declaring a non-match when in fact the records matched). Dynarski et al. (2013) evaluate the NSC matching algorithm in a state-level case-study for Michigan.

However, more recent information indicates that in addition to name and date of birth, SSN is available for linking. The NSC StudentTracker Detail Report as well as the Poverty Action Lab indicate that the SSN is available, though it is not entirely clear if this is a required data element for compliance reporting as detailed here.

Another factor that leads to some under-coverage in the NSC data is that under FERPA, a federal law that protects the privacy of student education records that applies to all schools that receive funds from the U.S. Department of Education, both students and schools can block NSC from releasing their enrollment and degree information. These “FERPA-blocks” come overwhelming from students, rather than institutions.

The NSC data on degrees awarded to students, student demographics, and additional student-level data are updated frequently. Data are typically updated 45 days after the start and finish of school semesters. The Poverty Action Lab reports that institutions typically submit data just after the beginning of the semester, after the add/drop deadline, and at the end of each semester. But schools are given 45 days to submit data. Fall enrollment data are typically available in late November, fall completion data are available in late February, spring enrollment data in early March, and spring completion data by the end of June.

Coherence

Some of the relevant variables included in the NSC data include school code, enrollment status, dates of attendance, institutions attended, anticipated graduation date, major course of study, degree, certificate, or credential title, student demographics, veteran’s status indicator, Pell Grant recipient flag, National Center for Education Statistics (NCES) Classification of Institutional Programs (CIP) code for major 1 and 2, high school name, disability code, Student Individual Taxpayer Identification Number (ITIN). These align well most of the school history variables in the NLSY97.

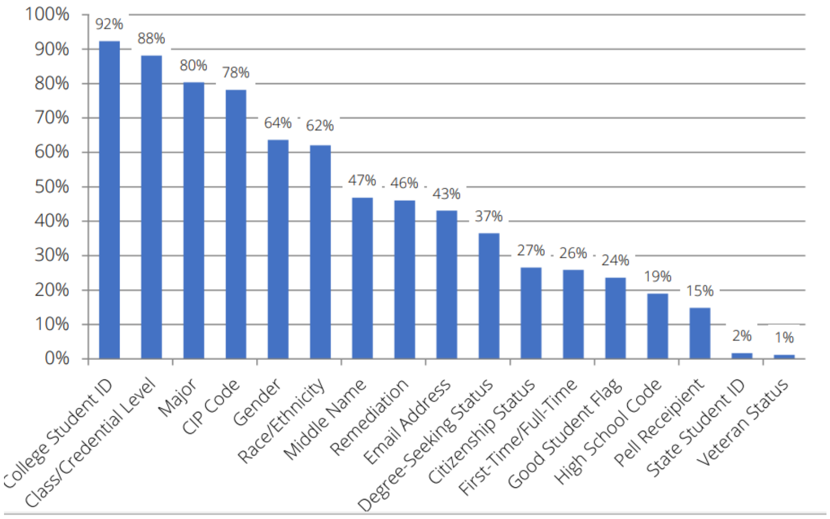
As part of an optional NSC service (DegreeVerifySM), institutions can also send to NSC detailed information on degrees awarded, including the degree type, level, and major, for each student (for a complete list see Appendix B). These data elements are currently provided for approximately 90 percent of all students in the data.

The data held by NSC include both mandatory and optional data elements. The following data are necessary for basic compliance reporting, which means that all participating institutions must report them for all students:

1. First name
2. Last name
3. Date of birth
4. Enrollment status (full-/part-time)
5. Dates of attendance
6. Graduation indicator and date

The optional, or additional, data elements are summarized in this document, and the availability of these data elements for 2020-21 enrollments are captured in the figure below:

Figure 1. Percentage of 2020-21 Enrollments with Selected Data Elements Reported



While the most of these additional elements are reported at well below a 100 percent rate, a potentially more promising avenue for BLS is described in the document summarizing the additional data elements available in the NSC: “the NSC Research Center is also able to leverage the program-level data elements … to make very precise determination of the student’s academic program level for special research projects.” We describe these program-level data further below.

As a result of changes in the ED’s compliance reporting requirements related to reporting to the National Student Loan Data System for federally aided students (commonly known as “150 percent program rules”) institutions started submitting **program-level enrollment data** to NSC in fall 2014.

Starting in the 2015-16 academic year, it became mandatory for participating institutions to report certain “program-level” data elements to the Clearinghouse to satisfy the National Student Loan Data System (NSLDS) reporting requirements related to the “150 percent rule.”[14] This means that the program-level data elements listed below are available for special research projects in **100 percent of the enrollment records reported since fall 2015**:

- Program indicator flag
- Program CIP Code
- Program Credential Level
- Program Enrollment Status

As in their basic enrollment data, schools submit this program-level data to NSC for all students, *not just federally aided students*.

[14] This refers to the rule that financial aid recipients will be terminated upon reaching 150 percent of the number of credits needed to complete their degree, diploma or certificate program.

Based on the NSC StudentTracker data dictionary, the following variables in the NLSY97 could be informed by NSC data (e.g., either via direct replacement or used to impute these variables):

- CV_COLLEGE_TYPE – Private or Public School
- SCH_COLLEGE_STATUS_X – Enrollment status (2-yr/4-yr/Graduate school) in month X
- YSCH-17400 – Quarters, Semesters, or Trimesters
- YSCH-21800 – Fulltime or Parttime Student
- YSCH-20400 – Enrollment Dates
- YSCH-23450 – Diploma or Degree Received from College X
- YSCH-21300 – Major Field of Study

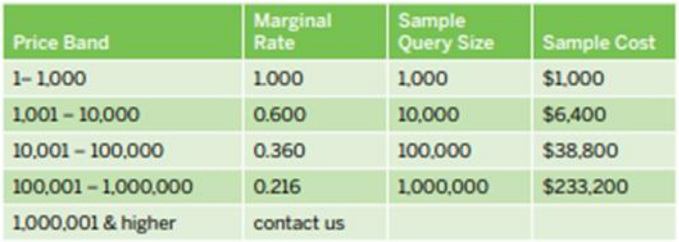
Feasibility

The cost of accessing NSC data depends on the type of data requestor and the nature of the research project. There are seven pricing models, each with different eligibility requirements, but the data returns are generally the same.

NSC designed the StudentTracker for "Other Educational Organizations" with academic researchers in mind. [15] Under this model, a fee is imposed based on the number of records the researcher submits for matching. Additionally, fees are imposed for operational expenses; there is a $500 set-up fee and a minimum fee of $425 per file for administrative costs.

Price per query is calculated in the following manner:

1. Determine the appropriate Price Band, based on the number of records being submitted for matching.
2. Add the Sample Cost of the previous Price Band to the result of the following calculation: Number of records submitted for matching minus Sample Query Size of previous Price Band, then multiplied by Marginal Rate of current Price Band.



Price Band	Marginal Rate	Sample Query Size	Sample Cost
1 – 1,000	1.000	1,000	$1,000
1,001 – 10,000	0.600	10,000	$6,400
10,001 – 100,000	0.360	100,000	$38,800
100,001 – 1,000,000	0.216	1,000,000	$233,200
1,000,001 & higher	contact us		

BLS estimation of 10-15k sample size for NLSY26

- Price band for 10,001-100,000 records is 0.36
- $1,000+(10k-6,400)*$0.36
- $1,000+(15k-6,400)*$0.36
- Cost between $2,296-$4,096

[15] It is not clear whether BLS would fall under this category, but it is the closest match of the available options provided on the website.

Confidentiality, Consent and Accessibility

In order to disclose personally identifiable information and be compliant with FERPA, the school, district, or education authority designates the Clearinghouse as a “school official” only for the specific purposes specified in the agreement. This means that the Clearinghouse can receive both directory, non-directory, and blocked information, but must respect the school’s directory information definition and blocks within the work it performs on behalf of the school.

FERPA also specifically allows schools to conduct research that would permit them to improve the instruction of future students. The agreements with the Clearinghouse meet the requirements for both of these exceptions to the consent requirement for the release of student records.

NSC agrees to only use the personally identifiable student information supplied by the school for the specified purposes[16] and to return or delete the personally identifiable information when the school is no longer under contract with the Clearinghouse; in this way, the school retains control over its data as required under FERPA.

When required under FERPA, a record is made that a student’s postsecondary education record was shared with the high school; the Clearinghouse reserves the right to share with the student the identity of any organization with which the student’s education record was shared.

Costs to access the data are important and often a source of confusion because of NSC’s diverse sources of revenue. Currently, higher education institutions pay no fee to participate in NSC or to submit data and reap the administrative benefits and cost savings that NSC offers. For example, NSC verifies student enrollment and degree information for loan servicers and guarantors, employers, and background check firms and charges a fee to these users. The services are provided on behalf of institutions that pay nothing. Indeed, in the absence of such a service, institutions would have to devote staff resources to handle these verifications, so the service allows them to serve students better, at lower costs.

NSC receives no funding or fees from the Department of Education (ED) for regular services. This includes the regulatory reporting and compliance services for the institutions, including Student Status Confirmation Reports (SSCRs), Federal Student Aid (FSA) enrollment roster reporting, and Gainful Employment reporting, all submitted directly to ED at zero cost either to the institutions or ED. ED has, on rare occasions, used StudentTracker services for its own research, such as enhancing survey responses for National Postsecondary Student Aid Study (NPSAS) and the Baccalaureate and Beyond Longitudinal Study (B&B), and measuring graduation outcomes for Pell recipients. In these cases, ED pays the standard fee per student searched, similar to what other researchers and education-related organizations pay.

The vast majority of institutions that participate in NSC also choose to access the StudentTracker service, which provides detailed, student-level data on the enrollments and credentials earned by eligible students. These specialized research reports, which combine and analyze the data other institutions provide, are offered free for all institutions, provided that they submit some optional data elements (beyond what they are required to submit for NSC’s minimum participation levels) and that they also participate in the free NSC verification services. As of 2016, two-thirds of the institutions did this and

[16] What exactly these specified purposes are is not clear. BLS may need to engage with the NSC to determine this.

receive StudentTracker research for free. The remaining one-third opt to pay a fee instead. In these cases, the fee is nominal: for most of the institutions, only $0.05 per enrolled student, assessed annually, for unlimited use of StudentTracker research. A small number of institutions with minimal participation in the NSC are required to pay $0.10 per enrolled student. Other types of organizations, such as high schools or outreach organizations pay different fees for this service.

Additional Resources

Dundar, Afet, Shapiro, Doug. "The National Student Clearinghouse as an Integral Part of the National Postsecondary Data Infrastructure." National Student Clearinghouse Research Center. 2016.

Dynarski, Susan M, et al. "The Missing Manual: Using National Student Clearinghouse Data to Track Postsecondary Outcomes." Educational Evaluation and Policy Analysis, vol. 37, no. 1S, 2015, pp. 53S–79S.

NCES Student Aid Study

User Experiences with National Student Clearinghouse Data

National Student Clearinghouse StudentTracker Overview

National Student Clearinghouse FERPA Compliance

National Student Clearinghouse How to Subscribe

National Student Clearinghouse StudentTracker Codebook

National Student Clearinghouse StudentTracker J-PAL

NSC Additional Data Elements

National Postsecondary Student Aid Study (NPSAS) Data File Documentation

Organizations looking to partner with NSC contact email: markk@studentclearinghouse.org

Parties interested in joining the Postsecondary Data Partnership contact email: PDPService@studentclearinghouse.org

National Student Loan Data System (NSLDS)

Relevance

The NSLDS is a federal database that tracks Pell grant and federal student loan award amounts and disbursements. It provides a centralized, integrated view of federal student aid loans and grants that are tracked through their entire lifecycle from aid approval through disbursement and repayment. The NSLDS is one of the data sources used in the National Postsecondary Student Aid Study (NPSAS), making use of NSLDS data related to postsecondary enrollment, loan repayment, income, and demographic information. Data are only available for federal student loan and Pell grant recipients.[17] The NSLDS Pell Grant and loan files include information on the year of interest and a complete federal grant and loan history for each student.

The NSLDS provides the most comprehensive source of data on the federal student loan program. It includes student- and borrower-level data that covers the entire life of a borrower's loans. It includes records and dates for each loan's status changes such as when the loan is disbursed; when it is in the in-school period; when it is paid in full; or if it enters repayment, default, deferment, or forbearance. It therefore provides information on patterns of repayment over long periods of time. NSLDS also includes information on the repayment plan for borrowers under the Direct Loan program.

One major limitation of the data is that NSLDS does not track cash flow. It reports a borrower's loan status, but not his monthly payments over time. Such information must be inferred from annual changes in the borrower's loan balance. Finally, the NSLDS only includes information on a borrower's loans, other federal student aid, and the school he attended. It does not include other information about the borrower during repayment, such as income, employment status, etc.[18]

Unfortunately, NSLDS does not make publicly available a codebook or data dictionary. And, in fact, very little specific information on the contents of the NSLDS is made available through NSLDS directly. However, NSLDS does indicate that information for NSLDS comes from the following sources:

- Guaranty Agencies, for information on the Federal Family Education Loan Program (FFELP)
- Department of Education Loan Servicers (ED Servicers)
- Department of Education Debt Collection Services (DCS), for information on defaulted loans held by the Department of Education
- Direct Loan Servicing (DLS), for information on Federal Direct Student Loans
- Common Origination and Disbursement (COD), for Federal Grant Programs information
- Conditional Disability Discharge Tracking System (CDDTS), for disability loan information
- Central Processing System (CPS), for aid applicant information
- Schools, for information on Federal Perkins Loan Program, student enrollment and aid overpayments.

Accuracy

NSLDS is updated frequently, though there is some variation for different data elements. Enrollment, for instance, updated at least monthly, while debt and loan information (from DCS and DLS, respectively) is reported weekly. NSLDS is heavily audited to ensure it is properly tracking loans and their repayment through time. For this reason, it is generally considered the best source of student loan data.

[17] Private student loan debt accounts for 8.4 percent of all outstanding student loan debt (https://educationdata.org/student-loan-debt-statistics).

[18] For more information, see the following Statement before the US House of Representatives Committee on Education and the Workforce on Data on the Federal Student Loan Program.

For the NPSAS linkage project, for example, federal student loan amounts are drawn only from NSLDS, even though other sources of information about student loans exist, including records held by universities and the students themselves. NPSAS deemed that using information from these sources would not improve the accuracy of the data (and reduce it in terms of student self-reports in the short NPSAS interviews) and increase the overall response burden of NPSAS. The NSLDS is updated each academic year.

In the case of the NPSAS, staff student-level data on Pell Grants and federal student loans from the NSLDS were matched to NPSAS sample members. The record match was a cooperative effort between NPSAS staff and the U.S. Department of Education. Sample members missing SSNs were not part of the match. The NPSAS study member had to have at least one valid grant or loan record within the NSLDS database to match successfully. All NSLDS data transfers used a password-protected NCES system transmitting over an encrypted SSL connection.

NPSAS match rates to NSLDS loan data (using data from the 2015-16 academic year) were 69 percent overall, while match rates to NSLDS Pell Grant data were about 60 percent overall. Match rates were considerably lower at public institutions than for-profit private institutions. It is not clear whether non-matches imply that the student did not have federal loans (or a Pell grant). More information on the NPSAS linkage effort and match rates to the NSLDS can be found in the NPSAS data file documentation.

As stated previously, NSLDS matching only returned records of sample members who, at some point in time, had received Pell Grant or federal student loan funding.

Coherence

No data dictionary or codebook was available.

Feasibility

We were unable to find specific information on obtaining access. The NSLDS has been used in some academic research (see Additional Resources section for examples), suggesting that data sharing for research purposes is possible.

Confidentiality, Consent, and Accessibility

The Common Rule and the Privacy Act both apply to the NSLDS. Your organization is responsible for maintaining an accurate and current listing of active users. Employees who have left your organization should immediately be removed from the system by your Primary Destination Point Administrator (PDPA).

Additional Resources

Connecting Student Loan Research and Federal Policy

What Accounts for Gaps in Student Loan Default, and What Happens After

NSLDS contact email: NSLDS@ed.gov

Veterans Affairs Data

The Veterans Health Administration at the VA is America's largest integrated health care system, providing care at 1,255 health care facilities, including 170 medical centers and 1,074 outpatient sites of varying complexity. It serves 9 million enrolled Veterans each year.

USVETS contains data acquired from over 35 sources, including the Decennial Censuses of the U.S. Census Bureau, the Department of Defense (DoD), Veterans Health Administration (VHA), National Cemetery Administration (NCA), Veterans Benefits Administration (VBA), and Social Security Administration (SSA). Some data are extracted from operational and transactional systems such as the Defense Manpower Data Center (DMDC) and the Veterans Assistance Discharge System (VADS). The VA/DoD Identity Repository (VADIR) data base was established to support the One VA/DoD data sharing initiative to consolidate data transfers between DoD and VA. These systems directly capture real-world events and interactions with service members, veterans, and their beneficiaries.

Relevance

USVETS includes all US Veterans. It contains 250 variables from over 35 data sources, including the VHA as well as different administrations with the VA. In addition to variables like name (including SSN), date of birth (and date of death), gender, race/ethnicity, and data indicators for year of inclusion in the data, USVETS data provides information on where veterans live, who they live with, their finances, education and employment, health insurance benefits and utilization, as well as enrollment in other veteran-specific programs.

In terms of Veteran health insurance, USVETS contains an indicator variable for whether a Veteran had active health insurance in the fiscal year (private, Medicaid, or Medicare parts A through D) and whether they participated in the Veteran Group Life Insurance (including the amount of their policy). It also contains information on the number of service-connected conditions a Veteran has.

Variables capturing benefit utilization include whether a Veteran used one of more VA benefits and services in the fiscal year, the date at which the Veteran enrolled in health care benefits, the kind of care received in the fiscal year, whether that care was inpatient or outpatient, and whether it occurred at a VA facility. Additionally, there is pharmacy utilization information specifically if the Veteran received VA pharmacy care in that fiscal year and the costs incurred by both inpatient and outpatient pharmacies.

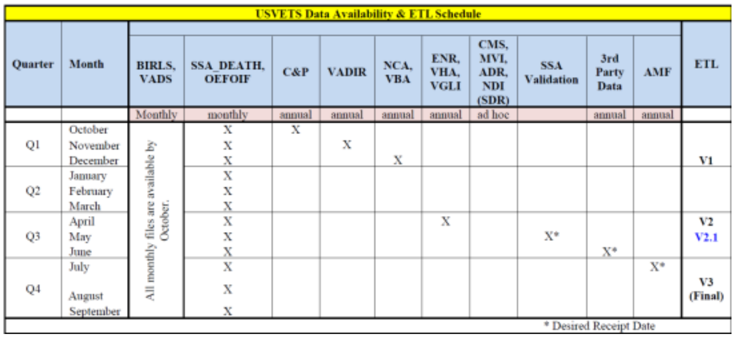
Accuracy

The National Center for Veterans Analysis and Statistics (NCVAS) has made progress increasing the reliability, accuracy, and timeliness of data to address information needs of users, adopting formal rules and evaluation procedures to integrate data from many sources for USVETS.

The VA Data Governance Council has adopted seven data quality dimensions—accuracy, completeness, consistency, traceability, uniqueness, validity, and timeliness.

Administrative data sources compiled for USVETS are themselves regularly audited, helping to ensure their accuracy for evaluating characteristics of the veteran population for most purposes. While no data lack errors, *USVETS may be considered the gold standard for veteran data*.[19] Moreover, USVETS has Social Security Administration (SSA) review SSN, name, gender, and DOB. Part of this review is to ensure that each record has an SSN and that there are no duplicate SSNs assigned to records.

Timeliness may be somewhat of an issue. In a spring 2019 presentation on the USVETS data, it was noted that the most recent available data was for fiscal year 2017. Part of the delay appears to be due to the SSA review process. The following chart, provided in a webinar on USVETS, details the schedule of data availability in USVETS:



USVETS Data Availability & ETL Schedule

Quarter	Month	BIRLS, VADS	SSA_DEATH, OEFOIF	C&P	VADIR	NCA, VBA	ENR, VHA, VGLI	CMS, MVI, ADR, NDI (SDR)	SSA Validation	3rd Party Data	AMF	ETL
		Monthly	monthly	annual	annual	annual	annual	ad hoc		annual	annual	
Q1	October	All monthly files are available by October.	X	X								
	November		X		X							
	December		X			X						V1
Q2	January		X									
	February		X									
	March		X									
Q3	April		X				X					V2
	May		X						X*			V2.1
	June		X							X*		
Q4	July		X								X*	
	August		X									V3 (Final)
	September		X									

* Desired Receipt Date

Feasibility

The VA Information Resource Center's (VIReC) VA/CMS Data for Research Project serves as the data custodian for Centers for Medicare and Medicaid Services (CMS) and United States Renal Data System (USRDS) data for VA research use. The project warehouses and provides data from CMS and USRDS to VA researchers. In addition, the project serves the VA research community by providing education and assistance to VA researchers using these data and conducting research on Veterans' use of Medicare and Medicaid services. Whether non-VA staff can also access this data source appears to be matter that the Office of Enterprise Integration at the VA would decide.[20] The data is hosted on the Department of Veterans Affairs Business Intelligence Service Line (BISL) SAS grid, which is currently non-research.

These private organizations require data use agreements that prohibit releasing the source files and detailed documentation. A similar arrangement with VBA exists that precludes the release and documentation of the source data.

The US Census Bureau is acquiring US Veterans Data through an MOU with the US Department of Veterans Affairs. The NCHS data linkage website notes that VA data is a "future" data linkage. VA data is also used in the NPSAS.

Confidentiality, Consent, and Accessibility

[19] United States Veterans Eligibility Trends and Statistics (USVETS) Support USVETS Data Quality Assessment Version 3.1, 20 MAR 2019

[20]

While the VA has engaged in several data sharing projects and appears to prioritize using VA data to improve knowledge about veterans' issues, some of the language in the VA data strategy report would seem to suggest that individual consent may be needed to share an individual's data:

Sharing of Veteran data – by VA or non-VA parties – when regulation and policy permit organizational discretion (for example, for purposes other than treatment, payment, health care operations, or meeting legal requirements), should be based on the Veteran's meaningful choice to permit sharing their information for that specific purpose. Timely, clear, relevant, concise, complete, and comprehensible information must be provided to the Veteran to serve as a basis for their free and informed choice. A Veteran's preference to change their mind about sharing or not sharing their information should be facilitated, with the understanding that information that has already been shared may not be able to be retrieved or retracted. A Veteran's choice(s) about data sharing must not be the basis to deny care or benefits to which they are otherwise entitled.

Additional Resources

2018 Health Services Research supplemental issue that focuses on the linkage of United States Department of Veterans Affairs (VA) and non-VA datasets to examine a range of topics.

Department of Veterans Affairs Data Strategy

Access to VA Data for Research and Quality Improvement Use

US Veterans Eligibility Trends and Statistics

Contact email for Open Data Lead: lisa.mavrogianis@va.gov

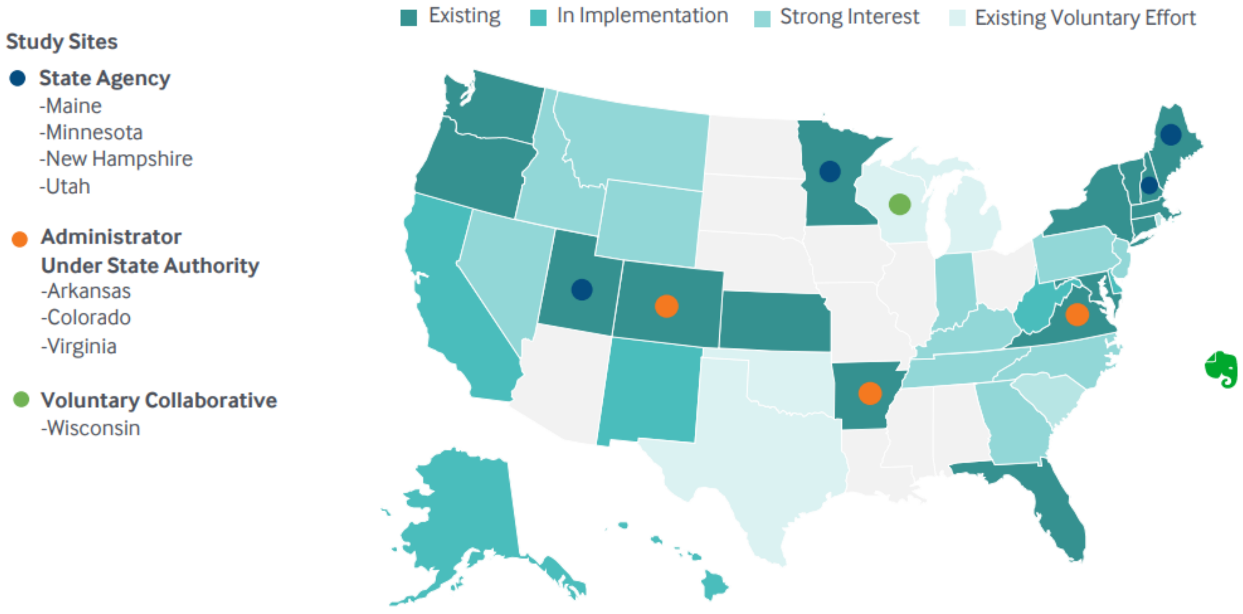
All-Payer Claims Databases (APCD)

All-Payer Claims Databases (APCDs) are large State databases that include medical claims, pharmacy claims, dental claims, and eligibility and provider files collected from private and public payers. APCD data are reported directly by insurers to States, usually as part of a State mandate.

Relevance

APCDs have been created or implemented in 21 states to collect and aggregate information on payment for health services from commercial health insurers, some self-insured employee benefit plans, and the Medicaid and Medicare programs (Exhibit 1). Another 11 states have indicated strong interest in implementing an APCD. And in some states, APCDs have been created through a voluntary effort by stakeholders such as health care systems and researchers.

Exhibit 1. State Activity on All-Payer Claims Databases



Source: Adapted from The APCD Council with permission. © 2009-2020 University of New Hampshire, The APCD Council, National Association of Health Data Organizations. All Rights Reserved.

The following report by The Commonwealth Fund provides detailed profiles of eight state APCDs.

Variables in APCDs typically include information on:

- Member Eligibility
- Medical Claims (service-level remittance with clinical diagnosis codes, medical procedure codes, and charges and payments data)
- Pharmacy Claims (service-level remittance with drug-dispensing, pharmacy and prescribing physician, and charges and payments data)
- Dental Claims (service-level remittance with clinical diagnosis codes, dental procedure codes, teeth treated, and charges and payments data)
- Providers (provider identifiers, such as the National Provider Identifiers (NPI), with provider name, practice location(s), and specialty data for all providers on all other files).

In some states, demographic variables, however, are limited and, when present, are not always of high quality. In some states it appears that PII is removed from the data before leaving the data submitters' systems (e.g., members' names and dates of birth). Some states (e.g., Minnesota) do not collect SSNs in any form.

More specifically, the APCD data structure for inpatient and outpatient claims incorporates data elements from the electronic CMS-1500 and UB04 claims forms. Each claim includes identifiers for patient, provider, and insurer, date of service, charges, diagnosis codes, and Current Procedural Terminology (CPT) and Healthcare Common Procedure Coding System (HCPCS) codes for medical procedures, services, and supplies. The pharmacy claims data structure is based on a National Council for Prescription Drug Programs (NCPDP) standard format, containing identifiers for patient, provider, and insurer, charges, National Drug Code (NDC), and drug name.

Some other benefits of APCD data:

- They include information on private insurance that many other datasets do not.
- They include data from most or all insurance companies operating in any particular State, in contrast to some proprietary datasets.
- They include information on care for patients across care sites, rather than just hospitalizations and emergency department visits reported as part of discharge data systems maintained by most States through State governments or hospital associations. They also include large sample sizes, geographic representation, and capture of longitudinal information on a wide range of individual patients.

Accuracy

In New York State, for instance, the relevant statutory authority requires the state agency to "implement quality control and validation processes to provide reasonable assurance that APD data released to the public is complete, accurate, and valid."

CMS has clear guidelines for claims submitted for Medicaid patients and penalties are incurred for erroneous claims, which should ensure a high-quality data source. The following Agency for Healthcare Research and Quality report from 2017 conducted an examination of the validity and coherence of APCD data for 7 states, looking at the accessibility of data, the basic usability of data and documentation, availability of key data elements, and data validity and accuracy.

In the linkage study using Utah APCD and Cancer Registry data, the authors reported an 82.4 percent linkage rate, of which 66 percent were perfect matches.

An AHRQ review of transparency in seven APCDs reviewed the range of missingness for key variables used for identifying and linking health care encounters. The study found 100 percent consistency (non-missingness) across quarters and years in unique patient identifiers, services date (day, month, and year) and claim status (though the available categories varied by APCD). Consistency for bill type, service type, and billing provider type was also very high (between 95 and 100 percent). APCDs were found to have different ways of populating this variable, though standard fields were populated similarly. However, admission date and type, discharge data and status, and admission source varied widely among APCDs in terms of missingness.

Feasibility

APCDs have been linked to electronic health records for research purposes, but these tend to use data from only a single state, such as this study that linked APCD from Utah to the Central Cancer Registry.

BLS would have to negotiate separate MOUs with each state. Given the need to engage with individual states, the risk of discontinuation is moderate to high. Turnover in agency staff or leadership could lead to shifting priorities or capacity to continue to provide data over a long period of time.

Confidentiality, Consent, and Accessibility

Rules and regulations governing the release and use of ACPD data vary by state. APCD legislation in various states is summarized here. In general, states must adhere to HIPAA as well as state-specific laws and regulations.

In New York State, for example, the release of APCD data is covered under Title 10 (Section 350.3) of the New York Codes, Rules and Regulations and allows for the state to release APD data, *including data with identifying elements*, to a New York State agency or the federal government in a manner that appropriately safeguards the privacy, confidentiality, and security of the data. It also allows the data to be released to *other data users* that have met the requirements for maintaining security, privacy, and confidentiality, and have an approved data use agreements with the New York State Department of Health.

In New York State, users who wish to gain access to data containing identifying data elements must submit application that includes an explicit plan for preventing breaches or unauthorized disclosures of identifying data elements of any individual in the data. State review of proposed projects includes an assurance that the "release of identifying data elements reflects the overall goals of confidentiality privacy, security, and benefits to public and population health." In New York State, the relevant statutory authority governing its ACPD is Public Health Law, Sections 206(18-a)(d) and 2816.

Additional Resources

Profiles of State All-Payer Claims Databases

All-Payer Claims Database Council FAQs

Linkage Between Utah All-Payers Claims Database and Central Cancer Registry

Agency for Healthcare Research and Quality FAQ Page

Health Care Cost Institute contact email: info@healthcostinstitute.org

Credit Agency Data

Publicly information on the data available from the three credit agency data sources was sparse. Here we briefly summarize the information we were able to find.

Relevance

Roughly 190 million US consumers have credit bureau files that meet the minimum criteria for calculating a FICO® Score. But 28 million consumers have files with insufficient data to meet these criteria. More than 25 million consumers have no bureau file at all.

The Experian data contains credit score and loan data and covers 300 million US consumers (or 95% of the US population). The data attributes available in the Experian data include consumer demographics like age, gender, marital status, children, and income. We were not able to confirm PII such as SSN, address, name is included, but we would assume these are collected.

The Equifax data includes credit risk scores, consumer age range, geography, debt balances and delinquency status at the loan level for all consumer loan obligations and asset classes. It appears that Equifax prepares an analytic dataset is created from an unbiased ten percent statistical sample of the U.S. credit active population across all geographic boundaries. The analytic data file contains historic data going back to 2005. Although this file is likely to be research-ready, the fact that it is only a 10 percent sample (albeit a geographically representative one) means that will not be suitable for integration into a new NLS cohort. BLS would have to determine whether individual records could be retrieved from the full data set.

Accuracy

No information.

Feasibility

For the Experian and Equifax data, data access entails a purchase agreement with the respective agencies. We were unable to find specific information on cost. The larger question is whether these agencies will share PII that will allow for linking to a survey. Equifax, for example, states that the data available are *anonymous*, non-aggregated granular consumer-level data.

Confidentiality, Consent, and Accessibility

No information.

Additional Resources

Experian Data Overview

Experian Data Infographic

Equifax Data Overview

FICO Score X Data Overview

FICO Score Overview

FICO Data White Paper

Equifax Contact Page

Experian Contact Page

FICO Contact Page

IBM MarketScan

The IBM® MarketScan® Research Databases contain individual-level, de-identified healthcare claims data including clinical utilization, expenditures, insurance enrollment/plan benefit for inpatient, outpatient, prescription drug, and carve-out services for a large population of individuals and their dependents with employer-provided commercial insurance in the United States. Additional information on the IBM MarketScan data set can be found in this whitepaper.

Relevance

The IBM MarketScan is an opportunity sample that is drawn from multiple data sources (including, example, employers, states, and health plans). Despite being an opportunity sample, it is quite large. It claims to contain data for more the 245 million covered individuals, 260 contributing employers, 40 contributing health plans, and 350 carriers. The database contains more than 32 billion service records.

IBM MarketScan consists of three core claims databases, a hospital discharge database and an EMR database, as well as several linked databases, data sets and files that combine claims data with other patient and employee data at the patient level.

Variables included in the IBM MarketScan include demographics, medical information, health plan, financial information, drugs, and enrollment.

Accuracy

MarketScan databases are based on a large convenience sample. Because the sample is not random, it may contain biases or fail to generalize well to other populations. However, these data can complement other data sets or be used as benchmarks against them. Data come mostly from large employers; medium and small firms may be underrepresented, although the MarketScan Research Databases include a large amount of data contributed from health plans.

Feasibility

Data access would appear to be straightforward and would entail incurring a licensing fee. The associated license fees depend on the number of data years and the number of data products requested.

Confidentiality, Consent, and Accessibility

The MarketScan Research Databases address and adhere to the requirement of the Health Insurance Portability and Accountability Act of 1996 (HIPAA). This means that they do not contain any of the data elements prohibited by HIPAA. IBM MarketScan notes that they have taken steps that go beyond HIPAA requires. In particular, MarketScan databases have undergone statistical analysis by a third part to verify that they meet HIPAA requirements for fully de-identified data sets.

Accessing the data appears to require data management software or programmer support.

Additional Resources

IBM MarketScan Research Databases

Academic Research Paper on IBM MarketScan Research Databases

IBM White Paper on IBM MarketScan Research Databases for researchers

Truven Health MarketScan Database

IBM Contact Page

Multidimensional Insurance Data Analytics Systems (MIDAS)

The original purpose of MIDAS was to provide reporting and analytical capabilities of key Affordable Care Act-related data to Centers for Medicare & Medicaid Services (CMS) and other stakeholders. CMS describes the purpose of MIDAS as "provid[ing] mission-critical functionality that Centers for Medicare & Medicaid Services (CMS) requires to implement and manage many provisions of the Affordable Care Act (ACA)."

Recent program needs have required that data in MIDAS also be used to support Marketplace related operational processes, which has changed how the data in MIDAS is being used and created a need to ingest data from new sources to support these operational processes.

The Multi-Dimensional Insurance Data Analytics System (MIDAS) serves as a central repository for capturing, organizing, aggregating, and analyzing CMS's Marketplace data. The data represent the number of unique individuals who have been determined eligible to enroll in a Qualified Health Plan.

All data in MIDAS originates from other internal or external operational systems supporting the Affordable Care Act (ACA). MIDAS ingest data from these upstream systems, such as Federally-Facilitated Marketplace (FFM) Data Services Hub (DSH) Health Insurance Oversight System (HIOS) Health Insurance Casework System (HICS) State-based Marketplaces (SBM) Issuer enrollment systems Integrated Marketplace Access System (IMAS) Small Business Health Opportunity Program (SHOP) Enrollment & Payment Store (EPS) External Data Gathering Environment (EDGE).

Relevance

MIDAS serves as a central repository for capturing, organizing, aggregating, and analyzing CMS's Exchange data for the **38 states** using HealthCare.gov (HC.gov) in 2020, and includes more than 1 million individuals in the system. MIDAS makes available Public Use Files (PUFs) that include data reported to CMS for State-based Exchanges (SBEs). SBEs operate their own Exchanges, with their own platforms, to conduct eligibility determinations, enrollment, and other related functions. In 2020, these states are **California, Colorado, Connecticut, District of Columbia, Idaho, Maryland, Massachusetts, Minnesota, Nevada, New York, Rhode Island, Vermont, and Washington.** In addition, the state-level PUF includes Basic Health Program (BHP) data from New York and Minnesota. The SBEs submit the data to CMS and verify its accuracy as of the date of publication. The PUFs contain data on individual Exchange activity, including health insurance applications, Qualified Health Plan (QHP) selections, and stand-alone dental plan (SADP) selections. They also include demographic characteristics of consumers who made a plan selection.

MIDAS has data from multiple ACA-related systems at Centers for Medicare & Medicaid Services (CMS) and data from external partners including issuers and state-based Marketplaces. The data contained in MIDAS includes:

- Consumer eligibility and enrollment data (includes names, addresses, email, phone, date of birth, Social Security Number (SSN), and consumer-provided income information)
- Issuer Plan Management data
- Consumer system account data (includes name and email address)
- Issuer Vendor Management data (includes financial account information)

The dataset contains the following variables:

- **County**: The County FIPS Code for the home address provided by the Marketplace applicant.
- **State**: The state of residence selected by the Marketplace applicant.
- **Plan Selections**: The total number of unique individuals who have a non-canceled plan selection coverage for the 38 states that use the HealthCare.gov platform, including the Federally facilitated Marketplace, State Partnership Marketplaces and supported State-based Marketplaces.
- **Advanced Premium Tax Credit (APTC):** A consumer was defined as having APTC if his or her Policy Applied APTC amount was greater than $0. Otherwise, a consumer was classified as not having APTC.
- **Cost-Sharing Reduction (CSR):** A consumer was defined as having CSR if his or her CSR variant value was greater than zero.
- **Metal Level**: A consumer's metal level corresponds to the plan policy that he or she selected. Metal level is based on plan level reference data, including Platinum, Gold, Silver, Bronze, and Catastrophic plans.
- **Type of Consumer**
- **Federal Poverty Level (FPL)**: A consumer household income as a percent of the Federal Poverty Level is set when a consumer provides his or her household income data on the application. Consumers provide household income data, along with the number of household member(s)
- **Race:** This field is not mandatory.
- **Age:** calculated based on reported birthdate of the consumer.

Accuracy

MIDAS does collect and share PII, including SSN, name, date of birth, mailing address, and taxpayer ID.[21] The social security number (SSN) is not used directly in MIDAS; however, the SSN can be included in detailed data extracts that MIDAS provides in support of operational processes. Whether the SSN can be used outside of operational processes (e.g., for research purposes) is not clear. Submission of PII is voluntary, however.

Data are directly comparable between the 38 states using HC.gov, but CMS does *not validate application and enrollment figures for SBEs using their own platforms*, and there appears to be some issue with standardizing data from states that use their own platform (identified under the Relevance section above). For example, CMS recommends that caution should be used when making comparisons between states using their own platforms as definitions may vary. More detail on differences in metrics for SBEs using their own platform is available in the PUF Definitions document.

Apart from cross-sectional consistency in the definition of data elements, there is an effort to ensure longitudinal within a given state, but some differences over time remain. CMS notes that metrics have the "same or very similar" definitions across years for the states that use HC.gov., and that SBEs also generally follow the same or similar CMS definitions across years. However, some year-to-year differences in certain metrics may arise from changes and clarifications to reporting. Data may also vary between SBEs due to differences in reporting systems. In addition, as SBEs operate under different Open Enrollment Periods, the length of the reporting periods can vary on a yearly basis.

Coherence

The variables in MIDAS appear to align well with insurance-related questions in the NLSY97, but much of the information in MIDAS goes well beyond what was captured in that NLS cohort.

Feasibility

MIDAS is used for a wide range of operational purposes, such as consumer assistance, qualified health plan certification, oversight and financial integrity, and coordination with Medicaid and CHIP. The data are also used internally by CMS data analysts for support analytics, reporting, research, and surveys. However, we did not find instances of MIDAS being used for research externally. And it is not clear that there is legal authority for CMS to share this data for anything other than operational purposes.

Confidentiality, Consent, and Accessibility

The legal authority to use and disclose the PII in MIDAS derives from the ACA (1411(g)) and must be used to ensure the efficient operation of the Exchange: 45 CFR 155.260. CMS has established Information Exchange Agreements (IEAs) with IRS, state-based exchanges and state Medicaid and CHIP agencies. CMS has Computer Matching Agreements (CMAs) with the SSA, IRS, Department of Defense, OPM, among others.

[21] Section 1414 of the Patient Protection and Affordable Care Act (ACA) provides the legal authority to use SSNs.

The ACA (1411(g)) permits the use and disclosure of personally-identifiable information (PII) collected or created by an Exchange to *ensure the efficient operation of the Exchange*. 45 CFR 155.260 was originally drafted with the understanding that Exchange minimum functions would ensure the efficient operation of the Exchange. However, the new version of the regulation permits the Secretary of the Department of Health and Human Services (DHHS) to determine that the disclosure of PII for purposes other than Exchange minimum functions can be made as long as certain substantive and procedural steps are followed and the **consent of the subject individuals is obtained**.

Internal Centers for Medicare & Medicaid Services (CMS) data analysts have access to data containing PII to support analytics, reporting, research and surveys. We were unable to find examples of MIDAS used for research purposes, so we anticipate that BLS would need to engage CMS to explore the possibility of accessing this data source for the NLS. The examples of CMS sharing PII from MIDAS with other federal agencies, cited in this document, appear to be limited to operational purposes.

MIDAS is hosted in a secure, Federal Information Security Management Act (FISMA)-compliant data center. Physical access to the system is limited to data center administrators only. User access is also dependent upon Centers for Medicare & Medicaid Services (CMS) approval and is not accessible to users outside of CMS networks.

Additional Resources

Multidimensional Insurance Data Analytics System (MIDAS) Overview

MIDAS Privacy Impact Assessment

CMS 2020-Issuer-Level-Enrollment-PUF-Methodology

State-based Exchange-Qualified Health Plan 2021 Data Dictionary

National Death Index (NDI)

The NDI is a centralized database of U.S. death records gathered from states' vital statistics offices. The NDI contains person-level information on date and causes of death collected from state death records.

Relevance

Data are updated annually with an approximately 15-month lag (e.g., CY 2019 was expected in March 2021). If the NLSY26 is a biennial survey, as the NLSY97 has become, then this may align with the survey's production schedule.

Death records are added annually, typically 15 or more months after the end of the calendar year. NCHS has an "Early Release" program in which researchers can request data sooner. Early Release files are made available when more than 90 percent of the previous year's death records have been processed, but no later than 6 months after the end of the calendar year. This is typically within a month of the end of a calendar year. Early Release files are updated at least twice before the release of final data. NCHS also tracks the percentage of records from each state that are available in the Early Release file.

Some of the variables in the NDI include:

- Date of death
- State of death
- Death certificate number
- Cause(s) of death International Classification of Diseases (ICD) codes.

A full list of variables can be found in the National Death Index User's Guide.

Accuracy

In order to qualify for an NDI match, the program requires at least one of the following combinations of data items:

- First and last name + SSN
- First and last name + month and year of birth
- SSN and DoB and sex.

That these data items are available for linking suggests that the linkage error will be low.

The NDI data originate as state-level administrative data. Such data are deemed to be highly accurate in the sense of being a comprehensive set of all death certificates recorded by state vital statistics offices. The NDI contains death certificate information for death records on file in state vital statistics offices for all 50 states, the District of Columbia, New York City, Puerto Rico, and U.S. Virgin Islands. Deaths that occur outside of these U.S. registration areas are not included in the NDI.

Coherence

The data in the NDI could be used to supplement, or replace, the variables that record the death of biological children (or other family) members. Because these are very salient events in respondents' lives, it is unlikely that the NDI would necessarily provide a great improvement in the accuracy of these variables.[22] In addition, the information in the NDI could be used to refine the survey sample from round to round.

Feasibility

NCHS has linked data from various surveys with death certificate records from the National Death Index (NDI) and makes available for research a restricted use Linked Mortality Files (LMF). NDI mortality data is already linked to NCHS survey participant data, which has allowed researchers to investigate the association of a wide variety of health factors with mortality. The NCHS linkage suggests that the feasibility of a similar linkage with a future NLS cohort is high.

[22] Though it could also be that respondents might be unwilling or unable to share this information during a s urvey interview.

The important role that the NDI plays, both in the NCHS match and for the many operational purposes in which it is used, suggests that the risk of discontinuation is low.

Confidentiality, Consent and Accessibility

The Public Health Service Act (42 U.S.C. 242m) provides in Section 308(d) that the data collected by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), may be used only for the purpose of health statistical reporting and analysis.

Furthermore, the applicant has assured NCHS that the identifying information:

1. Will be used only for statistical purposes in medical and health research.

2. Will not be used as a basis for legal, administrative, or other actions which may directly affect those particular individuals or establishments as a result of their specific identification in the study or project.

3. Will be used only for the study or project described in the approved NDI Application Form.

NCHS assures each NDI user that the identifiable data submitted on the user's study subjects to NCHS are kept confidential and secure before, during, and after the NDI computer matches. The user's data are protected by the Public Health Service Act [42 U.S.C. 242m Section 308(d)], as well as by the federal Privacy Act of 1974, both of which stipulate that data may only be used for the user's proposed study and may not be released to other parties without the user's permission.

To ensure confidentiality of data, NCHS provides safeguards including the removal of all personal identifiers from analytic files. Additionally, the files containing the linked 2014 NHCS-2014/2015 NDI data are only made available for research use at one of the NCHS Research Data Centers (RDCs) or one of the Federal Statistics Research Data Centers (FSRDCs) located across the country.

The fees for routine NDI searches consist of a $350.00 service charge plus $0.15 per user record per year of death searched. For example, 1,000 records searched against 10 years would cost $350 + ($0.15 x 1,000 x 10) or $1,850. Fees for the "NDI Plus" service—which includes cause of death codes for each record—are $0.21 per user record per year searched. For subjects that are known to be deceased NDI charges $0.15 per subject for routine searches and $5.00 per subject for NDI Plus searches, regardless of how many years are searched. Volume discounts are available. More information on user fees is available on NCHS' user fees worksheet.

Additional Resources

National Death Index J-PAL

National Death Index FAQ

National Death Index Repeat Request Form

National Death Index User's Guide - Chapter 2 - Preparing Your Records: Record Layout and Coding Specifications

Compendium of Administrative Data Sources

SNAP, WIC, and TANF

Studies of various welfare programs have found high error rates in the reporting of program receipt in different surveys and programs. For example, 60% of welfare recipients in the Current Population Survey (CPS) and the same share of pension recipients in the American Community Survey (ACS) fail to report receipt (Meyer and Mittag 2019b). Celhay, Meyer, and Mittag (2021) link multiple surveys to administrative records from New York State covering 2007–2012. They report that the probability of a false negative response of SNAP receipt varies from 19% in the SIPP to 42% in the CPS, while the false positive rate varies from just above 0.5% for cash welfare in the SIPP to just above 2% for SNAP in the ACS. Moreover, these studies also demonstrate that response errors are not independent of other respondent characteristics—or the true value of the variable—so that they like bias both causal and descriptive estimates obtained from survey data.

Relevance

Information about SNAP, WIC and TANF was collected in both the NLSY79 and the NLSY97. There is considerable interest in the effect of public assistance receipt on a range of outcomes (including interest in public assistance receipt as an outcome itself); however, research using household survey data has been hampered by the misreporting of program receipt. The useful of the SNAP and WIC-related questions in the NLSY97 were reduced when the questions about SNAP and WIC receipt were combined into a general food assistance question after 2009. This may have been driven in part by concern over respondents' ability to accurately distinguish between these two food assistance programs.

All members of the benefit unit, including children, are included in the administrative records. SNAP is a large program, providing assistance to more than 20 million households and 40 million people. WIC is a much smaller program, with about 6 million people. TANF is even smaller, with just under 2 million total participants and 800,000 families.

Accuracy

In general, the accuracy and completeness of those data elements that are used for programmatic purposes is high. Because recipients must provide their SSNs, and a range of PII, when applying for program benefits, linkage error tends to be low if the same PII is available in the other data. Match rates of SNAP data to Census Bureau's Numident file were on the order of 98 percent. Match rates to Census survey data, which no longer collects SSNs from respondents, was lower: on the order of 90 percent.[23]

Coherence

The NLSY97 provided monthly arrays of recipiency, benefit amount, and the benefit unit, for SNAP, WIC, and TANF through 2009. After 2009, only any receipt of program benefits from SNAP, WIC, and TANF

[23] Technically, this is the match rate of the survey to the Numident from which Protected Identification Keys (PIKs) are generated.

since last interview was reported. And after 2011, the respondents were no longer asked to distinguish between SNAP and WIC receipts.

Administrative records would align well with either the pre-2009 event history format or the post-2009 format. Despite variation in content and format, state program data contains, at a minimum, a record for every individual in the benefit unit, and for each month that the benefit unit received assistance, as well as the benefit amount in each month of receipt.

Feasibility

Obtaining nationally representative SNAP, WIC or TANF records would be a very difficult undertaking, as it would involve negotiating separate data sharing agreements with each state from which data is required. States differ greatly in their willingness to share data. This willingness depends on factors such agency leadership's interest in research products, agency staff capacity and technical skill, as well as political factors in the state. States generally require a program benefit from sharing their data. And the federal agency may require a general statement of benefit to the program as well.

Reasonable coverage could be achieved by targeting several populous states (e.g., California, Texas, Florida, New York, Illinois). However, the Census Bureau's experience in obtaining data from some of these large states (especially California, Florida, and Texas) suggests even this more modest aim may be difficult to achieve.[24] Moreover, the Census Bureau's data acquisition effort has not had as much success obtaining contemporaneous state data on an ongoing basis.

The risk of discontinuation from any given State for SNAP, WIC or TANF is substantial. Even with an MOU in place, there isn't any effective recourse if a state decides it is no longer in its interest to share its data. Changes in agency leadership, in the state's political or economic climate, and in agency staffing could prompt a state to reconsider the value of sharing its data. Even without a change in leadership or staff, state agencies may come to question the benefit to them of continuing to share data, especially if the data sharing arrangement is straining agency resources.

Confidentiality, Consent, and Accessibility

SNAP and WIC

SNAP law and regulations require that state agencies execute a data exchange agreement that specifies the information to be exchanged and the procedures to exchange such information. Therefore, as a first step, agency heads seeking to share SNAP data should agree on a process that will lead to an executed data sharing agreement that describes the information to be shared, with whom, and by what method. The agreement should clearly state the interest for the information sharing and the need to balance such interest with the interests of confidentiality and privacy. For successful implementation of data sharing, states or counties should consider forming two working groups: a Program Group and a Legal Group.

USDA Food and Nutrition Service (FNS) does not maintain a central repository of state SNAP or WIC administrative records, so NLS would have to negotiate separate MOUs with each state agency from which it would like to obtain data. For example, for the SNAP and WIC (and TANF) data it has obtained

[24] Even though Census has acquired SNAP and TANF data from New York, the continued provision of that data is uncertain.

thus far, the Census Bureau has signed MOUs with each of the relevant state agencies (which in some cases is a single state agency).

Additional notes:

- Recipients of data released by the SNAP program are required to protect the data against unauthorized disclosure (7 C.F.R. § 271.1(c)(2)).
- SNAP legislation and regulations actively safeguard the personally identifiable information (PII) of applicants and recipients of SNAP benefits while permitting data sharing with a number of other public programs.
- State SNAP agencies must execute data exchange agreements with other agencies before exchanging information, specifying information to be exchanged and procedures used for the exchange (7 C.F.R. § 273.2(a)(4));
- Recipients of data released by the SNAP program are required to protect the data against unauthorized disclosure (7 C.F.R. § 271.1(c)(2)).

One indication that FNS might be moving in the direction of centralizing SNAP data is in the 2018 Farm Bill, which allows state agencies to establish a longitudinal database containing information about households that receive benefits under SNAP.[25] The database must be used solely to conduct research on program participation and the operation of the SNAP program. Prior to approving the establishment of such a database, FNS must issue standards for the development of these state databases, including the way data security and privacy protections will be implemented and maintained. No PII (including social security number, home address, or contact information) may be included in those databases. FNS issued additional requirements in October 2020 and grants are expected to be awarded in late summer 2021.

TANF

States have autonomy in data sharing decisions, and these decisions are based on their own state laws, as well as applicable federal laws. Title IV-A of the Social Security Act gives states broad flexibility to implement their respective TANF programs, but also requires states to "take such reasonable steps as the State deems necessary to restrict the use and disclosure of information about individuals and families receiving assistance under the program…" The Privacy Act (5 U.S.C & 552a) permits disclosure of TANF data without an individual's consent for a "routine use" support by a System of Records Notice (SORN) published in the Federal Register. The TANF statute requires data sharing to support eligibility determination and child support enforcement.

Federal statutes also encourage data sharing for research purposes. Section 413 of the SSA encourages research on the impact of TANF on employment, self-sufficiency, child well-being, unmarried births, marriage, poverty, economic mobility, and other factors. A database of projects, called the Pathways to Work Evidence Clearinghouse, that used a "a proven approach or a promising approach in moving

[25] For more information on the proposed SNAP longitudinal database, see https://www.fns.usda.gov/snap/longitudinal-data-project-ldp.

welfare recipients into work, based on independent, rigorous evaluations of the projects" was created under this statute.

The TANF data collaborative (TDC), launched in late 2017, sought to foster the use of TANF administrative data for program improvement and evidence building at the federal, state, and local level. The TDC Pilot Initiative funded eight pilot agencies to support their efforts to build strategic partnerships for data sharing.

Additional Resources

Meyer, Bruce D., and Mittag, Nikolas. Using Linked Survey and Administrative Data to Better Measure Income : Implications for Poverty, Program Effectiveness and Holes in the Safety Net. AEI Press, 2015.

Celhay, Pablo A., Meyer, Bruce D., and Mittag, Nikolas. Errors in Reporting and Imputation of Government Benefits and Their Implications. National Bureau of Economic Review Working Paper, 2021.

Compendium of Administrative Data Sources

Appendix A

Data Source	Included in Final ADS Assessment	Reason for exclusion from Final ADS Assessment
2020 Census	No	To our thinking, survey data could be merged in to provide context at any point based on geography and does not need to be explored at this point.
ACA Enrollment Data	Yes	--
All-Payer Claims Databases	Yes	--
American Community Survey	No	To our thinking, survey data could be merged in to provide context at any point based on geography and does not need to be explored at this point.
Catalist	No	Want to focus on how ADS could be used in questionnaire/data files. At this time, not interested in how ADS could be used for sampling.
Census LEHD Partnership	No	Our understanding is that we'd still need agreements with each state *and* these data would be under the Census umbrella.
Criminal Justice Administrative Records System	Yes	--
Current Population Survey	No	To our thinking, survey data could be merged in to provide context at any point based on geography and does not need to be explored at this point.
Department of Housing and Urban Development Housing Rental Assistance Program Data	Yes	--
DMDC Military Files	No	Want to focus on how ADS could be used in questionnaire/data files. At this time, not interested in how ADA could be used for sampling.
EMR Data (IQVIA)	No	This product does not seem to be at a mature enough level of development for our use.

Equifax	Yes	--
Experian	Yes	--
FICO	Yes	--
Health Care Cost Institute	Yes	--
IBM MarketScan	Yes	--
IRS	No	Difficulty of reaching an agreement for microdata use.
Master Address Data (Black Knight)	No	Want to focus on how ADS could be used in questionnaire/data files. At this time, not interested in how ADS could be used for sampling.
Master Address File	No	Want to focus on how ADS could be used in questionnaire/data files. At this time, not interested in how ADS could be used for sampling.
Medicaid Analytic eXtract	No	Discontinued in 2015
Medicaid and CHIP Files	Yes	--
Medicare Data	No	We don't see this as useful for a new youth cohort; useful once sample members are substantially older. Is there something that we are missing?
Mercer	No	Coverage issues and potential availability of other sources.
Multidimensional Insurance Data Analytics System	Yes	--
National Death Index	Yes	--
National Directory of New Hires	Yes	--
National Immunization Surveys	No	Statutory prohibitions seem too difficult to surmount.
National Student Clearinghouse	Yes	--
National Student Loan Data System	Yes	--
National Vital Statistics Birth and Death Data	No	Want to focus on how ADS could be used in questionnaire/data files. At this time, not interested in how ADS could be used for sampling.
Postal Service Data	No	Want to focus on how ADS could be used in

		questionnaire/data files. At this time, not interested in how ADS could be used for sampling.
Quarterly Census of Employment and Wages	No	We would need to collect EIN or official business name as part of the interview. The better way to do this seems to be UI records or SSA earnings.
Realty Trac	No	Want to focus on how ADS could be used in questionnaire/data files. At this time, not interested in how ADA could be used for sampling. This source might also be used for finding addresses across rounds, but don't see the benefits over other sources for that information.
SNAP Administrative Records	Yes	--
Social Security Administration	Yes	--
SSA Numident	No	Want to focus on how ADS could be used in questionnaire/data files. At this time, not interested in how ADS could be used for sampling.
UI Wage Data	Yes	--
USVA Data	Yes	--
WIC Administrative Data	Yes	--

Appendix B

Relevance of Alternative Data Sources in the Life-Cycle of NLSY26 Respondents

Alternative Data Source	Parents/Guardians	Retrospective Pre-Adult[26]	Young Adult (Ages 18-25)	Prime Working Adult (Ages 25 to 55)	Later Working Adult (Ages 55 to 65)	Later Life (Ages 66+)
CJARS	Relevant, but historical data (e.g., pre-2020) may be more limited	Not available for minors	Relevant	Relevant	Less Relevant as respondents age out of risk for CJ encounters	Less Relevant as respondents age out of risk for CJ encounters
HUD	Relevant. Historical Data should be available for linking	Should not be necessary for *most* youths if parents are linked. Would be relevant for youths not living with parent or guardian.	Relevant for measuring respondent program participation	Relevant for measuring respondent program participation	Relevant for measuring respondent program participation	Relevant for measuring respondent program participation
Medicaid and CHIP	Relevant. Historical Data should be available for linking, though completeness and accuracy (and consistency) may be lower.	Should not be necessary if parents are linked	Relevant for measuring respondent program participation. It is also possible that information on respondents' children could be obtained from linked records.	Relevant for measuring respondent program participation. It is also possible that information on respondents' children could be obtained from linked records.	Relevant for measuring respondent program participation. It is also possible that information on respondents' children could be obtained from linked records.	Relevant, even though Medicaid and CHIP participation declines for this age group
NDI	Relevant for obtaining accurate timing and cause of death for parents (and possibly even grandparents)	Less relevant as mortality risk is low	Less relevant as mortality risk is low	More relevant as mortality risk increases	More relevant as mortality risk increases	Relevant
NDNH	Relevant for matching employment and earnings histories.	Less Relevant. Unlikely that many respondents will have UI-covered earnings at this age.	Relevant for employment status and earnings at this life stage.	Relevant for employment status and earnings at this life stage.	Relevant for employment status and earnings at this life stage.	Less Relevant as respondents age into retirement
NSC	Relevant but data coverage may be more limited.	N/A unless respondent enrolls in postsecondary institution early	Relevant	Less relevant as respondents age of out of prime postsecondary schooling age	Not Relevant	Not Relevant
NSLDS	Less relevant	Not Relevant unless respondent enrolls in	Relevant as respondents may incur and pay down student loan debt	Relevant as respondents may continue to incur and pay down	Less Relevant, though some respondents may still have SL debt at this age	Less Relevant

[26] For this life stage, applicable consent (and therefore access to linkable data) may depend on which parent or guardian originally provides consent for linkage, and what transitions occur in the patterns of co-residence or legal custody between the youth and the consenting parents or guardians.

		postsecondary institution early		student loan debt in later adulthood		
SSA Earnings	Relevant. Complete & accurate information on employment status, earnings available for linking; lack of temporal granularity less of a concern for family earnings records.	Less Relevant. Unlikely that many respondents will have UI-covered earnings at this age	Relevant for employment status and earnings at this life stage.	Relevant for employment status and earnings at this life stage.	Relevant for employment status and earnings at this life stage.	Less Relevant as respondents age into retirement
SSA SSI/OASDI records	Relevant. Historical records available for linking.	Should not be necessary if parents match (and R is part of benefit unit)	Relevant for SSI receipt	Relevant for SSI receipt	Relevant as respondents age into OASDI eligibility	Relevant as respondents age into OASDI eligibility
VA Data	Relevant, but historical data may not be as complete or accurate.	Not Relevant	Relevant	Relevant	Relevant	Relevant